A general quantum circuit framework for Extended Wigner's Friend Scenarios: logically and causally consistent reasoning without absolute measurement events

V Vilasini, Mischa P Woods

A 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*

Extended Wigner's Friend Scenarios (EWFSs) go beyond the standard usage of quantum theory where agents are treated classically, and model agents as unitary evolving quantum systems. This has been the subject of several no-go results: Frauchiger and Renner (FR) suggested that quantum agents reasoning using quantum theory will arrive at logical paradoxes, while other works, e.g. the Local-Friendliness theorem, highlight challenges for having an objective notion of measurement events and for causal reasoning in EWFSs. This raises the question: Is it possible to reliably make and test scientific predictions, and consistently reason about the world when applying quantum theory universally, and without assuming that observed measurement outcomes are absolute? We give a positive answer by developing a general quantum circuit framework for EWFSs. We formalise the concept of Heisenberg cuts by mapping them to distinct channels in a quantum circuit, and prove that FR-type paradoxes can be fully resolved by making explicit the conditioning on the quantum channels that are used in the reasoning process. We also provide concrete rules by which quantum agents can reason and make predictions in a logically and causally consistent manner. Our framework describes all perspectives and predictions of an EWFS within a single, well-defined causal structure, although it allows events to be fundamentally subjective. Moreover, we show that an objective notion of measurement events nevertheless emerges in real-world experiments. Our work demonstrates the possibility of a relational yet operational framework overcoming challenges to scientific reasoning in EWFSs, without modifying the Born rule, quantum unitarity or the axioms of classical logic and probability theory applied to measurement outcomes. This enables analysis and comparison of different EWFS arguments and yields a formal platform to extend existing quantum information methods and studies consistently to the domain of Wigner's Friend Scenarios.

CONTENTS

* vilasini@inria.fr
† mischa.woods@gmail.com

I. INTRODUCTION

Quantum theory is among the most successful physical theories, accurately describing microscopic phenomena. In recent years, efforts have been underway to observe quantum phenomena in larger systems, both for scalable quantum computing and for fundamental tests of physical laws. Hence, it is crucial to consider the implications if quantum theory were universally valid. It is natural to expect a complete and universally valid theory of the physical world to be able to consistently model observers or agents as physical systems of the theory, at least in principle.

Wigner was among the first to concretely consider this question back in the 1960s, through an intriguing thought experiment [1]. Wigner's thought experiment highlights challenges in applying quantum theory to agents: an agent (the Friend) measures a quantum system and observes a classical outcome, which is at odds with the view of an outside agent (Wigner, the "super-agent") who models the Friend's lab as a closed quantum system evolving unitarily, and can perform quantum operations on the lab's quantum superposition state, including "undoing" the unitary evolution of the Friend's measurement. This tension lies at the core of the quantum measurement problem, and Wigner's scenario beautifully highlights how this can have empirical consequences for observers in quantum theory.

More recent works (e.g., [2–4]) explore Extended Wigner's Friend Scenarios (EWFSs) involving additional agents, suggesting even more radical implications for physics. Frauchiger and Renner (FR) suggested that in EWFSs, agents who model each other as quantum systems and, at the same time, reason about each other's knowledge, would arrive at logical contradictions [2]. Specifically, they claimed that the following assumptions cannot all hold simultaneously: the physical predictions follow the quantum Born rule[1] (Q), agents (reasoning using the same theory) can inherit each other's conclusions (C), and measurements yield single non-contradictory outcomes in each run (S). In [5], additional assumptions involved in the FR argument were made explicit: agents' labs can evolve unitarily (U) and the distributive axiom of logic holds for classical outcomes (D). Thus, FR's result suggests that Q, U, C, and D together imply a violation of S, i.e., a paradox where a measurement yields contradictory outcomes.

Other no-go results, starting from the work by Brukner [3] and including the more recent Local-Friendliness (LF) theorem [4], address more ontological aspects of quantum theory through EWFSs. These suggest that observed measurement events cannot be regarded as absolute and objective (under certain compelling metaphysical assumptions relating to causality and free choice). This challenges the notion that objective events, for instance "the light is on in this room", may only hold true relative to something (such as an agent), complicating causal reasoning in EWFSs [6, 7].

FR's claim raises concerns about the consistent usability of quantum theory in EWF scenarios where agents are modelled as quantum systems. The non-absoluteness of events in EWFSs raises concerns about whether the results of scientific experiment can be considered objective facts about the world. More generally, it therefore becomes imperative to formally address the following question:

$\mathcal{Q}$: How can we continue to reliably do science, that is, consistently reason about the world, make and test physical predictions, if unitary quantum theory were valid at the level of agents who have full quantum control over each others' labs and where one does not assume an absolute notion of measurement events?

This issue is pertinent as a sufficiently large quantum computer could act as an agent in these EWFS arguments, necessitating a clear resolution.

These EWFS no-go theorems have been met with several responses, with different interpretations of

[1] Strictly speaking, FR require a weaker version of the Born rule, restricted to 0 or 1 probabilities.

quantum theory suggesting distinct resolutions, especially for FR's paradox. Responses range from conceptual discussions on the implications of FR's results for physics (e.g., [5, 8, 9]), suggestions for additional reasoning rules to avoid the paradox in FR's scenarios (e.g., [10–12]), to arguments against the validity of FR's theorem due to implicit assumptions (e.g., [13–16]).

However, there is no concrete framework for identifying which implicit assumptions (if any) are necessary to recover the apparent FR paradox. In the context of previous reasoning rules, it is often suggested (e.g., [10, 12, 17]) that the validity of an agent Alice's prediction or the ability to reason about Alice's outcome should depend on whether a super-agent undoes Alice's measurement in the future. While this can ensure logical consistency, it raises concerns for causality principles and the efficiency of reasoning. How many possible future operations must be tracked when reasoning about present measurement outcomes? If Bob is spacelike separated from Alice, how does the possibility of a super-agent acting on Alice's lab affect the validity of Bob's conclusions about Alice?

Causality issues are also central to LF-type no-go arguments. Challenges for causality from LF indicate that current quantum causal modeling frameworks cannot account for non-absolute measurement events [6, 7]. These formalisms impose absoluteness of events by assuming the existence of a single global probability distribution over the outcomes of all agents in a given scenario. In [7], the authors show that causal models in any theory assuming absoluteness of events, relativistic causality, and free choice cannot explain LF inequalities' violations. This holds even when allowing cyclic causal structures, raising the challenge of whether a consistent causal modeling formalism exists in quantum theory gives up absoluteness of events, while preserving the other fundamental relativistic and operational principles.

FR and LF-type EWFS arguments are often studied separately, with FR seen as agent-centric and LF as more metaphysical. Despite extensive literature, these often focus on the specific 4-agent EWFS of FR and LF, and there is a lack of a comprehensive and unified framework that addresses both relational aspects from non-absoluteness and operational aspects of agents' reasoning in general EWFSs.

We observe that at the core of all the no-go arguments lies the ambiguity in how a measurement is to be modelled: as an irreversible evolution leading to classical records or a reversible unitary evolution of a closed system (the agents' lab). In quantum theory, this can also be understood as a choice of Heisenberg cut, that distinguishes which parts of an experiment are modelled as classical vs quantum. So far, this has remained more of a philosophical concept that does not formally appear in formalisms for quantum theory. However, Wigner's thought experiment highlights the need to take this seriously and to make explicit how measurements are modelled.

Desiderata for a consistent formalism for EWFSs Keeping these discussions in mind, we motivate some important desiderata for a formalism to address such questions.

Firstly, the formalism must clearly formulate the predictions that quantum theory implies in a Wigner's Friend Scenario, as both FR and LF-type arguments stem from such predictions. It must do so while formally and explicitly specifying the assumptions about how measurements are modelled and what knowledge about measurement outcomes is known, when deriving said predictions.

Secondly, it must be applicable to general EWFSs and must allow for the possibility of modeling agents' labs as unitarily evolving systems on which another agent can have full quantum control. Since this allows a measurement to be "undone" by reversing its unitary evolution, this means that an absolute notion of measurement events and records is not assumed by the formalism.

Thirdly, the formalism must guarantee consistency of the predictions as well as logical statements agents can make using said predictions, within a clear causal semantics: the rules of logical reasoning must be compliant with causality principles such as the impossibility of signaling faster than light.

Fourthly, it is crucial for such a framework which does not assume absolute events to explain how objectivity of measurement outcomes emerges in existing real-world experiments, and reproduce the observable predictions of quantum theory in such experiments.

Finally, it is desirable to have an interpretation-independent formalism, such that the identification of relevant assumptions, and resolution of paradoxes will apply across interpretations of quantum theory, facilitating agreement on the matter.

Within the standard usage of quantum theory, where agents are not regarded as quantum systems, the quantum circuit framework satisfies several of these desiderata. It provides a clear causal semantics, an unambiguous manner to compute predictions using the Born rule while ensuring that these predictions are well-defined and form the basis of consistent reasoning. The circuit implies an information-theoretic causal structure that tells us about the flow of information between different systems, which is compatible with the direction of time. Moreover, choices of future quantum operations do not influence outcomes of earlier measurements. Furthermore, a subscriber of any interpretation of quantum theory can at the least, use such a circuit description as a tool to make and test empirical predictions and reason about the world, as this is independent of whether they believe the circuit to be a true representation of an "ontological state of reality".

Here, we develop a framework incorporating all the desiderata motivated above, by consistently generalizing the quantum circuit formalism to general EWFSs where agents' labs/memories are explicitly included as wires in the circuit. In particular, this yields a concrete and constructive solution to the question $\mathfrak{Q}$, among several other results.

Overview of contributions We provide an overview of the main contributions of this work below.

- **Quantum circuit framework for EWFSs:** In Section II, we review Wigner's thought experiment, then in Section III, we build a comprehensive circuit framework for all EWFSs in quantum theory, accommodating any number and configuration of agents and super-agents. This framework formalises Heisenberg cuts by mapping them to different channels, labeled by a parameter (the setting) that distinguishes whether one refers to the classical outcome of a measurement vs whether one regards it as a purely quantum, unitary evolution. Thus we show that every EWFS in quantum theory can be represented in terms of a single *augmented circuit* which allows to compute well-defined normalised probabilities, relative to a choice of settings.

- **Completeness, consistency, and causality:** In Section IV A, we prove three key properties of the augmented circuit: completeness (all quantum predictions in EWFSs can be recovered within the single augmented circuit), consistency (no contradictory predictions can arise in EWFSs using the augmented circuit), and causality (outcomes depend only on past choices of Heisenberg cuts relative to the causal order of the protocol). These results hold without assuming absolute measurement events or the existence of a unique joint probability distribution for all agents' outcomes.

- **Resolution and root of EWF reasoning paradoxes:** In Section IV B, we apply our formalism to agents' reasoning, demonstrating that the augmented circuit allows agents in any EWFS to consistently reason while simultaneously using the quantum Born rule, unitary evolution of closed quantum systems, classical logic and probability theory applied to observed outcomes. In Section IV C, we prove that any apparent inconsistencies can only arise in a scenario when an additional assumption **I**, which provably fails in that scenario, is imposed. **I** captures that physical predictions are independent of Heisenberg cuts (or how a measurement is modelled), and the result concretely identifies the failure of this assumption as a core reason for apparent EWFS paradoxes.

- **Detailed analysis of the FR scenario and claims:** Our above results establishing the general consistency of quantum theory and logical reasoning, are contrary to FR's claim that "Quantum theory cannot consistently justify the use of itself" [2]. We apply our framework to analyse the FR arguments in full detail, considering the entanglement version of their scenario in Section V (and the original prepare and measure version in Appendix F 2). We show that our formalism yields a simple resolution of the FR paradox even though it reproduces, in an explicit form, every statement made in FR's arguments, and without placing any restrictions on agents' reasoning. In Section V C, we provide a more refined understanding of the message of FR's result, by discussing the role of the **I** assumption in FR's scenario, and more generally in relation to the meta-physical concept of absoluteness of events considered in other EWFS no-go theorems (e.g., [3, 4]).

- **Emergence of objective measurement events and role of causality:** In Section VI, we address how subjective events in EWFS reconcile with objective measurement results in real-world experiments. We distinguish between standard and Wigner's Friend type experiments by identifying concrete criteria for "super-agency" based on the causal structure. We prove that in standard experiments (where agents do not measure each other's memories/labs in a non-trivial manner), predictions become Heisenberg-cut independent, and objectivity of measurement events emerges.

- **Discussions on multi-agent reasoning and physical interpretations** In Section VII A, we discuss classical multi-agent scenarios that can lead to inconsistencies when agents overlook common knowledge and implicit assumptions. This helps us contrast the genuinely quantum aspects of EWFS arguments from classical ones. Based on these insights, we outline a general paradigm for scientific reasoning to ensure consistency in multi-agent contexts, showing how this is incorporated into our formalism. We then comment on the generalisation of our approach to ensure logically and causally consistent, and efficient reasoning in scenarios where agents only have partial knowledge of the protocol. In Section VII C, we discuss the physical interpretations of the concept of settings introduced in our work. In particular, we outline how setting choices in reasoning can be updated over time, similar to Bayesian updates, in light of new observations. Finally, in Section VII B, we discuss the interpretation-independence of our results.

- **A more unified picture:** Our framework provides a more unified platform for several aspects of EWFSs while shedding light on their relations. Specifically, in Appendix H, we discuss the links between FR's argument and Hardy's logical proof of Bell non-locality, highlighting the relations between Heisenberg cuts (given by our settings) and measurement contexts. In Appendix I we provide a more unified view of the relations between our framework and different classes of previous responses to FR's results, and we also discuss how different interpretations of quantum theory could apply our formalism consistently. Although we have focused more on the application of our general framework to resolving FR-type reasoning paradoxes, the framework can also describe the LF scenario. In forthcoming work [18], we apply the same formalism to analyse the LF scenario and derive further insights on the (non-)absoluteness of observed events.

II. BACKGROUND INFORMATION

In Section II A, we start with a brief overview on the role of conditional probabilities in ensuring consistency in theories described in terms of circuits. We then review Wigner's original argument in Section II B. In both cases, we highlight the salient features and insights which are core to understanding our general framework and solution to EWFS paradoxes.

A. Conditional probabilities in operational theories

Operational procedures in any theory involve state preparations, transformations, and measurements on physical systems. These procedures are often represented through circuit diagrams and can be formalised within several existing frameworks such as generalised probabilistic theories and process theories (e.g., [19–21]), the details of which are not pertinent here. The common feature across all these frameworks is that they provide rules for computing the probabilities of measurement outcomes.

These probabilities are conditioned on the relevant preparations, transformations, and measurements in the circuit. For example, if a system S is prepared in state ρ, evolved using transformation $\mathcal{U}$, and measured with $\mathcal{M}$ to obtain outcome a, the probability of a taking value a is $P(a = a|\rho,\mathcal{U},\mathcal{M})$. This probability generally depends on the specific choices of $\rho, \mathcal{U}$, and $\mathcal{M}$; for instance, $P(a = a|\rho,\mathcal{U},\mathcal{M}) \neq P(a = a|\rho,\mathcal{U}',\mathcal{M})$ if $\mathcal{U}$ and $\mathcal{U}'$ are different transformations.

Consider a classical scenario where S is a bit, $\rho = 0$, and $\mathcal{M}$ measures the bit's value, yielding $a \in \{0, 1\}$. Let $\mathcal{U}$ be the identity transformation and $\mathcal{U}'$ be a bit flip. Then (1) $P(a = 1|\rho,\mathcal{U},\mathcal{M}) = 0$ while (2) $P(a = 1|\rho,\mathcal{U}',\mathcal{M}) = 1$. Ignoring the conditioning on $\mathcal{U}$ and $\mathcal{U}'$ would lead to a paradox where $a = 1$ both never occurs (according to (1)) and certainly occurs (according to (2)). This demonstrates the necessity of considering the conditioning information to avoid contradictions, even in simple classical scenarios.

Remark II.1. *It is not always necessary to condition on all the information in a circuit when computing probabilities. If some preparations or transformations are fixed and unchanging during the analysis, conditioning on them can be safely omitted without causing inconsistencies. In a multi-agent protocol, this corresponds to fixed elements that are common knowledge.*

Here and in the rest of Section II, we use the names of the preparations, transformations, and measurements in the conditional probabilities. Later, when developing our framework, we will simplify this by labeling certain transformations with binary labels, $x \in \{0, 1\}$, as only two choices will be relevant in the scenarios of interest.

B. Wigner's friend scenarios: a tale of two evolutions

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 Section 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 quantum theory is universally applicable to measurement devices, agents performing the measurements, and their laboratories. Notably, an agent in this context need not be a conscious human (although Wigner speculated about this aspect in his work); a sufficiently advanced quantum computer capable of measuring another system, storing the outcome in quantum memory, and performing basic computations can serve as an agent in Wigner's Friend no-go theorems and arguments.

Suppose Alice (Wigner's friend) measures a quantum system S prepared in the state $|\psi\rangle = \alpha\,|0\rangle_{\mathrm{S}} + \beta\,|1\rangle_{\mathrm{S}}$ in the computational basis, obtaining a classical outcome a with values $a \in 0, 1$. She stores this outcome in her memory A, initialised to the state $|0\rangle_{\mathrm{A}}$. 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

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

where $|a\,a\rangle_{\mathrm{SA}}$ represents the state of Alice's lab observing the outcome $a = a$. When the memory is initialised to $|0\rangle_{\mathrm{A}}$, the unitary evolution $\mathcal{M}^{\mathrm{A}}_{unitary}$ 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 $|\psi\rangle_{\mathrm{S}}\,|0\rangle_{\mathrm{A}} \mapsto |0\rangle_{\mathrm{S}}\,|0\rangle_{\mathrm{A}}$, and for $a = 1$, it is $|\psi\rangle_{\mathrm{S}}\,|0\rangle_{\mathrm{A}} \mapsto |1\rangle_{\mathrm{S}}\,|1\rangle_{\mathrm{A}}$. These are trace-decreasing evolutions, but summing over all possible outcomes gives the trace-preserving evolution:

$$\begin{aligned}\mathcal{M}^{\mathrm{A}}_{projection} : |\psi\rangle\langle\psi|_{\mathrm{S}} \otimes |0\rangle\langle 0|_{\mathrm{A}} \mapsto \\ |\alpha|^2\,|00\rangle\langle 00|_{\mathrm{SA}} + |\beta|^2\,|11\rangle\langle 11|_{\mathrm{SA}}\end{aligned} \quad (2)$$

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.

This highlights that depending on whether a measurement is regarded as producing classical records (as

in the projection postulate) or as a purely unitary evolution of quantum systems, one would describe it through distinct evolutions of the same initial state. Wigner's thought experiment shows that this ambiguity can have observable consequences if quantum theory is universally valid.

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 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 modelled: using the unitarity (Equation (1)) or the projection postulate (Equation (2)).

Deutsch's version of the thought experiment [22] adds a twist. Suppose Alice leaves a note saying "I observed a definite outcome" without specifying the value. This note is then unentangled from Alice's memory which stores the outcome value, and remains unaffected after Wigner measures Alice's lab. If Wigner's measurement confirms the superposition state of Alice's system and memory, it is at conflict with Alice's note that confirms a definite outcome was observed.

Frauchiger-Renner's work [2] elevated this ambiguity into an apparent logical paradox, by extending Wigner's original set-up to include two friends and two Wigners (or superagents). They propose a no-go theorem which claims that when agents model each others' labs as quantum mechanical systems and reason about each others' knowledge of measurement outcomes using classical logic, they arrive at logical contradictions. They demonstrate this through a particular thought experiment with four agents, where agents reasoning in this manner arrive at an apparent paradox. A review FR's claimed theorem, the associated assumptions and scenario can be found in Appendix C.

On the other hand, the Local-Friendliness theorem [4] extends Wigner's original thought-experiment in a similar manner with four agents, but to address a different aspect than agents' reasoning, proving that a set of seemingly reasonable metaphysical assumptions about a physical theory cannot mutually hold. One of these assumptions is about the absoluteness of observed events, originally discussed in [3].

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 also connecting the predictions of quantum theory, agents' knowledge and statements, 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 (EWFS) is commonly used in the literature to describe specific scenarios such as the Frauchiger-Renner and Local-Friendliness scenarios. However, a general definition has been lacking. We propose a general definition of EWFSs within quantum theory, encompassing all finite multi-agent quantum protocols where agents' memories (in which they store the measurement outcome) or equivalently agents' labs are modelled as quantum systems, and where one agent can have full quantum control over the labs of other agents in the scenario. The formal definition is provided below, and its generality is justified in Appendix A.

Definition III.1 (Extended Wigner's Friend Scenario (EWFS))**.** *An EWFS is a quantum protocol that consists of*

1. *A finite set* $\mathtt{S} := \{\mathrm{S}_1, ..., \mathrm{S}_m\}$ *of systems,*
2. *A finite set* $\mathtt{A} := \{\mathrm{A}_1, ..., \mathrm{A}_N\}$ *of agents,*
3. *A set* $\mathtt{M} := \{\mathrm{M}_1, ..., \mathrm{M}_N\}$ *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* $\mathtt{S}_i \subseteq \mathtt{S} \cup \mathtt{M} \backslash \{\mathrm{M}_i\}$ *of systems (which can include the memories of other agents) that they measure at time* t_i *according to a measurement* $\mathcal{M}^{\mathrm{A}_i}$*, 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* $\mathtt{O}_i := \{0, 1, ..., d_{\mathtt{S}_i} - 1\}$ *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* $\mathtt{S}_i \cup \{\mathrm{M}_i\}$ *that captures the possibility of performing a further operation on the system* $\mathtt{S}_i$ *after the measurement, depending on the measurement outcome.*
7. *A joint initial state* $\rho_{\mathrm{S}_1,...,\mathrm{S}_m}$ *of all the systems* $\mathtt{S}$ *at time* $t_0 < t_i$ *for all* $i \in \{1, ..., N\}$

The protocol takes the form of Figure 1, and we can consider projective measurements $\mathcal{M}^{\mathrm{A}_i} = \{\pi^{\mathtt{S}_i}_{a_i} = |a_i\rangle\langle a_i|_{\mathtt{S}_i}\}_{a_i \in \mathtt{O}_i}$ *without loss of generality.*

The general form of an EWFS is illustrated in Figure 1. 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 modelled as a single measurement by encoding the measurement choice in an initial state, as we show

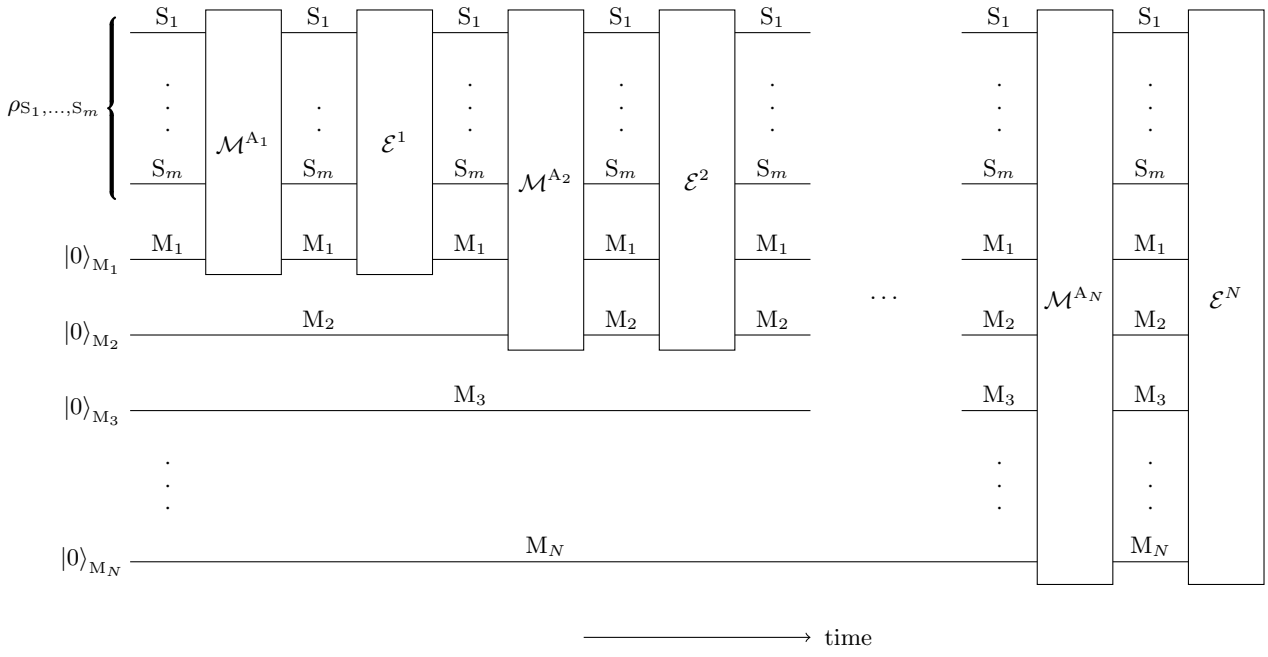

FIG. 1: General form of an EWFS involving N agents and m systems described in the main text. All agents agree on the initial state $\rho_{S_1,\ldots,S_m}$ of the systems and initialise their memories to $|0\rangle$. Each agent A_i performs a measurement on some subset of the systems and on a subset of 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 modelled, depending on their perspective some agents may describe a measurement as a purely unitary evolution as in Equation (3) while others may describe the same measurement as a decoherent process by assigning projectors as in Equation (8). Due to this ambiguity how a measurement is modelled, this diagram is not a fully specified quantum circuit.

explicitly in an upcoming work [18] for the LF scenario, which includes different measurement choices.

In an EWFS, the unitary description $\mathcal{M}^{A_i}_{unitary}$ of a measurement $\mathcal{M}^{A_i}$ corresponds to a CNOT from $\mathtt{S}_i$ to the memory M_i (chosen to be of appropriate dimensions) in the basis $\{|a_i\rangle_{\mathtt{S}_i}\}_{a_i\in\mathtt{O}_i}$ of the measurement. For any state $|\psi\rangle_{\mathtt{S}_i} = \sum_{a_i} c_{a_i}|a_i\rangle_{\mathtt{S}_i}$ expressed in this basis, we have[2]

$$\mathcal{M}^{A_i}_{unitary} : |\psi\rangle_{\mathtt{S}_i}\otimes|0\rangle_{M_i} \mapsto \sum_{a_i} c_{a_i}|a_i a_i\rangle_{\mathtt{S}_i M_i}\,. \tag{3}$$

As highlighted by Wigner's thought experiment, in Wigner's Friend Scenarios (WFS), 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 $\mathcal{M}^{A_{j_1}},\ldots,\mathcal{M}^{A_{j_p}}$ of measurements, it will be useful to denote the corresponding set of outcomes and set of values using vectors

$$\begin{aligned}\vec{a}_j &:= a_{j_1},\ldots,a_{j_p}\\ \vec{a}_j &:= a_{j_1},\ldots,a_{j_p}.\end{aligned} \tag{4}$$

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

Definition III.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 $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, k = \hbar)$, where $\vec{a}_j$ and $\vec{a}_l$ represent the set of outcomes associated with any two disjoint subsets $\mathcal{M}^{A_{j_1}},\ldots,\mathcal{M}^{A_{j_p}}$ and $\mathcal{M}^{A_{l_1}},\ldots,\mathcal{M}^{A_{l_q}}$ (latter possi-*

[2] In the following equation, we have used a pure initial state $|\psi\rangle_{\mathtt{S}_i}$ to make the equations more concise, but it is easy to verify that analogous equations and the same arguments hold for mixed initial states $\rho_{\mathtt{S}_i}$.

bly empty) of measurements in the EWFS, and k is a (possibly empty) set of random variables whose values $\hbar$ encode additional information about the scenario (the exact description of which is to be specified by the rules $\mathcal{R}$). Whenever $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, k = \hbar)$ takes values in $\{0, 1\}$, we refer to it as a logical prediction.

We have already seen an example of the k parameters in Section II A, where they represented the states, transformations and measurements we were conditioning on. We will see more examples later.

Remark III.1. *In probability theory, a conditional probability is by definition "well-defined". This means it is uniquely determined by the event space and is normalised, 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 will see examples where the rules produce both well-defined and not well-defined conditional probabilities.*

Definition III.3 (Statements associated with predictions)**.** *Every prediction $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, k = \hbar)$ 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 $k = \hbar$, then the outcomes $\vec{a}_l$ take values $\vec{a}_l$ with a probability $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, k = \hbar)$."

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 III.4 (Logical statements)**.** *Statements associated with logical predictions are called* logical statements*. When we have a logical prediction $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, k = \hbar) = 1$ (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 $\vec{a}_l$ (do not) take values $\vec{a}_l$." For logical predictions, we can also express the statements using logical operators $\wedge$ (and), $\Rightarrow$ (implies) and $\neg$ (negation). Specifically, the statements associated with $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, k = \hbar) = 0$ and $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, k = \hbar) = 1$ would respectively be*

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

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 III.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 $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, k = \hbar)$ and $P'(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, k = \hbar)$ respectively where $P \neq P'$.*

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

$$S \in \Sigma \quad \Rightarrow \quad \neg S \cap \Sigma = \emptyset, \tag{6}$$

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

At this point, it is important to note that there are two distinct ways in which one can be certain of an outcome value, as illustrated by the following simple example.

- **Scenario 1:** Alice has a fair coin yielding outcome c taking values in $\{heads, tails\}$. She flips the coin and observes $c = heads$ in a round and is then certain that $c = heads$.
- **Scenario 2:** Alice has a biased coin that reads heads on both sides, $P(c = heads) = 1$ and she is certain that $c = heads$ 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 only covers 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 generalisation of the results to include Σ_{obs} is presented in the appendix. This is because a theory only permits observation of certain outcomes when its predictions assign a non-zero 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 considered 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.[3] These probabilities typically do not condition on additional variables k or channels in the circuit, their expressions only refer to 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 formalise 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 III.6 (Conventional predictions in an EWFS)**.** *A conventional prediction in an EWFS is a conditional probability $P_{conv}(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l)$ evaluated by applying the following rules.*

1. *In the given EWFS (see also Figure 1), all the measurements $\mathcal{M}^{\mathrm{A}_i} \notin \{\mathcal{M}^{\mathrm{A}_{j_1}},...,\mathcal{M}^{\mathrm{A}_{j_p}}\} \cup \{\mathcal{M}^{\mathrm{A}_{l_1}},...,\mathcal{M}^{\mathrm{A}_{l_q}}\}$ (those not appearing in the prediction) are modelled as unitary evolutions of their labs (according to the channel Equation (3)).*

2. *For each $\mathcal{M}^{\mathrm{A}_i} \in \{\mathcal{M}^{\mathrm{A}_{j_1}},...,\mathcal{M}^{\mathrm{A}_{j_p}}\} \cup \{\mathcal{M}^{\mathrm{A}_{l_1}},...,\mathcal{M}^{\mathrm{A}_{l_q}}\}$, the projective measurement $\{\pi^{\mathtt{S}_i}_{a_i} = |a_i\rangle\langle a_i|_{\mathtt{S}_i}\}_{a_i\in\mathtt{O}_i}$ is applied on the corresponding system $\mathtt{S}_i$, followed by a CNOT in the same basis from the system $\mathtt{S}_i$ 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 $P_{conv}(\vec{a}_j = \vec{a}_j, \vec{a}_l = \vec{a}_l)$ is then calculated by applying the Born rule to this circuit, using the measurement projectors given above.*[4]

4. *The prediction $P_{conv}(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l)$ is then obtained through the usual conditional probability rule.*

[3] Collapse theories, for instance, would violate this assumption and prescribe different probabilities and rules for computing them.

[4] 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 will decline from doing so at this point, as we will show that these conventional predictions since are recovered as a special case of our to-be-described and rigorously defined general formalism (Theorem IV.1).

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 modelled 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 modelling the measurements (as discussed in Section II B), 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 modelled as unitary evolutions of agents' labs and also preserving the validity of the Born rule.

For this, recall that for a measurement $\mathcal{M}^{\mathrm{A}_i}$ in a basis defined by the projectors $\{\pi^{\mathtt{S}_i}_{a_i} = |a_i\rangle\langle a_i|_{\mathtt{S}_i}\}_{a_i\in\mathtt{O}_i}$ acting on a state $|\psi\rangle_{\mathtt{S}_i} = \sum_{a_i} c_{a_i}|a_i\rangle_{\mathtt{S}_i}$, the associated unitary evolution $\mathcal{M}^{\mathrm{A}_i}_{unitary}$ is as given in Equation (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 $|a_ia_i\rangle\langle a_ia_i|_{\mathtt{S}_i\mathrm{M}_i}$ 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.

$$\mathcal{M}^{\mathrm{A}_i}_{projection}(|\psi\rangle\langle\psi|_{\mathtt{S}_i}\otimes|0\rangle\langle 0|_{\mathrm{M}_i}) = \sum_{a_i}|c_{a_i}|^2\,|a_ia_i\rangle\langle a_ia_i|_{\mathtt{S}_i\mathrm{M}_i}\,. \tag{7}$$

Notice that $\mathcal{M}^{\mathrm{A}_i}_{projection}(|\psi\rangle\langle\psi|_{\mathtt{S}_i}\otimes|0\rangle\langle 0|_{\mathrm{M}_i})$ above can be equivalently expressed as

$$\sum_{a_i}\pi^{\mathtt{S}_i\mathrm{M}_i}_{a_i}\Big(\mathcal{M}^{\mathrm{A}_i}_{unitary}\big(|\psi\rangle\langle\psi|_{\mathtt{S}_i}\otimes|0\rangle\langle 0|_{\mathrm{M}_i}\big)\mathcal{M}^{\mathrm{A}_i\dagger}_{unitary}\Big)\pi^{\mathtt{S}_i\mathrm{M}_i}_{a_i}, \tag{8}$$

where $\{\pi^{\mathtt{S}_i\mathrm{M}_i}_{a_i} := |a_ia_i\rangle\langle a_ia_i|_{\mathtt{S}_i\mathrm{M}_i}\}_{a_i\in\mathtt{O}_i}$, 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 $\mathcal{M}^{\mathrm{A}_i}$ are now given by Equation (8) and Equation (3). 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 evolutions Equation (3) and Equation (8) respectively.

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

$$\Pi^{\mathrm{A}_i}_{x_i} := \begin{cases} \{\pi^{\mathrm{A}_i}_{a_i,x_i}\}_{a_i\in\{\perp\}} = \{\pi^{\mathrm{A}_i}_{\perp,0} := 1_{\mathtt{S}_i\mathrm{M}_i}\} & \text{if } x_i = 0 \\ \{\pi^{\mathrm{A}_i}_{a_i,x_i}\}_{a_i\in\mathtt{O}_i} = \{\pi^{\mathrm{A}_i}_{a_i,1} := \pi^{\mathtt{S}_i\mathrm{M}_i}_{a_i}\}_{a_i\in\mathtt{O}_i} & \text{if } x_i = 1. \end{cases} \tag{9}$$

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 $1_{\mathtt{S}_i\mathrm{M}_i}$. Meanwhile, for $x_i = 1$, a_i takes values a_i in the set $\mathtt{O}_i$ (reflecting the different possible classical outcomes of the measurement, see Definition III.1). We will only use the terminology *outcome*, to refer to the value of a_i when $x_i = 1$. In the case $x_i = 0$, $a_i = \perp$ and we will not regard this as an outcome (we will sometimes refer to this case as the *trivial outcome*).

Then we can model each measurement $\mathcal{M}^{\mathrm{A}_i}$ as the corresponding unitary $\mathcal{M}^{\mathrm{A}_i}_{unitary}$, followed by a (setting-dependent) projective measurement $\Pi^{\mathrm{A}_i}_{x_i}$. The pure unitary picture is always recovered by setting $x_i = 0$, in which case only $\mathcal{M}^{\mathrm{A}_i}_{unitary}$ unitary is applied. The set of choices of settings for all N agents can be represented by a vector $\vec{x} = (x_1, \ldots, x_N)$. With this explicit model, we can transform any EWFS into a standard temporally ordered quantum circuit parametrised 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 IV.1). We call this the *augmented circuit* of the EWFS. This is illustrated in Figure 2, and summarised in the following definition.

Definition III.7 (Augmented circuit of an EWFS)**.** *Given an EWFS of the form of Figure 1, we can associate with it an augmented circuit of the form of Figure 2 which is obtained from the EWFS by replacing each measurement $\mathcal{M}^{\mathrm{A}_i}$ by the corresponding unitary description $\mathcal{M}^{\mathrm{A}_i}_{unitary}$ (Equation (3)), followed by the setting-dependent enlarged set of projectors $\Pi^{\mathrm{A}_i}_{x_i}$ (Equation (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 non-trivial measurement outcome $a_i \in \mathtt{O}_i$ whenever $x_i = 1$. We will 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 non-trivial projectors) being 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 post-measurement 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 V.1, where we show how our framework can also be applied to EWF scenarios where the Born is 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 will see in Theorem IV.1, formalising conventional predictions (Definition III.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 Section 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 non-relational ones).

Definition III.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 III.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, 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 D. Whenever $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi}) \in \{0, 1\}$, we will refer to it as a* logical setting-conditioned prediction.

Definition III.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 III.3 assigns statements to predictions $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, k = \hbar)$, but for setting-conditioned predictions, the scenario parameter values $k = \hbar$ 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 III.4 but with $k = \hbar$ replaced by $\vec{x} = \vec{\xi}$.

Definition III.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 III.9. Similarly, restricting to the set of all logical predictions, we have the corresponding set of all logical statements of the scenario which will be denoted as $\Sigma^{aug}_L \subseteq \Sigma^{aug}$.*

Remark III.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 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, $P(\vec{a}_j = \vec{a}_j) = \sum_{\vec{\xi}} P(\vec{a}_j = \vec{a}_j|\vec{x} = \vec{\xi})P(\vec{x} = \vec{\xi})$.*

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 generalised 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 will 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, we formalise 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 noting that our framework does not assume an absolute and objective notion of measurement events, unlike the majority of existing frameworks for describing quantum information protocols. It does not require the existence of a single objective joint probability distribution $P(a_1, .., a_N)$ over the (non-trivial) outcomes $a_i \in \mathtt{O}_i$ of all measurements in the scenario. Rather, by introducing settings $\vec{x}$, which as we will discuss later Section VII C, model choices of Heisenberg cuts, the framework allows outcome probabilities to be fundamentally relational.

For example, suppose we have an EWFS with two measurements $\mathcal{M}^{\mathrm{A}_1}$ and $\mathcal{M}^{\mathrm{A}_2}$. If we treat the joint system $\mathrm{S}_1\mathrm{M}_1$ (modelling the lab of the first agent) as storing classical outcome records after the measurement while regarding the $\mathrm{S}_2\mathrm{M}_2$ as a purely unitarily evolving quantum system, then $x_1 = 1$ and $x_2 = 0$ and we can only compute probabilities involving the non-trivial measurement outcomes $a_1 \in \mathtt{O}_1$. If we treat both measurements as being associated with classical records, then we have $x_1 = x_2 = 1$ and can compute joint probabilities of a_1 and a_2, the probability for a_1 in the two cases need not generally agree since we use a different setting x_2 in computing this probability in the two cases.

We now show that even though our augmented circuit formalism does not a-priori impose absoluteness of events, 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, we clarify what we mean by causality principles. To formalise this, we begin by defining a directed acyclic graph (DAG)[5] that corresponds to every EWFS. This DAG captures the potential information flow within the EWFS protocol, adhering to the time-ordered sequence of operations. We will refer to this as the causal structure of the EWFS.

Recall that by definition, an EWFS with N agents is modelled as having $2N$ operations w.l.o.g., as each agent performs a measurement $\mathcal{M}^{\mathrm{A}_i}$ 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 will denote by $\mathtt{S}_{\mathcal{O}}$ the subset of all systems and memories $\mathtt{S} \cup \mathtt{M}$ such that the operation $\mathcal{O}$ acts non-trivially on all systems in $\mathtt{S}_{\mathcal{O}}$ and as the identity on the rest of $\mathtt{S} \cup \mathtt{M}$.

Definition IV.1 (Causal structure of an EWFS)**.** *The causal structure of an EWFS with N agents is a directed acyclic graph (DAG) G with the following properties*

1. *G has $2N$ vertices given by the set Vert(G) := $\{V_i^{\mathcal{M}}, V_i^{\mathcal{E}}\}_{i=1}^{N}$, where $V_i^{\mathcal{M}}$ and $V_i^{\mathcal{E}}$ are respectively associated with the operations $\mathcal{M}^{\mathrm{A}_i}$ and $\mathcal{E}_i$, and both pairs of such vertices are associated with the time step t_i, for each i.*

[5] In simple terms, a DAG is a graph with additional structure: 1) The edges have a direction associated with them. 2) By "following the edges in the indicated direction" one can never get back to the starting point, i.e. no "directed loops".

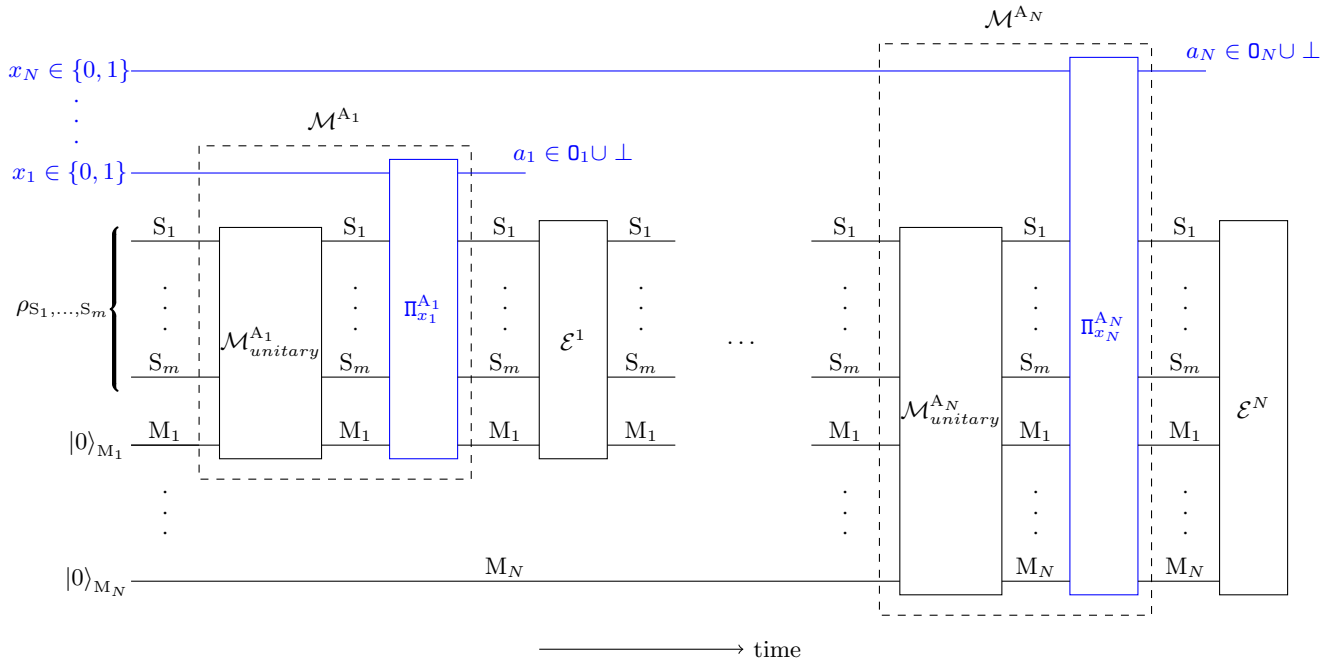

FIG. 2: Augmented circuit for the general form of an EWFS illustrated in Figure 1 that makes explicit the implicit setting choices needed to model the measurement of each agent. This circuit makes it clear that each agent has a choice in how they describe each measurement in the scenario, when the setting $x_i = 0$, the corresponding measurement $\mathcal{M}^{\mathrm{A}_i}$ is modelled as a unitary evolution $\mathcal{M}^{\mathrm{A}_i}_{unitary}$ (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 modelled as the unitary $\mathcal{M}^{\mathrm{A}_i}_{unitary}$ followed by non-trivial projectors that identify the classical measurement outcome $a_i \in \{0, 1, \ldots d_{\mathsf{S}_i} - 1\} := \mathsf{O}_i$. 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 $x_j = 0$.

2. *G contains a directed edge $V \to V'$ for $V, V' \in Vert(G)$ whenever $t_V < t_{V'}$ and $\mathsf{S}_{\mathcal{O}_V} \cap \mathsf{S}_{\mathcal{O}_{V'}} \neq \emptyset$, 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 IV.2 (Directed paths and partial order)**.** *The DAG G associated with an EWFS in Definition IV.1 defines a partial order relation $\prec$ on agents in the EWFS, with $\mathrm{A}_i \prec \mathrm{A}_j$ if and only if there is a directed path from the measurement vertex $V_i^{\mathcal{M}}$ of A_i to the measurement vertex $V_j^{\mathcal{M}}$ of A_j. Furthermore, we will write $\mathrm{A}_i \prec^S \mathrm{A}_j$ whenever there is such a directed path and the system S is included in at least one of the set intersections $\mathsf{S}_{\mathcal{O}_V} \cap \mathsf{S}_{\mathcal{O}_{V'}}$ involved in that path. We use $\mathrm{A}_i \not\prec \mathrm{A}_j$ and $\mathrm{A}_i \not\prec^S \mathrm{A}_j$ to denote the absence of directed paths with the above-defined properties.*

Crucially, notice that the above graph G and induced partial order $\prec$ represent objective properties inherent to the EWFS, independent of the specific settings chosen to model the measurement $\mathcal{M}^{\mathrm{A}_i}$ in the augmented EWFS. This is because the definition only relies on 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 III.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 $\mathrm{A}_i \not\prec \mathrm{A}_j$. As our circuits are acyclic and operations therein have a clear time ordering, we have $\mathrm{A}_i \prec \mathrm{A}_j$ implies $t_i < t_j$, and it is easy to see that this ensures that G is indeed a directed acyclic graph and that $\prec$ 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 Section VII C).

Finally we note that $t_i < t_j$ does not necessarily imply $\mathrm{A}_i \not\prec \mathrm{A}_j$ as it is possible to have two measurements acting at different times but on non-overlapping 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 space-like 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 $\prec$ is compatible with the causal struc-

ture of the spacetime [23, 24].

Theorem IV.1.

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* $P_{conv}(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l)$ *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* $i \in \{j_1, ..., j_p, l_1, ..., l_q\}$ *and* $x_i = 0$ *for all* $i \notin \{j_1, ..., j_p, l_1, ..., l_q\}$.

2. *Consistency: For any EWFS, the set of all statements* Σ^{aug} *obtained in the corresponding augmented circuit (Definition III.10) are consistent according to Definition III.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* $\mathrm{A}_i \not\prec \mathrm{A}_k$ *for all* $k \in \{j_1, ..., j_p, l_1, ..., l_q\}$*, the prediction is independent of the setting* x_i*. That is, for all such* i*, we have the following, where we denote* $P(a = a)$ *as* $P(a)$ *for short and note that* $\vec{x} = (x_1, ..., x_N)$.

$$\begin{aligned}\forall \xi_i, \xi_i', \quad & P(\vec{a}_j|\vec{a}_l, (\xi_1, ...\xi_i, ..., \xi_N)) = \\ & P(\vec{a}_j|\vec{a}_l, (\xi_1, ...\xi_i', ..., \xi_N)).\end{aligned} \tag{10}$$

Whenever a setting-conditioned prediction is independent of a setting x_i, i.e., satisfies Equation (10), we will simply drop x_i from that prediction and denote it as follows.

$$\begin{aligned}& P(\vec{a}_j|\vec{a}_l, (\xi_1, ..., \xi_N)) = \\ & P(\vec{a}_j|\vec{a}_l, (\xi_1, ...\xi_{i-1}, \xi_{i+1}..., \xi_N))\end{aligned} \tag{11}$$

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 the subject of agents' reasoning in EWFS. One immediate corollary of our general consistency result of Theorem IV.1, for agents' reasoning is the following.

Corollary IV.1. *If any two agents use the same choice of settings* $\vec{x}$ *for all measurements* $\{\mathcal{M}^{\mathrm{A}_i}\}_i$ *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 formalised through the settings). More generally, agents can choose different settings depending on their perspective or prediction they wish to compute. Below, we show that our formalism enables agents in an EWFS to reason consistently even when they apply typical axioms of classical logic to observed classical outcomes, and can model each others' labs as unitarily evolving quantum systems and have fundamentally subjective perspectives. In particular, we will consider the logical axioms used in FR's argument, which relate to the inheritance of knowledge (of other trusted agents) and the distributivity of logical statements. We first formalise the relevant logic axioms and the quantum theory dependent assumptions and in the context of our framework.

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 [5], this assumption C was formalised within the framework of modal logic. Here $K_{\mathrm{A}_i}(S)$ 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 [5] 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

$$K_{\mathrm{A}_i} K_{\mathrm{A}_j}(S) \Rightarrow K_{\mathrm{A}_i}(S), \tag{12}$$

More specifically, it is pointed out in [5] that such an inference only needs to be made 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 a that of FR, [5] instantiate the trust structure by considering pairs of agents performing compatible measurements, and only apply C for such pairs. The trust structure will not be relevant for the general solution that we propose here because, as we will show (c.f. Theorem IV.1), the inclusion of the settings in our framework ensures the general validity of Equation (12) in an augmented EWFS independently of the trust structure.

Definition IV.3 (Assumption **C**)**.** *For any two agents* A_i *and* A_j*, Equation* (12) *holds for all statements* $S \in \Sigma^{aug}$.

As noted in [5], the FR argument also uses the distributive axiom of logic. We instantiate this in our formalism below.

Definition IV.4 (Assumption **D**)**.** *For any set of agents reasoning using the augmented circuit, if* $S_1 \in \Sigma_L^{aug}$*, and* $S_1 \Rightarrow S_2$*, then* $S_2 \in \Sigma_L^{aug}$ *holds and we have*

$$K_{\mathrm{A}_i}(S_1 \wedge (S_1 \Rightarrow S_2)) \Rightarrow K_{\mathrm{A}_i}(S_2), \tag{13}$$

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 III.5 as the latter applies to all predictions while the former only to logical predictions.

Definition IV.5 (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, for 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 IV.6 (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 $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi}) = P$ can be derived by applying the Born rule to the EWFS (as detailed in Definition III.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 formalisation of the **U** assumption, we make explicit another implicit assumption in the previous literature, which relates to quantum control over other agents' labs.

Definition IV.7 (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 $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi})$, both choices $x_i \in \{0, 1\}$ are allowed for all $i \notin \{j_1, ..., j_p, l_1, ..., l_q\}$. Moreover, agents can have full quantum control over the labs of other agents, i.e., the measurement of an agent A_i can act non-trivially on the total system $\mathtt{S}_j\mathrm{M}_j$ comprising the lab of another agent A_j.*

We then obtain the following corollary, which follows by construction of our framework along with the general consistency result of Theorem IV.1. We nevertheless include a proof in Appendix J for completeness.

Corollary IV.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 formalisation **Q**, **U**, **C**, **D**, **S** of the FR assumptions Q, U, C, D and S within our framework, showing our version of the 5 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 formalised, allowing room for interpretation (see also Appendix I for references to previous responses to FR's arguments). See Section V C for further discussion on the interpretation of FR's claims in light of our results.

Physically, our assumptions **Q**, **U**, **C**, **D**, **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 [5]), 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, even at a physical level, 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 IV.1 and Corollary IV.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 modelled 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 formalise 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 Section V C after analysing the example of the FR scenario).

Definition IV.8 (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.[6] *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 IV.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 firstly, using Corollary IV.2 we have that the augmented EWFS is perfectly consistent with all five assumptions **Q**, **U**, **C**, **D** and **S**. Further, the consistency of our framework captured by Theorem IV.1 and Corollary IV.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 [2]), 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.

V. A SIMPLE RESOLUTION TO THE FR APPARENT PARADOX

Having developed a fully general framework, here we apply it to an example to show it in action. In particular, we provide a simple resolution to the FR paradox. We focus here on the entanglement version of the FR protocol illustrated in Figure 3 (and reviewed in detail in Appendix C 2), describing the main idea behind the resolution. This version of FR's protocol was proposed by Luis Masanes and Matthew Pusey in their talks. In Appendix F 1, a detailed analysis of the entanglement version of the FR scenario can be found, which explicitly shows all the calculations backing the main points.

Furthermore, while many consider the entanglement version to be equivalent to the original prepare and measure version, it has been suggested by the authors of the FR paper that the entanglement version misses important subtleties regarding the timing information involved in the reasoning process, to which a lot of care has been given in the original FR formulation of the experiment. Our results are fully general and resolve (in particular) the apparent paradoxes arising in both these situations. In Appendix F 2, we resolve the original prepare and measure version of the FR paradox, while giving a statement-by-statement analysis and comparison to show how the paradox completely disappears in our framework even though all the agents can freely reason using unitary quantum theory, the standard Born rule and classical logic, and all the individual statements of FR's original argument can be reproduced.

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 Figure 3. 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 [9]), 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 $\mathcal{M}^{\mathrm{A}_i}_{unitary}$ and when set to $x_i = 1$ models the measurement as the unitary $\mathcal{M}^{\mathrm{A}_i}_{unitary}$ 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 $u = w = \mathrm{ok}$. 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)[7] and ignore the explicit communication channel for the announcement as this is a post-processing. Then we only need to assign settings x_1 and x_2 to the agents Alice and Bob.

[6] Recall that in our framework, the settings $x_i = 1$ for all outcomes a_i belonging to the outcome sets $\vec{a}_j$ or $vec a_l$ appearing in the given prediction $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi})$, the remaining components of $\vec{x}$ can be varied.

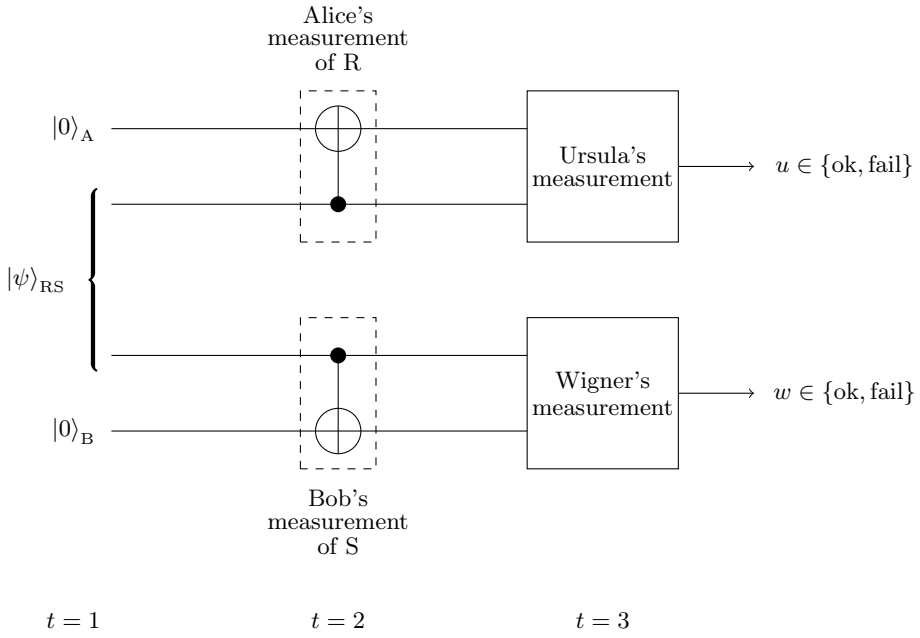

FIG. 3: Circuit that describes the entanglement version of the FR protocol, from the view of the superagents Ursula and Wigner who describe the measurement of Alice and Bob as unitary evolutions. The protocol proceeds as follows: Alice and Bob share a bipartite state $|\psi\rangle_{\mathrm{RS}} = \frac{1}{\sqrt{3}}(|00\rangle + |10\rangle + |11\rangle)_{\mathrm{RS}}$, 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 $\{|ok/fail\rangle := \frac{1}{\sqrt{2}}(|00\rangle \mp |11\rangle))\}$ basis to obtain the outcomes $u, w \in \{\mathrm{ok}, \mathrm{fail}\}$. Note that this circuit alone does not allow the superagents to reason about the classical measurement outcome of Alice and Bob as these are modelled purely unitarily, and hence no classical measurement outcomes a and b are identified.

We can now write the enlarged set of projectors $\Pi^A_{x_1}$, and $\Pi^B_{x_2}$ (c.f. Equation (9)) as

$$\begin{aligned}
\Pi^A_0 &:= \{\pi^A_{0,\perp} = 1_{\mathrm{RA}}\},\\
\Pi^A_1 &:= \{\pi^A_{1,0} = |00\rangle\langle 00|_{\mathrm{RA}}, \pi^A_{1,1} = |11\rangle\langle 11|_{\mathrm{RA}}\}\\
\Pi^B_0 &:= \{\pi^B_{0,\perp} = 1_{\mathrm{SB}}\},\\
\Pi^B_1 &:= \{\pi^B_{1,0} = |00\rangle\langle 00|_{\mathrm{SB}}, \pi^B_{1,1} = |11\rangle\langle 11|_{\mathrm{SB}}\}
\end{aligned} \tag{14}$$

These capture the fact that when Alice's setting $x_1 = 0$, her measurement is modelled as a unitary evolution and the a is set deterministically to the trivial value $\perp$. When $x_1 = 1$, Alice's measurement is first modelled unitarily and then her measurement outcomes $a = 0$ and $a = 1$ are identified by the projectors $\pi^A_{1,0}$ and $\pi^A_{1,1}$. 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 Figure 4. 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 of Figure 2. 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 C 2, the four logical statements involved in the entanglement version of the FR paradox (see Figure 3 for a quick recap of the protocol) are as follows.

$$\begin{aligned}
u = ok \wedge w = ok\\
u = ok \Rightarrow b = 1\\
b = 1 \Rightarrow a = 1\\
a = 1 \Rightarrow w = fail
\end{aligned} \tag{15}$$

The paradox ensues since the last three statements can be combined using classical logic to yield $u = ok \Rightarrow w = fail$, which contradicts the first statement. These statements follow due the following four conventional predictions (Definition III.6) of the 4-agent

[7] We need not consider the case where Ursula and Wigner's measurements are modelled unitarily (setting 0) as this is only the case if there were further super-super agents who measured Ursula and Wigner's labs.

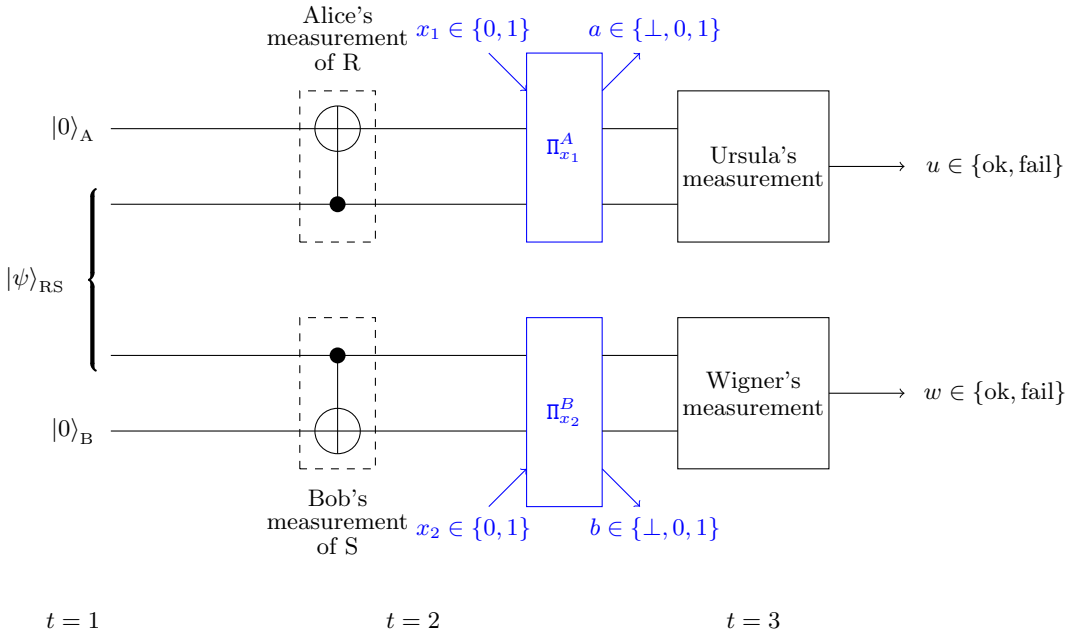

FIG. 4: Augmented circuit for the entanglement version of the FR protocol. Unlike the circuit of Figure 3, this circuit allows agents to model the measurements of Alice and Bob as unitary evolutions and at the same time also reason about their classical measurement outcomes. To model Alice unitarily, 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 set-up 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.

EWFS specified by this protocol, while post-selecting on an experimental run where $u = w = ok$ is obtained by the superagents Ursula and Wigner. The first prediction in the list below guarantees this post-selection will eventually succeed.

$$\begin{aligned} P_{conv}(u = ok, w = ok) &= \frac{1}{12} \\ P_{conv}(b = 1|u = ok) &= 1 \\ P_{conv}(a = 1|b = 1) &= 1 \\ P_{conv}(w = fail|a = 1) &= 1 \end{aligned} \tag{16}$$

From our main theorem, Theorem IV.1, it follows that these conventional predictions are equivalent to the following four setting conditioned predictions (that can be computed from the augmented circuit of Figure 4).

$$\begin{aligned} P(u = ok, w = ok|(x_1, x_2) = (0, 0)) &= \frac{1}{12} \\ P(b = 1|u = ok, (x_1, x_2) = (0, 1)) &= 1 \\ P(a = 1|b = 1, (x_1, x_2) = (1, 1)) &= 1 \\ P(w = fail|a = 1, (x_1, x_2) = (0, 1)) &= 1 \end{aligned} \tag{17}$$

The 3 logical statements associated with the last three setting-conditioned predictions, together with the corresponding original 3 logical statements of FR are given in Table I for comparison.

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

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.

Then we can immediately see that while the original statements can be chained together using classical logical rules to yield $u = w = \text{ok} \Rightarrow w = \text{fail}$ (Equation (C7)), 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 Equation (16) and Equation (17) indeed match can seen by explicitly writing out the expression of the conventional prediction using the Born rule, which we show in full detail in Appendix F 1. We illustrate this here for some of these

cases. Consider the conventional prediction $P_{conv}(u = ok, w = ok)$. This is computed in the circuit of Figure 3 i.e., applying the two CNOT gates to the initial state of $|\psi\rangle_{\mathrm{RS}}$ and memories initialised to $|0\rangle_{\mathrm{A}}$ and $|0\rangle_{\mathrm{B}}$, one obtains $|\psi\rangle_{\mathrm{RASB}} = \frac{1}{\sqrt{3}}(|0000\rangle + |1100\rangle + |1111\rangle)_{\mathrm{RASB}}$. The measurements of Ursula and Wigner act on this state. Indeed our augmented circuit of Figure 4 for the case $(x_1, x_2) = (0, 0)$ is equivalent to the circuit of Figure 3. Thus writing our the two predictions, we have

$$\begin{aligned} &P_{conv}(u = ok, w = ok) \\ =&P(u = ok, w = ok|(x_1, x_2) = (0, 0)) \\ =&|\langle ok|_{\mathrm{RA}} \otimes \langle ok|_{\mathrm{SB}} . |\psi\rangle_{\mathrm{RASB}}|^2 \end{aligned} \tag{18}$$

For comparison, consider the conventional prediction $P_{conv}(a = 1|b = 1)$. This is fully specified by $P_{conv}(a = 1, b = 1)$, 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 $|\psi\rangle_{\mathrm{RS}}$, which is equivalent to applying the $\{|00\rangle, |11\rangle\}$ basis measurement to $|\psi\rangle_{\mathrm{RASB}}$. This is exactly what one would get when using the settings $(x_1, x_2) = (1, 1)$ in the augmented circuit of Figure 4. We have

$$\begin{aligned} &P_{conv}(a = 1, b = 1) \\ =&P(a = 1, b = 1|(x_1, x_2) = (1, 1)) \\ =&|\langle 11|_{\mathrm{RA}} \otimes \langle 11|_{\mathrm{SB}} . |\psi\rangle_{\mathrm{RASB}}|^2. \end{aligned} \tag{19}$$

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

Remark V.1 (On the role of the projection postulate)**.** *As seen in Section II B, Wigner's Friend Scenarios generally involve an interplay of 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.*

Notice that the apparent FR paradox in the entanglement version discussed in this section, arises through the combination of the logical statements in Equation (15) *which are fully implied by the conventional predictions of Equation* (16)*. Computing these predictions (which are measurement probabilities) requires applying the Born rule and unitary modelling of agents' measurements, but does not rely on the projection postulate which also specifies a post-measurement state. However, our resolution of the apparent paradox is also entirely at the level of measurement probabilities given by the setting-conditioned predictions Equation* (17)*. 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 (they specify the measurement basis).*

For instance, for the prediction concerning Ursula and Bob's outcomes u and b, Alice's measurement is modelled unitarily (setting $x_1 = 0$) and the basis in which we describe this unitary does not matter for the predictions, however the basis information for Bob (encoded in the projectors $\{\pi^B_{1,0} = |00\rangle\langle 00|_{\mathrm{SB}}, \pi^B_{1,1} = |11\rangle\langle 11|_{\mathrm{SB}}\}$ of Equation (14) *associated with $x_2 = 1$) is needed for identifying the outcome b and computing its probability. Therefore the setting 1 case need not be thought of as modelling an objective "collapse", but can be regarded as encoding knowledge about a measurement outcome and the basis needed to identify that classical record (independently of the post-measurement 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 highlights that even versions of the FR argument that do not invoke the projection postulate or associated state update rule can be resolved in a similar manner within our approach, without invoking these assumptions, but by being careful about conditioning on the relevant knowledge used in the reasoning.

C. Setting-dependence: a refined interpretation of the FR paradox

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

In Appendix F 3, we show the setting-dependence of predictions in the FR scenario (i.e., a violation of **I**), by explicitly computing them. In particular, recall that we recovered the conventional predictions of the FR scenario equivalently as specific setting conditioned predictions in Equation (17). For instance it follows from the detailed analysis of Appendix F 1 that $P_{conv}(w = \text{fail}|a = 1) = P(w = \text{fail}|a = 1, (x_1, x_2) = (1, 0)) = 1$. $P(w = \text{fail}|a = 1, (x_1, x_2) = (1, 0))$ is indeed setting independent, as one can verify (see Appendix F 3) that $P(w = \text{fail}|a = 1, (x_1, x_2) = (1, 1)) = \frac{1}{2} \neq 1$. Similarly, the setting-dependence of other predictions of the FR scenario can also be verified in our framework.

Using the general results of Section IV and the above analysis of the FR experiment, 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 F 2).

Theorem V.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.*

A refined interpretation 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 formalised a version **Q**, **U**, **C**, **D** and **S** of these, showing that they can always be applied consistency 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 should be interpreted as capturing the validity of unitary "quantum theory" and "classical logic" applied to the knowledge of agents. However, the assumptions are not sufficiently rigorously formalised, due to ambiguities in defining "agents' knowledge", and especially how agents model measurements in each statement they make.

Specifically our results show that if the FR assumptions are interpreted as just capturing the validity of quantum theory and of classical logic, then FR's claimed theorem would be wrong, as we have shown rigorously the general consistency of these assumptions within our framework by developing an explicit consistent model for reasoning in quantum theory where all these assumptions are satisfied.

In order for FR's theorem to be correct, FR's assumptions should be interpreted as imposing a version of quantum theory that ignores choices of Heisenberg cuts (as allowed by our **I** assumption). This distinction necessitates recognizing two versions of quantum theory in the context of EWFSs: (1) *Heisenberg cut independent* and (2) *Heisenberg cut dependent* versions.

In essence, FR's result can be interpreted as revealing a contradiction between version (1) of quantum theory and classical logic within a specific EWFS, while our findings establish the general consistency between version (2) of quantum theory and classical logic across all EWFSs. Then the two sets of results are mutually consistent.

Moreover, the violation of **I**—the dependence of predictions on Heisenberg cuts in EWFSs—is not unexpected once the concept of such cuts is formalised 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 Section II B, the core message is that the ambiguity in how a measurement is modelled, 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.

Explicitly, recall that the two evolutions of an initial state $\sqrt{\frac{1}{2}}(|0\rangle+|1\rangle)_{\mathrm{S}}$ measured by an agent Alice in the computational basis, lead to the following final states of her system S and memory A. Here we label the two cases with the corresponding settings of our framework, where x denotes the setting of Alice's measurement.

$$\begin{aligned}\sqrt{\frac{1}{2}}(|0\rangle+|1\rangle)_{\mathrm{S}} &\xrightarrow{x=0} \sqrt{\frac{1}{2}}(|00\rangle+|11\rangle)_{\mathrm{SM}},\\ \sqrt{\frac{1}{2}}(|0\rangle+|1\rangle)_{\mathrm{S}} &\xrightarrow{x=1} \frac{1}{2}(|00\rangle\langle 00|+|11\rangle\langle 11|)_{\mathrm{SM}}.\end{aligned} \tag{20}$$

If the superagent Wigner now measures SA in the $|\mathrm{ok}\backslash\mathrm{fail}\rangle := \{\frac{1}{\sqrt{2}}(|00\rangle \mp |11\rangle)\}$ basis to obtain the outcome w, clearly, we have $P(w=\mathrm{ok}|x=0)=0$ and $P(b=\mathrm{ok}|x=1)>0$. Therefore **I** is violated, and if we nevertheless ignore the settings, we obtain an apparent paradox with $P(w=\mathrm{ok})=0$ and $P(b=\mathrm{ok})>0$.

This highlights that in EWFSs, when assuming the universal validity of unitary quantum theory, it is natural to consider version (2) of quantum theory, which incorporates the Heisenberg cut dependence of predictions. And we have shown that this version of quantum theory, appropriately formalised, is perfectly consistent in all EWFSs. Both Wigner's original result as well as FR's result can be regarded as a cautionary note on the dangers of insisting to use version (1) of quantum theory in such EWFSs. However, FR's version arrives at the apparent contradiction while only requiring agents performing compatible measurements to reason about each other in each statement, while the apparent paradox obtained as above in Wigner's experiment requires a super-agent (Bob) to issue statements about an agent (Alice) who performs an incompatible measurement.

Comment on absoluteness of events Although FR's arguments center of 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 the draw insights on the failure of an absolute notion of events here, and its relation to the settings we have introduced in this work.

Absoluteness of observed events (AoE) entails that the predictions of the scenario can be derived from a single joint probability distribution on the observed (non-trivial) outcomes of all agents. However, the quantum predictions of the FR scenario given in Equation (16) 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 Equation (16) are equivalent to the setting-conditioned predictions of Equation (17). 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 Equa-

tion (17), 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.

Due to space considerations, we leave a detailed discussion of other no-go results for EWFSs relating to the absoluteness of events [3, 4], particularly the Local-Friendliness (LF) theorem [4], and the relation between AoE and **I** to a follow-up paper [18]. 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 this section, we address how our formalism recovers the perceived objectivity of measurement outcomes and the standard predictions of quantum theory in present real-world experiments, even though the formalism can be used in a fundamentally relational manner for general EWFSs where absoluteness of events may not hold.

In Wigner's Friend Scenarios, relational approaches typically propose to avoid paradoxes by 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 [2, 5, 9]). However, this leaves open the crucial question of how one can recover predictions of realistic quantum experiments where we perceive measurement outcomes and probabilities to be non-relational and objective. To address this question, we develop criteria to distinguish genuinely Wigner's Friend type experiments from standard quantum scenarios, when agents do not measure each other's memories/labs in a non-trivial 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 $\mathtt{S}_i$ which A_i measures together with their memory M_i constitute an idealised model of A_i's lab. We now proceed to formally define what is meant by "acting non-trivially on an agent's memory".

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 "non-trivial" or "non-standard" ways in which the operation performed by one agent in a Wigner's Friend Scenario can act on the memory/lab of another agent.

We provide such a formal definition, and then show that indeed the settings of the augmented circuit (which are the only perspectival/relational part of our general framework) can be safely dropped in standard quantum experiments while preserving the predictions of the augmented circuit.

A. Causal criteria for superagency: distinguishing standard and genuinely Wigner's Friend scenarios

We apply the notions of causal structure and directed paths introduced in Definition IV.1 and Definition IV.2 to establish a criterion distinguishing when one agent does not act as a superagent to another in an EWFS. This criterion helps differentiate parts of a general EWFS as standard quantum sectors as opposed to genuine Wigner's Friend experiments.

We term this concept a *non-superagent structure* ($n\mathcal{SA}$). The main idea is that for an agent A_j not to act non-trivially on the memory of 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[8]), 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 scenarios where agents may communicate different information based on the outcomes stored in their memories. The post-measurement state of an agent's memory M_i and measured system S_i after a measurement $\mathcal{M}^{\mathrm{A}_i}$ 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 Equation (3) and Equation (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.

[8] 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 modelled 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.

However, a scenario where Wigner performs a Hadamard operation on a Friend's brain, or undoes a Friend's measurement, involves a non-trivial 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). With these physical intuitions in mind, we provide the following technical definitions.

Definition VI.1 (Operationally equivalent EWFSs)**.** *We say that two N-agent EWFSs involving sets $\{\mathrm{A}'_1, ..., \mathrm{A}'_N\}$ and $\{\mathrm{A}_1, ..., \mathrm{A}_N\}$ of agents are operational equivalent if and only if the following hold*

- *There is one-to-one identification between the systems $\mathtt{S}' = \{\mathrm{S}'_1, ..., \mathrm{S}'_m\}$ and $\mathtt{S} = \{\mathrm{S}_1, ..., \mathrm{S}_m\}$, agents $\{\mathrm{A}'_1, ..., \mathrm{A}'_N\}$ and $\{\mathrm{A}_1, ..., \mathrm{A}_N\}$, memories $\{\mathrm{M}'_1, ..., \mathrm{M}'_N\}$ and $\{\mathrm{M}_1, ..., \mathrm{M}_N\}$, measurements $\{\mathcal{M}^{\mathrm{A}'_1}, ..., \mathcal{M}^{\mathrm{A}'_N}\}$ and $\{\mathcal{M}^{\mathrm{A}_1}, ..., \mathcal{M}^{\mathrm{A}_N}\}$, subsets $\mathtt{S}'_i \subseteq \mathtt{S}' \cup \mathtt{M}' \backslash \{\mathrm{M}'_i\}$ and $\mathtt{S}_i \subseteq \mathtt{S}' \cup \mathtt{M} \backslash \{\mathrm{M}_i\}$ of systems on which each measurement acts non-trivially, and sets of operations $\{\mathcal{E}'_1, ..., \mathcal{E}_N\}$ and $\{\mathcal{E}_1, ..., \mathcal{E}_N\}$, with outcome sets $\mathtt{O}'_i$ and $\mathtt{O}_i$ being equivalent.*
- *The augmented circuits of the two EWFS 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*

$$\begin{aligned} &P(\vec{a}_j = \vec{a_j} | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi}) \\ =&P(\vec{a}'_j = \vec{a_j} | \vec{a}'_l = \vec{a}_l, \vec{x}' = \vec{\xi}). \end{aligned} \tag{21}$$

Definition VI.2 (Non-superagent structure)**.** *We say that an EWFS involving a set $\{A_1, ..., A_N\}$ of agents respects a non-superagent structure $n\mathcal{SA}$ whose elements are pairs $(\mathrm{A}_i, \mathrm{A}_j)$ of agents, if the given EWFS is operationally equivalent to another EWFS involving a set $\{A'_1, ..., A'_N\}$ of agents, such that for all $(\mathrm{A}_i, \mathrm{A}_j) \in n\mathcal{SA}$ 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*
   - $\mathrm{A}'_i \not\prec \mathrm{A}'_j$
   - *If $\mathrm{A}'_i \prec \mathrm{A}'_j$, then $\mathcal{E}'_i$ acts trivially on the memory M'_i and $\mathrm{A}'_i \not\prec^{\mathrm{M}'_i} \mathrm{A}'_j$*

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

Note that the $n\mathcal{SA}$ of an EWFS is also a physical, objective and perspective-independent property of the protocol. Although the notion of operational equivalence refers to settings, the property is in fact independent of settings as it must hold for all settings.

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 only acts on some quantum system (that does not include Wigner's memory/lab), applying this definition, we find that $(\mathrm{F}, \mathrm{W}) \notin n\mathcal{SA}$ and $(\mathrm{W}, \mathrm{F}) \in n\mathcal{SA}$. 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 non-trivial 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 they augmented circuit as the original scenario.

Therefore $(\mathrm{W}, \mathrm{U}), (\mathrm{U}, \mathrm{W}) \in n\mathcal{SA}$ while $(\mathrm{F}, \mathrm{U}) \notin n\mathcal{SA}$ and $(\mathrm{W}, \mathrm{U}) \in n\mathcal{SA}$, and of course $(\mathrm{F}, \mathrm{F}), (\mathrm{W}, \mathrm{W}), (\mathrm{U}, \mathrm{U}) \in n\mathcal{SA}$ 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 C 2, we have $n\mathcal{SA} = \{(\mathrm{A}, \mathrm{B}), (\mathrm{B}, \mathrm{A}), (\mathrm{U}, \mathrm{W}), (\mathrm{W}, \mathrm{U}), (\mathrm{U}, \mathrm{B}), (\mathrm{B}, \mathrm{U}), (\mathrm{W}, \mathrm{A}), (\mathrm{A}, \mathrm{W}), (\mathrm{U}, \mathrm{A}), (\mathrm{W}, \mathrm{B}), (\mathrm{A}, \mathrm{A}), (\mathrm{B}, \mathrm{B}), (\mathrm{U}, \mathrm{U}), (\mathrm{W}, \mathrm{W})\}$, but this is not a standard quantum experiment as the pairs (A, U) and (B, W) don't appear in $n\mathcal{SA}$, the latter agent acts as a superagent to the former. The $n\mathcal{SA}$ of the prepare and measure version of the FR protocol is identical.

We are now ready to formally define what we mean by a standard quantum experiment.

Definition VI.3 (Standard quantum scenario)**.** *An EWFS with N agents $\{\mathrm{A}_1, ..., \mathrm{A}_N\}$ is said to correspond to a standard quantum scenario if $(\mathrm{A}_i, \mathrm{A}_j) \in n\mathcal{SA}$ for all A_i, $\mathrm{A}_j \in \{\mathrm{A}_1, ..., \mathrm{A}_N\}$. In case this holds for a subset of the first k agents $\{\mathrm{A}_1, ..., \mathrm{A}_k\}$ (who act at time steps $t_1 < ... < t_k$) in the context of a larger N-agent EWFS we call $\{\mathrm{A}_1, ..., \mathrm{A}_k\}$ a standard quantum sector of the EWFS. Otherwise we call it a non-standard 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.

Non-standardness as a signature of genuine Wigner's Friend-ness? We have provided a concrete definition of standard quantum scenarios among a general class of EWFSs, based on the operational causal structure of the scenario. An interesting question is whether scenarios that are non-standard according to this definition can be considered as exhibiting a genuinely Wigner's Friend-type aspect. One intuition in favour of this is that by definition, non-standardness 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 non-trivial manner that cannot be regarded as A_j simply "asks" A_i their outcome (for the latter can be simulated in an operationally equivalent scenario with trivial action on memory, as shown before). However, this point needs to be further investigated both at a conceptual and technical level to make conclusive statements, after all there can be different equally well-motivated criteria for "genuineness" of a non-classical resource, as the vast literature on quantum non-locality 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 from causal principles, the quantum resource associated with Wigner's Friend type no-go results that fundamentally distinguish them from existing quantum no-go results 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 formalise Heisenberg cuts in our framework) 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 (i.e., where agents do not have full quantum control over each other's labs/memories), we must show that the setting-variables can be safely dropped from the predictions and the augmented circuit when analysing such experiments. The following results formalise this intuition concretely using our definition of standard quantum scenarios. At a foundational level, this will shed light on how the perceived objectivity or non-relationalism of observed measurement events emerges within standard quantum sectors.

Theorem VI.1 (Non-action on memory and setting-independence)**.** *Consider an EWFS and a subset $\mathrm{A}_{\mathcal{K}}$ of agents therein. Suppose that $\mathrm{A}_i \notin \mathrm{A}_{\mathcal{K}}$ is another agent in the EWFS such that no agent in $\mathrm{A}_{\mathcal{K}}$ acts as a superagent to A_i i.e., $(\mathrm{A}_i, \mathrm{A}_k) \in n\mathcal{SA}\ \forall A_k \in A_{\mathcal{K}}$. Then for every partition $A_{\mathcal{K}} = \{\mathrm{A}_{j_1}, ..., \mathrm{A}_{j_p}\} \cup \{\mathrm{A}_{l_1}, ..., \mathrm{A}_{l_q}\}$ of $\mathrm{A}_{\mathcal{K}}$, 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,*

$$\begin{aligned} &P(\vec{a}_j|\vec{a}_l, (\xi_1, ..., \xi_N)) = \\ &P(\vec{a}_j|\vec{a}_l, (\xi_1, ...\xi_{i-1}, \xi_{i+1}..., \xi_N)). \end{aligned} \tag{22}$$

Recall that this expression is equivalent to the conditional independence given in Equation (10)*.*

A proof of the above theorem can be found in Appendix J.

Then, as a corollary of Theorem VI.1, we can immediately recover the full setting-independence of predictions in standard quantum experiments.

Corollary VI.1 (Full setting-independence in standard quantum theory)**.** *In any EWFS corresponding to a standard quantum scenario, every non-trivial setting-conditioned prediction, is setting indepndent. 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 $i \notin \{j_1, \ldots, j_p, l_1, \ldots, l_q\} := \mathcal{JL}$ i.e., Equation* (22) *holds for all such i. Specifically, in such standard scenarios, all non-trivial 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.*

$$\begin{aligned} &P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi}) \\ =&P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, x_i = 1, \forall i \in \mathcal{JL}) \\ :=&P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l). \end{aligned} \tag{23}$$

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 modelled (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: we can either model each measurement as acting only on the measured system and yielding a (non-trivial) 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 E, 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 Figure 6 and Figure 7 in the same appendix.

VII. DISCUSSIONS

A. Sound scientific reasoning in EWFS and analogies to classical multi-agent reasoning

The results of Section III guarantee that our formalism allows quantum agents to make predictions and rea-

son consistently about physical experiments and each other's knowledge, even when unitary quantum theory is universally valid.

Here, we illustrate a general paradigm for scientific reasoning indicated by our results and discuss their wider scope, contrasting the quantum and classical aspects. Specifically, while there are potentially genuine quantum aspects related to agents' reasoning in EWFS, by formalising EWFSs as done here, all considerations regarding the consistency of agents' reasoning can be reduced to analogous issues arising in classical multi-agent reasoning. This allows to extend the scope of our proposal to more general multi-agent scenarios by considering how analogous generalisations would work in the classical case.

Genuinely quantum and relational aspects We have seen a concrete rule for choosing settings to make predictions in universal quantum theory (Theorem IV.1), equivalent to conventional quantum predictions in the EWFS literature. This default rule models the maximum number of measurements as pure unitaries. It is implied by the choice of prediction one wishes to compute: if we wish to compute the probability of measurement outcomes a_1 and a_2, the settings x_1 and x_2 of the corresponding measurements are $x_1 = x_2 = 1$, while the settings for all other measurements are set to 0. This does not allude to agents or depend on them. However, when agents incorporate this rule while reasoning about each other's knowledge, the setting choices depend on the agent who is reasoning and the agent being reasoned about.

For example, if Alice reasons about Bob's outcome b based on her outcome a, she will use a prediction $P(b|a,\vec{x})$ choosing $x_A = x_B = 1$ and $x_C = 0$ for the measurement of Charlie. If Alice reasons about Charlie instead of Bob then the corresponding prediction $P(c|a,\vec{x})$ will have $x_A = x_C = 1$ and $x_B = 0$. Within the quantum formalism, we can interpret our settings as choices of Heisenberg cuts (this interpretation need not hold in hidden variable models for reproducing quantum predictions, see Appendix I). Then, the default rule captures subjective choices of Heisenberg cuts: each agent places themselves and the agent whose classical outcome they are reasoning about on the "classical side" of the cut, and everyone else on the "quantum side" of the cut. This allows the notion of an observed event (classical outcome) to be subjective and relative to a choice of cut.

This type of relationalism, non-absoluteness of observed events and the general setting-dependence (or Heisenberg-cut dependence) of predictions in EWFS are arguably non-classical aspects. This is because classical theories lack a non-trivial concept of Heisenberg cuts, generally have no ambiguities in how a measurement has to be fundamentally modelled.

Moreover, for one agent A_j to act as a super-agent to another agent A_i (i.e., $(\mathrm{A}_i, \mathrm{A}_j) \notin n\mathcal{SA}$), A_j must perform a non-trivial joint operation on the lab of A_i. If both agents measure in the same basis, we have already seen in section VI that $(\mathrm{A}_i, \mathrm{A}_j) \in n\mathcal{SA}$ as the operation can be simulated by acting on part of the lab, but this is not the case when A_j measures A_i's lab in a complementary superposition basis. This suggests that some notion of measurement complementarity (not typically a classical feature) may be necessary for having a non-trivial WF-like scenario. There is scope for future work on identifying and characterising non-classical resources in EWFS as discussed in Section VI A and Section VIII, which would be needed for making these observations fully rigorous.

Aspects reducible to classical issues Our formalisation of Heisenberg cuts in quantum theory as different choices of channels in a circuit implies, based on classical probability theory and logic, that one must generally condition on these choices in their reasoning unless (1) it is known that all agents in the scenario employ the same fixed choices, or (2) it has been established that these choices do not matter.

Suppose Alice and Bob who are reasoning about the output of a physical classical channel acting on an input state ρ that they previously agreed upon, but they assume different noise models N_A and N_B for the channel. Then they will generally arrive at distinct output states/probabilities, and their conclusions can seem inconsistent if they do not communicate the conditioning on the assumed noise model. One can obtain logical paradoxes akin to FR in such classical scenarios (see Appendix G for an explicit example). Note that in this example, the agents need not communicate the initial state ρ as it is common knowledge.

If Alice and Bob perform their analysis with the same noise model but one of them before and the other after lunch, there would be no inconsistencies in their conclusions even if they forget to mention whether they had eaten, as the predictions are independent of this parameter. Finally, if Alice observes that the channel's output differs from her prediction, this can falsify her assumption regarding the noise model N_A. If she believes the experiment was performed correctly, she would update her knowledge of the noise model based on a closer examination of the experimental results.

General reasoning paradigm These classical examples highlight a general paradigm for sound scientific reasoning that ensures agents do not arrive at inconsistencies:

1. Identify fixed common knowledge vs variable parameters
2. Drop redundant parameters
3. Fix a choice of remaining parameters and condition on them
4. Check if choices are falsified by observed data and update them if needed

Such careful conditioning on all relevant variable parameters ensures that the set of statements is consistent according to Definition III.5, ensuring logical and probabilistic consistency.

All these aspects are incorporated in our formalism and results. We take the protocol description of an

EWFS (Definition III.1) to be the common knowledge of all agents and construct the augmented circuit from this knowledge alone. The only variable parameters are the settings $\vec{x}$. Our results (Theorem IV.1 and Theorem VI.1) provide concrete criteria for identifying redundant settings based on the operational causal structure and non-super agent structure, both of which are objective properties of the protocol. Our default rule then fixes a choice of the remaining settings for every prediction/statement and explicitly conditions on them.

Finally, the discussions throughout our paper (see also Section VII C and Appendix I) highlight how one can falsify setting conditioned predictions through experimental data. Our default rule is based on the premise of universal validity of quantum theory, but a yet-undiscovered physical mechanism for objective collapse may falsify these predictions (obtained through the default rule) in a future experiment and one would then have to update the setting choices given by this rule (changing $x_i = 0$ to $x_i = 1$) for certain settings, if the experimental demonstration of the falsification is deemed "loophole free" and sufficiently convincing.

Efficiency of reasoning Our results guarantee that the reasoning rules we propose for quantum agents remain consistent and respect causality principles in general EWFSs. We now discuss the computational efficiency of this reasoning process, aiming to illustrate that the complexity of applying our reasoning rules is comparable to standard quantum or classical reasoning under similar assumptions.

First, computing setting-conditioned predictions in any EWFS involves applying the Born rule to a circuit with the same number of gates as the physical operations in the protocol, making it as complex as standard quantum reasoning in standard quantum scenarios. The additional rules for processing and assigning settings do not introduce any inefficiencies, as they simply involve reading off independences from the protocol's causal structure and non-super agent structure, which can be largely pre-computed.[9]

The only potential extra resource cost is in communicating and storing non-redundant settings in EWFS. However, this cost is minimal since the vector of non-redundant settings has binary entries and dimensions typically smaller than N, the number of agents. Moreover, as illustrated before, even in classical examples, under similar ambiguities in how a (classical) channel is modelled, consistency necessitates that agents communicate and keep track of the information that removes this ambiguity. Further, we have also shown that the no additional rules are needed for combining predictions/statements about measurement outcomes in quantum EWFS once the settings are accounted for, just classical probability theory and axioms of classical logic.

Finally, we have shown that reasoning in our framework respects the **C** assumption which formalised FR's C rule for setting-conditioned predictions. This means that when a setting-conditioned prediction or statement is communicated to agent A_i by another agent A_j, A_i does not need to recalculate the prediction by simulating A_j's reasoning process, and can directly inherit this knowledge. Notice that despite this lack of agent-labels on predictions, the agent-dependence and relationalism are incorporated through the fact that our default rule allows different agents to choose different settings in their reasoning.

Scope of our resolution In this work, we have taken the protocol description to be in the common knowledge of all the agents for simplicity and to illustrate the core features of Wigner's Friend scenarios that extend beyond more standard experiments. However, situating our framework within the broader reasoning paradigm described above, generalizing our consistency results to partial knowledge scenarios is entirely analogous to generalizing a classical circuit framework (for reasoning in a classical world) from full knowledge to partial knowledge scenarios.

In both classical and quantum cases, allowing agents only partial knowledge about the protocol increases the number of variable parameters that need to be conditioned on to avoid inconsistencies, as common knowledge is reduced. In the earlier classical example, if Alice and Bob had not agreed on the input state ρ to the channel, or if they are unsure whether the other person knows this state, they would need to condition on their choice of state in their communicated statements in order to avoid paradoxes.

Our quantum reasoning rules for the complete knowledge case can be straightforwardly generalised to partial knowledge scenarios using the ideas described here. We consider this generalisation to be entirely analogous to how it would be implemented in a purely classical theory of multi-agent reasoning, and therefore do not detail it here.

Status of proposed reasoning challenges With these contributions, we believe that challenges related to the logical reasoning of quantum agents in EWFS (including partial knowledge scenarios) are fully addressed to the same extent that analogous challenges are considered resolved in purely classical theories.

For example, Renner and del Rio [25] have recently proposed a challenge for agents' reasoning in quantum theory, based on the FR paradox. This challenge includes scenarios where agents have partial knowledge about a protocol. However, due to ambiguities—similar to those in the original FR paper—regarding assumptions about the Heisenberg cut, the formal modeling of measurement channels, and conditioning on agents' knowledge, we believe it is currently unclear whether the challenge is well-defined. Inconsistencies in agents' predictions can also arise in purely classical theories when similar assumptions are not carefully accounted

[9] The irrelevance of certain settings for specific predictions is derived from the protocol's structural properties alone and can naturally be included in the protocol description initially given to agents.

for, especially in scenarios where agents may have incomplete information about the protocol. As a result, we believe it remains an open question whether such a challenge admits solutions even in classical theories.

Therefore, we consider the challenges for quantum theory raised by FR's original paper [2] and the recent article [25] as effectively addressed by the framework and results proposed here, in the sense explained above (while noting the subtleties surrounding classical multi-agent inconsistencies). A further point supporting our conclusion is that we have demonstrated in detail, the consistent resolution of the FR paradox—on which such challenges are based—through our formalism (Section V for the entanglement version and Appendix F 2 for the original prepare-and-measure version). Moreover, to our knowledge, no example of a quantum EWFS (including those involving partial knowledge) exists in which our approach fails to ensure consistent reasoning or violates any fundamental physical principle, such as causality.

While these consistency issues for reasoning quantum agents appear to be resolved, several intriguing open questions remain regarding EWFSs, including both agents' reasoning and meta-physical aspects such as the absoluteness of observed events. Our work offers a solid foundation with which these questions can be formally explored. We discuss in Section VIII.

B. Interpretations of quantum theory

Our results are presented in three levels of generality that allow to distinguish different conceptual points. Firstly there is a general formalism for defining EWFSs, which allows for different rules that one may prescribe to compute probabilities $P(\vec{a}_j|\vec{a}_l, k)$ of some outcomes $\vec{a}_j$, $\vec{a}_l$ given certain parameters k describing the assumptions made about the scenario (such as its states, channels etc). Even when specialising these k parameters to the settings $\vec{x}$ of our augmented circuit in the second step, there is freedom is choosing the values of these settings. Different interpretations of quantum theory can suggest different sets of parameters k to be considered (e.g., certain interpretations such as Bohmian mechanics may require the description of additional hidden variables λ to be included in k), and can also assign different values to the settings (e.g., collapse theories would suggest $x_i = 1$ modelling the projection postulate, for all measurements, while a many-worlds type interpretation would suggest $x_i = 0$ or unitary evolution for measurements).

Our consistency results of Theorem IV.1 apply to all choices of settings in the augmented circuit, and therefore show that logical and probabilistic inconsistencies can be avoided across interpretations by applying our formalism. In Appendix I we discuss in detail how different interpretations of quantum mechanics (such as many-worlds, collapse theories, hidden variable theories, relational quantum mechanics and QBism) could apply our framework consistently to resolve all apparent FR-type paradoxes. In this sense, our general formalism is interpretation independent.

In the third step, we prescribe reasoning rules for universal quantum theory, without assuming absolute events and while maintaining causality principles. The assumption of universal quantum theory as well as adherence to causality principles excludes certain interpretations such as collapse theories or retrocausal interpretations, this part of our results is still applicable with different interpretations such as many-worlds, QBism or relational quantum mechanics as the applicability of the reasoning rules do not depend on whether the states and channels of the augmented circuit are interpreted in an ontological or epistemic manner.

Finally, we note that there are certain previous works that, at a rather high level, may appear similar to some of our results, especially with regards to our resolution of the FR paradox. This includes the consistent histories interpretation of quantum mechanics [26, 27], works on quantum decoherence [28] and another reasoning rule proposed for quantum theory to avoid FR paradoxes [10]. We discuss the relation to these previous proposals in Appendix I.

C. Physical interpretation of settings

Before discussing interpretations of settings, it is important to remind the reader of the role they already play in the existing literature of EWFSs.

A setting choice is *required* for a prediction to be made in EWFS experiments. Settings specify how a measurement is modelled (as purely quantum unitary evolution or as being associated with classical records identified by projectors), and this specification is necessary for computing any prediction in quantum theory using the Born rule[10]. We have shown that conventional predictions considered in the literature when referring to "predictions of quantum theory" in an EWFS are in fact equivalent to specific setting-conditioned predictions (Theorem IV.1). This highlights that particular setting choices are already present in EWFS arguments. This applies to FR's arguments related to consistent reasoning as well as arguments related to the absoluteness of events, as discussed in Section V.[11] Our framework makes the assumptions about the settings explicit and highlights that this information cannot generally be neglected in EWFSs.

This serves as an important prelude to the discussion of the question about physical interpretations of the settings. Due to the interpretation-independence aspect of our formalism, there isn't a unique answer

[10] Note that this is the case even in scenarios where one does not apply the projection postulate to specify the post-measurement state, see Remark V.1.

[11] Although the setting-dependence can have different consequences for the conclusions drawn from these types of arguments.

to this. The interpretation of the settings depend on the assumptions we make about the completeness of quantum theory and our beliefs about how far quantum theory might extend to the macroscopic domain.

a. **Heisenberg cuts and classical records** Let us assume that quantum theory is complete (i.e., that measurement outcomes are not described by hidden variables; see Section VII B for when this does not hold). Then, from a more physical perspective, the settings can also be associated with the Heisenberg cut. Setting $x_i = 0$ for a measurement $\mathcal{M}^{\mathrm{A}_i}$ implies that the memory M_i of the agent A_i (or more generally their lab) is inside the Heisenberg cut and is treated as a quantum system, while setting $x_i = 1$ implies that the agent's memory M_i can be treated as a classical database (in the measurement basis) that is outside the cut.

This choice of Heisenberg cut or settings can generally be dependent on the prediction being considered, or the perspective of an agent. For instance, when considering a prediction about Alice's outcome a, the measurement producing the outcome a is modelled with setting 1 as we refer to its classical record. This would also be consistent with the perspective of Alice who would perceive her own memory (which stores the classical outcome a that she observes) as being outside the Heisenberg cut and would thus assign setting 1 to her own measurement in her reasoning process. This perspective and setting choice is still consistent with the universal validity of unitary quantum theory because the perceived classicality is only relative to the basis in which the agent (here, Alice) accesses their memory. However, one agent can still model the measurements of other agents as unitary evolutions of the respective labs and therefore assign setting 0 to these, as they may not have access to these classical records and may be able to perform arbitrary quantum operations on those labs that can destroy the associated classical measurement records.

Crucially, we note that if we assume the universal validity of unitary quantum theory, the setting choice $x_i = 1$ of our framework does not correspond to a projector that was "actually performed" or an objective "wave function collapse", as no such objective account may exist. Rather, the projectors associated with $x_i = 1$ are only required for calculating the probability of the outcome of the agent A_i through the Born rule, which is necessary when one wishes to reason about said outcome.

b. **Falsifying setting choices through experiments** We have shown that every EWFS can be equivalently formulated in terms of a unique quantum circuit, which we called an augmented EWFS. It is important to note that these settings do not correspond to different choices of operations that are actually performed in the protocol. Rather, they can capture choices made by agents when scientifically reasoning about the protocol (as discussed in the previous paragraph), or the fundamental dynamics imposed by different extensions of quantum theory to the macroscopic domain. The latter becomes relevant when we do not a priori assume that unitary quantum theory is universally valid.

Different interpretations of quantum theory can generally predict different types of fundamental dynamics for macroscopic quantum systems such as agents' labs, and can thus assign different settings. Consider Wigner's original experiment, with Alice being the friend and Wigner the super-agent. Objective collapse models defy unitary quantum theory beyond a certain scale; if agents and measurement devices involved are larger than this scale, Alice's measurement would fundamentally correspond to $x_A = 1$ evolution in this case, such that the associated projectors are "actually implemented" if the physical world followed such a theory. The prediction for Wigner would then be $P(w|x_A = 1)$. In many-worlds type interpretations, one would believe that fundamental dynamics is generally unitary and assign $x_A = 0$ to Alice's measurement, computing the prediction $P(w|x_A = 0)$ for Wigner's outcome.

As it is still an open question whether or not quantum theory is universally applicable, we do not know which of these predictions will be confirmed by a hypothetical future experiment of this Wigner's Friend Scenario. If there is a collapse mechanism that breaks unitary quantum theory, then the data of the experiment could falsify the prediction $P(w|x_A = 0)$, and if unitary quantum theory prevails, then the prediction $P(w|x_A = 1)$ could be falsified by experimental data.

Different setting choices lead to different predictions, and as such, one may be inclined to believe that there is one setting choice for each measurement which is actually the "correct" one. This would be the case for models that objectively fix the Heisenberg cut for all measurements, such as the objective collapse case. This is definitely a valid interpretation of the settings, but it is not the only one. As we have seen, our formulation does not impose the absoluteness of events and permits a relational interpretation where there is no "one correct" or "one absolute" setting assignment.

This discussion highlights that even though setting choices are linked to agents' reasoning and beliefs rather than their choices of different physical operations, they do still have physical and empirical consequences in EWFSs.

c. **Time dependence of setting choices and knowledge update** The falsification of predictions goes hand-in-hand with the updating of knowledge. If one makes a prediction that is falsified by experimental data, one is forced to question the assumptions under which said prediction was made.

Given an EWFS, if a prediction made under a certain setting choice is not consistent with observed experimental outcomes in a physical realisation of the EWFS, one can update one's knowledge about the settings. Such a situation can only arise in a scenario that has setting-dependence, as setting-independent predictions would not allow us to infer anything about the setting choices.

For example, in FR, predictions concerning Alice's

outcome are independent of Bob's setting and vice-versa[12], and hence Alice's outcomes cannot be used to falsify Bob's setting choice and vice-versa. However, we have seen that at later times, when super-agents Ursula and Wigner perform their measurements, the joint predictions for the outcomes of the two super-observers are dependent on Alice's and Bob's settings. In particular, there are certain experimental outcomes which are inconsistent with Alice's and Bob's settings being 1, observing which can falsify this setting choice.

Note that updating knowledge in light of new data is also common in classical theories. However, the difference is that in such classical theories, one updates their knowledge about something that exists and to which there is an "absolute fact of the matter" regardless of whether it is measured. On the other hand, in our formalism for EWFSs, the knowledge update represents our belief about the existence/nonexistence of classical records of measurements, which may no longer be an absolute, fact of the matter and which can be updated in time in light of new data.

For example, we may safely treat all the measurements performed by experimentalists around the world as having setting 1, as we do not believe that anyone currently has access to a device that could potentially Hadamard the memories of other agents (or quantum computers that can act as measuring reasoning agents), even if some of us may believe that the world is ultimately quantum mechanical.[13] However, once we are sufficiently convinced in the future that such devices do exist, we would need to consider assigning setting 0 to some measurements in order to consistently explain the results of future experiments where such devices can "undo" or erase classical records of measurements.

This suggests yet another operational interpretation of the settings. The setting 1 case could also be interpreted as a result of unitary evolution: to do so, one simply uses the model where these incoherent measurement channels are purified via an ancillary system. Then the setting being 1 can be understood as reflecting one's belief that a superobserver does not have access to the ancillary system and thus does not have the ability to destroy the classical measurement records through a non-trivial joint operation on Alice's lab (which now includes the ancilla). Thus, the classical records of what we presently observe persist for as long as this belief holds true, but may no longer be a matter of the fact if a powerful future super-agent gains control over sufficient quantum degrees of freedom. This also complies with a decoherence type interpretation, where the world is fundamentally unitary but classical records emerge due to decohering interactions with an environment (here, ancilla). However, such interpretations typically do not consider the premise that the environment could be accessed by a future super-agent, and the resulting time-dependence with regard to the persistency of classical records.

d. **Analogy to the Maxwell's demon paradox** The broader message delivered by our resolution of the apparent FR paradox is analogous to the resolution of Maxwell's demon paradox in thermodynamics. In short, the latter is a thermodynamical thought experiment where a microscopic demon with access to knowledge of the microstates of gas molecules in a box could exploit this knowledge to apparently extract work from the gas and violate the second law of thermodynamics. However, the paradox is resolved once we are consistent in the perspective (of the microscopic demon with knowledge of microstates or a macroscopic observer without this knowledge) that is taken while calculating thermodynamic quantities such as entropies, we then find that no such violation of the second law ensues.

From the perspective of the macroscopic observer, the problem never arises in the first place and from the perspective of the demon, work must be performed in order to erase the information gained in the process which in turn ensures that the second law is not violated through the work extraction. It is only when we argue from both perspectives while ignoring the perspective that was used, that we run into an apparent paradox—the work is extracted from taking the demon's perspective and there is apparently no work performed in the process when taking the macroscopic observer's perspective.

This is analogous to the message of Corollary IV.3. The settings of our framework are related to the perspective of agents and their Heisenberg cuts as discussed above. Taking this into account allows us to make predictions that are consistent with a given perspective. We have shown that this ensures that no logical contradictions arise. It is only when we use different settings to derive a set of predictions and then ignore this setting choice through the assumption **I** that we obtain an apparent paradox.

VIII. CONCLUSIONS AND OUTLOOK

Wigner's thought experiment [1] exposes fundamental challenges in applying quantum theory to observers or agents. Recent no-go arguments (e.g., [2–4]) extend this experiment to multiple agents, suggesting that the universal validity of unitary quantum theory radically challenges our understanding of logic, scientific reasoning, causality, and the absoluteness of observed events.

In this work, we have developed a comprehensive theoretical framework for Extended Wigner's Friend Scenarios (EWFSs), enabling sound scientific reasoning without assuming absolute measurement events. The main theorems establish the framework's consistency

[12] In the entanglement version this follows from the causal structure, and in the prepare and measure version, this follows from the non-super agent structure, even though the causal structure permits a dependence due to communication.

[13] Notice that even if measurements are modelled as unitary evolutions of systems and memories/labs, as long as there are no "super-agents" who perform non-trivial operations on another agent's whole lab, we may safely treat this as the setting 1 case as we have setting-independence (c.f. Corollary VI.1).

and preservation of causality, distinguishing objective (e.g., causal structure) and subjective (e.g., predictions, agents' knowledge) aspects of EWFSs. These results can be applied to ensure global consistency of scientific reasoning in EWFSs, in a manner independent of the particular interpretation of quantum theory that one subscribes to. Further, we have discussed in Remark V.1 how our solution ensures consistent reasoning in FR-like EWFSs even when measurements are modelled as unitary evolutions of agents' labs and where the Born rule is applied to reason, without invoking the projection postulate or state update rule of quantum theory.

As a key application, our framework fully resolves all FR-type logical paradoxes in quantum theory, and more generally ensures both logical and probabilistic consistency in all EWFSs. We provided a physically motivated set of reasoning rules for quantum agents that are simultaneously consistent with the universal validity of quantum theory and classical logic applied to observed measurement outcomes, and also the fundamental relationalism of agents' perspectives in EWFSs. This demonstrates in a constructive manner that quantum theory is perfectly consistent in all EWFSs that one can construct within the theory, and that there is no threat to the ability to consistently program future quantum computers that play the role of agents. This is contrary to FR's broader claim [2] that "Quantum theory cannot consistently justify the use of itself", and we have discussed a refined interpretation of FR's arguments in light of our results, such that FR's statement about the apparent paradox does not conflict with our statements about consistently.

The key insight is to make explicit how measurements are modelled in each reasoning step, as physical predictions in EWFSs do depend on whether a quantum measurement is regarded as producing a classical outcome or as a purely quantum unitary evolution (capturing the choice of Heisenberg cut associated with the measurement). This sheds light on the core reason for apparent FR paradoxes, as arising from ignoring the choices of Heisenberg cuts used in the reasoning.

Extending beyond agent reasoning, our framework explains the emergence of objective, Heisenberg-cut-independent predictions in real-world quantum experiments. This provides a concrete view on how quantum theory can naturally accommodate the non-absoluteness of events and fundamental relationalism in general EWFSs (where agents can have arbitrary quantum control over each others' labs), while remaining consistent with objective scientific observations made so far. We outline future research directions and some broader implications of these findings for the field of EWFSs below.

a. **Local-Friendliness and non absolute events** Building on Brukner's no-go theorem for the absoluteness of observed events (AoE) [3], the Local-Friendliness (LF) no-go theorem [4] imposes strong constraints on any physical theory satisfying AoE (and other reasonable physical assumptions relating to causality and free choice), demonstrating that such theories cannot explain quantum predictions in a specific EWFS. This EWFS is similar to the entanglement version of FR, but importantly, allows super-agents to choose different measurements to perform on the agents' labs. Although we have not discussed scenarios with physical measurement choices here due to space constraints, we apply our framework to model the LF scenario and analyse their no-go theorem in forthcoming work [18], highlighting the rather distinct yet relevant role played by setting or Heisenberg-cut dependence (violation of **I**) in the LF theorem compared to FR's analysis.

In future work, it would be interesting to further explore the implications of these links between the **I** and AoE assumptions, for characterizing novel scenarios that yield no-go theorems for AoE in combination with other fundamental assumptions and quantum phenomena. These include measurement complementarity, contextuality, indefinite causal order and relativistic causality principles. We discuss some of these possible directions below.

b. **Quantum and relativistic causality in EWFSs** Causal models offer a rigorous framework to connect observed data with causal explanations, widely used in classical data-driven fields [29]. Bell's theorem exposes fundamental challenges for classical causal models in explaining quantum correlations consistently with relativity, driving the development of quantum causal modelling frameworks [30, 31] within an information-theoretic paradigm.

The LF theorem [4] is suggested to present more radical challenges for causality [6], even to existing quantum causal models as these assume AoE. AoE violations suggest that the causal structure may become subjective, affecting the notion of spacetime events and relativistic principles [6]. This raises the need for a framework for quantum causal modelling and relativistic causality that does not assume AoE.

Our framework shows that all predictions in an EWFS, possibly subjective, can be recovered within a single objective causal structure of the protocol. The augmented circuit respects this causal structure, forming an acyclic circuit that can be appropriately embedded in spacetime to preserve relativistic causality. The settings which model the Heisenberg cut choices are explicit inputs in this circuit, and the only part that can be subjective are the priors which specify the agents' choices of settings or Heisenberg cuts.

This paves a concrete pathway for developing a quantum causal modelling framework for EWFSs that does not assume AoE, which is relational, perspectival, and operational (considering agents' interventions and knowledge) and fully consistent with free choice and relativistic causality principles in space-time. This is the subject of a follow-up work.

The EWFSs in our framework correspond to protocols where agents' operations occur in a fixed, acyclic order, consistent with the time direction. Quantum theory permits so-called indefinite causal order pro-

cesses [32–34] involving quantum superpositions of the order of agents' operations, and cyclic generalisations of quantum causal models enable a description of these [23, 24, 35]. A consistent formalism for reconciling such cyclic and indefinite causal structures (which are defined through an information-theoretic notion of causality) with spacetime (which has a definite causal structure according to a relativistic definition of causality) was developed in [23, 24]. Combining techniques from our present formalism with [23, 24] offers scope for generalisation towards a unified framework for quantum agents, quantum causality and spacetime structure, providing a platform for exploring phenomena at the intersection of EWFSs, quantum processes without a definite order of operations, relativity and quantum correlations in space and time.

c. **EWFSs beyond quantum theory** Both arguments related to agents' reasoning (such as FR) and those related to AoE (such as LF) have been studied in broader theoretical contexts beyond quantum theory.

EWFS beyond quantum theory were first considered in [36], where agents memories are modelled as physical systems of a given theory. A theory-independent analogue of the projective and unitary perspectives on quantum measurements was formalised, with the latter represented by the concept of an information-preserving memory update. Using this, it was shown that agents sharing a PR-box (a post-quantum resource) [37] can encounter an apparent contradiction similar to the FR scenario. This shows that FR-type apparent paradoxes are not exclusive to quantum theory. Moreover, it was shown in [38] that any physical theory allowing for information-preserving memory updates, Bell non-classical correlations, and satisfying a locality principle would lead to violations of AoE akin to those witnessed in the quantum LF scenario [4]. Thus, the measurement problem is also not unique to quantum theory.

Future work can explore generalizing the current framework beyond quantum theory to resolve multi-agent paradoxes in post-quantum theories and illuminate the nature of AoE violations and the measurement problem therein. Many concepts and tools developed here are amenable to such generalisation.

d. **Resource-theoretic characterisation of EWF results** What are the information-theoretic resources in EWFSs responsible for FR and LF type results, and which distinguish Wigner's Friend setups from standard quantum experiments?

While we have shown that apparent Wigner's Friend paradoxes can always be resolved in quantum theory with careful specification of Heisenberg cuts, it is intriguing to characterise scenarios where this choice can be safely ignored without leading to inconsistencies. Our results Theorem IV.1 and Theorem VI.1 provide sufficient conditions for safely ignoring settings based on general structural properties (the causal and non-superagent structures), but there is scope to study the role of state and measurement dependence and derive tight necessary and sufficient conditions for scenarios that violate **I** (have setting-dependence).

Specifically, in the FR scenario, the correlations are known to be identical to those in Hardy's proof of quantum contextuality (see Appendix H for a discussion). Is contextuality a necessary feature for apparent Wigner's Friend paradoxes? In an upcoming work [39], it is shown that for a large class of multi-agent paradoxes within EWFSs in general physical theories (including FR's quantum paradox and [36]'s PR-box based paradox, but excluding Wigner's original 2-agent quantum scenario), a logical form of contextuality is a necessary property. Studying the necessary and sufficient conditions for different classes of multi-agent reasoning paradoxes in a theory-independent manner, comparing the structure of such apparent paradoxes in quantum vs more general theories, and relating them to physical principles and informational resources of the theory remain interesting future directions.

While we have provided a general resolution to all EWF quantum paradoxes, we have also highlighted that there remain interesting and less-explored questions relating to the apparent paradoxes initiated by the FR paper [2]. Our work provides a formal and consistent toolkit for exploring these other promising avenues towards understanding the structure of quantum correlations and measurements, through the study of EWFSs and Heisenberg-cut dependence of predictions.

Similar questions can also be posed for understanding the limits of AoE. In [39, 40], it is shown that Bell non-locality is not necessary for FR type paradoxes, as contextuality without Bell non-locality suffices. Is it possible to construct no-go theorems for AoE using contextuality as a resource and is this a necessary feature? More broadly, the non-super agent structure introduced here to distinguish standard quantum scenarios from genuine Wigner's Friend scenarios could be relevant for developing a resource theory of EWFSs (as discussed in Section VI).

ACKNOWLEDGMENTS

We thank Renato Renner, Victor Gitton, Yìlè Yīng, Marina Maciel Ansanelli, Joe Renes, Eric Cavalcanti and Nuriya Nurgalieva for insightful discussions. V.V. acknowledges support from an ETH Postdoctoral Fellowship. M.P.W. acknowledges funding from the Swiss National Science Foundation (AMBIZIONE Fellowship, No. PZ00P2_179914). Both authors acknowledge support from NCCR QSIT.

APPENDIX

Appendix A: Generality of the definition of EWFSs

Here we discuss the justification for the generality of Definition III.1 of an Extended Wigner's Friend Scenario that we have proposed in this paper. The idea is that Definition III.1 encompasses all finite multi-agent quantum protocols where agents' memories (in which they store the measurement outcome) are modelled 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 $\mathtt{A} = \{\mathrm{A}_1, \ldots, \mathrm{A}_N\}$ and a set of m systems $\mathtt{S} = \{\mathrm{S}_1, \ldots, \mathrm{S}_m\}$ under study and a set $\mathtt{M} := \{\mathrm{M}_1, \ldots, \mathrm{M}_N\}$ 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 $\mathtt{S}_i$ that they measure, along with their memory M_i. Each agent $\mathrm{A}_i \in \mathtt{A}$ performs a measurement $\mathcal{M}^{\mathrm{A}_i}$ on some subset $\mathtt{S}_i \subseteq \mathtt{S} \cup \mathtt{M} \backslash \{\mathrm{M}_i\}$ of systems, which can include the memories of other agents, and stores the outcome of the measurement in their memory M_i.[14] Note that $\mathtt{S}_i$ 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 bought to this form by modelling 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 modelled 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). Further, we can model each measurement $\mathcal{M}^{\mathrm{A}_i}$ as acting on the whole set $\mathtt{S}$ of systems and the set of all memories $\mathrm{M}_1, \ldots, \mathrm{M}_{i-1}$ of previous agents since any operation on a subset $\mathtt{S}_i$ 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 $\mathtt{S} \cup \mathtt{M}$) in between time steps t_i and t_{i+1}, for instance, agent A_1 may measure the system $\mathrm{S}_1 \in \mathtt{S}$ 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 of Figure 1 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 $\{\mathcal{M}^{\mathrm{A}_i}\}_{i=1}^N$ denote the measurements, one for each agent.

Furthermore, consider that an agent A_i measures a subset $\mathtt{S}_i$ of the systems through a measurement $\mathcal{M}^{\mathrm{A}_i}$. Let a_i be the random variable associated with the measurement outcome, which takes values a_i in the set $\{0, 1, ..., d_{\mathtt{S}_i} - 1\}$, where $d_{\mathtt{S}_i}$ is the Hilbert space dimension (assumed to be finite) of the system $\mathtt{S}_i$ 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. [41, Sec. 9-6]. Explicitly, we can use the projective measurement $\{\pi_{a_i}^{\mathtt{S}_i} = |a_i\rangle\langle a_i|_{\mathtt{S}_i}\}_{a_i \in \mathtt{O}_i}$, where $\{|a_i\rangle_{\mathtt{S}_i}\}_{a_i \in \mathtt{O}_i}$ forms an orthonormal basis of $\mathtt{S}_i$. While this looks like a computational basis due to the choice of outcome labels $\{0, 1, ..., d_{\mathtt{S}_i} - 1\}$, 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 III.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 III.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 generalise to ensure consistency of predictive and observational statements together. To make the distinction clear, in this section, we will explicitly write $\Sigma_{pred} := \Sigma$.

Definition B.1 (Observational statements)**.** *Observational statements are statements in the set* $\Sigma_{obs} := \{$ *"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 will denote elements of the set as* $\vec{a}_j = \vec{a}_j|_{obs}$*.*

[14] The memory M_i of an agent A_i is chosen to be a system of at least the same dimension as the system $\mathtt{S}_i$ that the agent A_i measures.

Definition B.2 (Consistency of observational statements)**.** *A set Σ_{obs} of observational statements is consistent iff $\vec{a}_j = \vec{a}_j|_{obs} \in \Sigma_{obs}$ implies that $\vec{a}_j = \neg\vec{a}_j|_{obs} \notin \Sigma_{obs}$.*

Definition B.3 (Global consistency)**.** *A set $\Sigma_{pred} \cup \Sigma_{obs}$ 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 Definition III.5 and Definition B.2 respectively,* and *additionally, $S := \vec{a}_j = \vec{a}_j|_{obs} \in \Sigma_{obs}$, then $S' \in \Sigma_{pred}$ where S' is a statement associated with a prediction $P(\vec{a}_j = \vec{a}_j|k = \hbar) > 0$.*

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 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 either 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 IV.1: 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 was 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 B.3: if these premises are satisfied, then Bob can observe $b = \mathfrak{b}$ only if $P(b = \mathfrak{b}|x_A = 0) > 0$. 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 $P(b = \mathfrak{b}|x_A = 1)$ (according to Definition B.3), and need not be globally consistent relative to $P(b = \mathfrak{b}|x_A = 0) > 0$. 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 non-trivial 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 Section 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 $\mathcal{R}^{def}$ for selecting the settings Theorem IV.1 (as also discussed in Section 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 modelled as pure unitary evolutions. Therefore, if unitary quantum theory were indeed valid universally, our consistency result of Theorem IV.1 (which only refers to 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 realisation of that EWFS. Therefore, we have no loss of generality in restricting only to predictive statements in the main text.

Consistency and agents' knowledge This distinction between predictive and observational statements emphasised here, provides a precise understanding of what does not 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 only refers to 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 $c = heads$ in one round, she would assign probability 1 to $c = heads$ in that round while Bob would assign a uniform probability

to $c = heads$ 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 $c = heads|_{obs}$. This example leads to consistent predictions since both agents would predict a uniform probability for $c = heads$ 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 $c = heads$ (and possibly other information $k = \hbar$ that she may know about the scenario). This would then correspond to a prediction $P(b = win|c = heads, k = \hbar)$. On the other hand, Bob who does not know the outcome of the coin flip but has the same background information $k = \hbar$ about the scenario, would make a prediction $P(b = win|k = \hbar)$ which can be different from $P(b = win|c = heads, k = \hbar)$. 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: 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 [5]. The FR protocol and formal statement of their no-go theorem will be reviewed in Appendix C 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 that *Statement* $\mathrm{A}^{(\mathrm{i})}$: "System S is in state $|\psi\rangle_{\mathrm{S}}$ at time t_0." Suppose furthermore that agent A knows that *Statement* $\mathrm{A}^{(\mathrm{ii})}$: "The value x is obtained by a measurement of S w.r.t. the family $\{\pi_x^{t_0}\}_{x\in\chi}$ 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 $\xi \in \chi$, then agent A can conclude that *Statement* $\mathrm{A}^{(\mathrm{iii})}$: "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 that *Statement* $\mathrm{A}^{(\mathrm{i})}$: "I am certain that agent A', upon reasoning within the same theory as the one I am using, is certain that $x = \xi$ at time t". Then agent A can conclude that *Statement* $\mathrm{A}^{(\mathrm{ii})}$: "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 that *Statement* $\mathrm{A}^{(\mathrm{i})}$: "I am certain that $x = \xi$ at time t" The agent A must necessarily deny that *Statement* $\mathrm{A}^{(\mathrm{ii})}$: "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 [5], 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 [5]. We will 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 [5], 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 [5] by formalising 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 [5] for the formal statement of this theorem in this mathematical language. We will not review the full modal logic framework here but will refer to aspects of it wherever necessary.

At a broad level, the proof proceeds by considering the FR thought experiment 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 experiment

In Section V and Figure 3 of the main text, we provided an overview of 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 thought-experiment 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 Lluis Masanes (based on a talk), and was also mentioned by Matthew Pusey in [8]. It is much simpler than FR's original (prepare and measure based) protocol, but makes the important/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 to the original prepare and measure version of FR's arguments. For the interested reader, we provide a review of the original FR thought-experiment in Appendix C 3 and apply our framework to resolve it, in Appendix F 2.

In this section we follow the notation and agent naming conventions of [5, 36]. 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 (pre-selection) at an initial time $t = 1$, intermediate local measurements by Alice and Bob on the state at time $t = 2$, a final measurement and post-selection by two superagents Ursula and Wigner at time $t = 3$.

- **Pre-selection at time $t = 1$:** An initial state $|\psi^{t=1}\rangle_{\mathrm{RS}} := \frac{1}{\sqrt{3}}(|00\rangle_{\mathrm{RS}} + |10\rangle_{\mathrm{RS}} + |11\rangle_{\mathrm{RS}})$ 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 initialised to $|0\rangle_{\mathrm{A}}$ and $|0\rangle_{\mathrm{B}}$. Hence the initial preparation (i.e., pre-selected state) on RASB is given by

$$|\psi^{t=1}\rangle_{\mathrm{RASB}} := \frac{1}{\sqrt{3}}(|0000\rangle + |1000\rangle + |1010\rangle)_{\mathrm{RASB}} \tag{C1}$$

- **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 modelled as a unitary $\mathcal{M}^{\mathrm{A}}_{unitary} := (CNOT)_{\mathrm{RA}}$ on the joint system RA which performs a CNOT operation with R as control and A as target, while Bob's measurement is similarly modelled as a unitary $\mathcal{M}^{\mathrm{B}}_{unitary} := (CNOT)_{\mathrm{SB}}$.

- **Post-selection at time $t = 3$:** The super-observers Ursula and Wigner post-select on the following final state $|\phi^{t=3}\rangle_{\mathrm{RASB}} := |\mathrm{ok}\rangle_{\mathrm{RA}} |\mathrm{ok}\rangle_{\mathrm{SB}}$, which they achieve by measuring RA and SB respectively in the basis $\{|\mathrm{ok}\rangle := \frac{1}{\sqrt{2}}(|00\rangle - |11\rangle), |\mathrm{fail}\rangle := \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)\}$ and halting when both of them obtain the outcome $u = \mathrm{ok}$, $w = \mathrm{ok}$ (where u and w denote Ursula's and Wigner's outcomes, the halting condition is checked by announcing these measurement outcomes in each run).

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

$$|\psi^{t=2}\rangle_{\mathrm{RASB}} = (\mathcal{M}^{\mathrm{A}}_{unitary} \otimes \mathcal{M}^{\mathrm{B}}_{unitary}) |\psi^{t=1}\rangle_{\mathrm{RASB}} = \frac{1}{\sqrt{3}}(|0000\rangle + |1100\rangle + |1111\rangle)_{\mathrm{RASB}}. \tag{C2}$$

The super-observers can then calculate the probability of success of the post-selection given the pre-selection and intermediate unitary evolution as

$$P(u = w = \mathrm{ok}|\psi^{t=2}) = |\langle\phi^{t=3}|\psi^{t=2}\rangle|^2 = |\langle\phi^{t=3}| (\mathcal{M}^{\mathrm{A}}_{unitary} \otimes \mathcal{M}^{\mathrm{B}}_{unitary}) |\psi^{t=1}\rangle|^2 = \frac{1}{12}. \tag{C3}$$

By this, they establish that they have a non-zero 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 $K_{\text{A}}(S)$ denotes that Agent A knows the statement S.

- Upon obtaining $u = \text{ok}$ on measuring RA, Ursula reasons using the joint state $|\psi^{t=2}\rangle$ (Equation (C2)) that Bob must have certainly obtained the outcome $b = 1$ upon measuring S, since $\langle \text{ok}|_{\text{RA}} \langle 00|_{\text{SB}} |\psi^{t=2}\rangle_{\text{RASB}} = 0$. This gives

$$K_U(u = w = \text{ok} \Rightarrow b = 1) \tag{C4}$$

- 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 $\langle 00|_{\text{RA}} \langle 11|_{\text{SB}} |\psi^{t=2}\rangle_{\text{RASB}} = 0$.

$$K_U K_B(b = 1 \Rightarrow a = 1) \tag{C5}$$

- 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 $\langle 11|_{\text{RA}} \langle \text{ok}|_{\text{SB}} |\psi^{t=2}\rangle_{\text{RASB}} = 0$. This gives

$$K_U K_B K_A(a = 1 \Rightarrow w = \text{fail}). \tag{C6}$$

As shown in [5], the above three statements can be combined using the assumption C of the FR paper (in the form of Equation (12)) and the distributive axiom D to yield the following paradoxical chain of statements. To obtain this result, the assumption C in the form of Equation (12) only needs to be used between the following pairs of agents: Alice and Bob, Alice and Wigner, Ursula and Bob, Ursula and Wigner, as other pairs of agents need not trust each other [5, 36].[15]

$$K_U(u = w = \text{ok} \Rightarrow b = 1 \Rightarrow a = 1 \Rightarrow w = \text{fail}), \tag{C7}$$

or in short $K_U(u = w = \text{ok} \Rightarrow w = \text{fail})$. 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 [5]) regarding the incompatibility of Q, U, C, D and S.

3. Prepare and measure version of the FR experiment

The scenario in [2] describes a protocol realised by agents F, $\overline{\text{F}}$, W and $\overline{\text{W}}$. Agents F and $\overline{\text{F}}$ have their own individual labs while W and $\overline{\text{W}}$ are so-called super-observers, i.e. W can perform arbitrary measurements on F and the lab of F, while $\overline{\text{W}}$ can perform arbitrary measurements on $\overline{\text{F}}$ and the lab of $\overline{\text{F}}$. The labs of F and $\overline{\text{F}}$ are completely isolated from W and $\overline{\text{W}}$ until W and $\overline{\text{W}}$ measure at the end of the protocol. It is also assumed that the labs F and $\overline{\text{F}}$ are initially in pure states. The lab systems of F and $\overline{\text{F}}$ are denoted L and $\overline{\text{L}}$ respectively. $\overline{\text{L}}$ includes everything in $\overline{\text{F}}$'s lab such as the agent $\overline{\text{F}}$ and a random generator R that they use, but excludes a spin qubit S which will start off in the lab of $\overline{\text{F}}$ 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 analysed. The n^{th} round of the protocol from the perspective of W and $\overline{\text{W}}$ is as follows:

Before time $n\!:\!00$, $\overline{\text{F}}$ 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.

$$|\text{init}\rangle_{\text{R}} := \sqrt{\frac{1}{3}} |\text{heads}\rangle_{\text{R}} + \sqrt{\frac{2}{3}} |\text{tails}\rangle_{\text{R}} . \tag{C8}$$

---

[15] In [5], 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 Equation (12) 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.

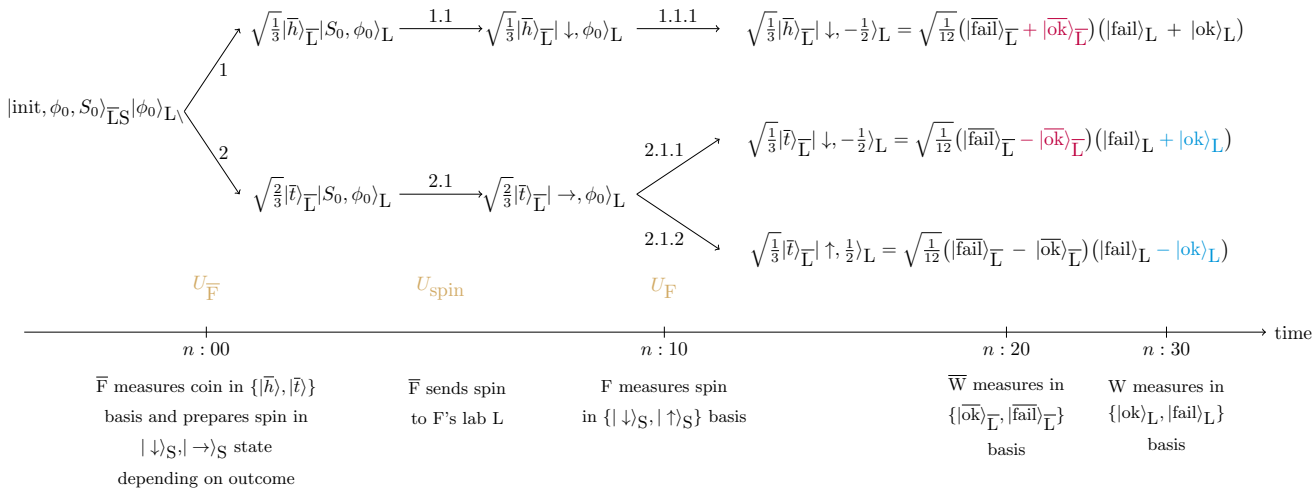

FIG. 5: The FR thought experiment expressed using branching notation. The quantum state as per a unitary description from the viewpoint of W and $\overline{\text{W}}$ at any time, is the sum of the states of each branched at said time. Arrows indicate the splitting of the amplitudes due to the actions 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.

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

$$U_{\overline{F}} : |\text{init}, \phi_0, S_0\rangle_{\overline{\text{L}}\text{S}} \rightarrow \left(\sqrt{\frac{1}{3}}\,|\overline{h}\rangle_{\overline{\text{L}}} + \sqrt{\frac{2}{3}}\,|\overline{t}\rangle_{\overline{\text{L}}}\right)|S_0\rangle_{\text{S}}\,, \tag{C9}$$

where $|\text{init}, \phi_0, S_0\rangle_{\overline{\text{L}}\text{S}} := |\text{init}\rangle_{\text{R}}\,|\phi_0\rangle_{\overline{\text{L}}\backslash}\,|S_0\rangle_{\text{S}}$ with $|\phi_0\rangle_{\overline{\text{L}}\backslash}$ the initial state of $\overline{\text{F}}$'s lab excluding the spin system and the random generator; the latter two being $|S_0\rangle_{\text{S}}$ and $|\text{init}\rangle_{\text{R}}$ respectively, where $|S_0\rangle_{\text{S}}$ is some initial state of S. The kets $|\overline{h}\rangle$, $|\overline{t}\rangle$ represent the state of $\overline{\text{F}}$'s lab (excluding the spin qubit) after her measurement of the random variable ($|\overline{h}\rangle$ in the case her measurement revealed heads while $|\overline{t}\rangle$ when she obtained tails).

Following the coin toss, $\overline{\text{F}}$ prepares a spin qubit in her lab in state $|\downarrow\rangle_{\text{S}}$ if she gets heads, while $|\rightarrow\rangle_{\text{S}}$ if she gets tails. From the perspective of F, W and $\overline{\text{W}}$, 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 $U_{\overline{F}}$, this implements the evolution

$$U_{\text{spin}}U_{\overline{F}} : |\text{init}, \phi_0, S_0\rangle_{\overline{\text{L}}\text{S}} \rightarrow \sqrt{\frac{1}{3}}\,|\overline{h}, \downarrow\rangle_{\overline{\text{L}}\text{S}} + \sqrt{\frac{2}{3}}\,|\overline{t}, \rightarrow\rangle_{\overline{\text{L}}\text{S}}\,, \tag{C10}$$

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

$$U_{\text{F}}U_{\text{spin}}U_{\overline{F}} : |\text{init}, \phi_0, S_0\rangle_{\overline{\text{L}}\text{S}}\,|\phi_0\rangle_{\text{L}\backslash} \rightarrow \sqrt{\frac{1}{3}}\Big(\,|\overline{h}\rangle_{\overline{\text{L}}}\,|\downarrow, -1/2\rangle_{\text{L}} + |\overline{t}\rangle_{\overline{\text{L}}}\,|\downarrow, -1/2\rangle_{\text{L}} + |\overline{t}\rangle_{\overline{\text{L}}}\,|\uparrow, 1/2\rangle_{\text{L}}\,\Big), \tag{C11}$$

where $|\downarrow, -1/2\rangle_{\text{L}}$ denotes the spin in the down state $|\downarrow\rangle_{\text{S}}$ and the state of the rest of lab L when F obtains measurement outcome $-1/2$; similarly for $|\downarrow, -1/2\rangle_{\text{L}}$. Between times $n:20$ and $n:30$, $\overline{\text{W}}$ measures lab $\overline{\text{L}}$ in the basis $\{|\overline{\text{ok}}\rangle_{\overline{\text{L}}}\,, |\overline{\text{fail}}\rangle_{\overline{\text{L}}}\}$, where $|\overline{\text{ok}}\rangle_{\overline{\text{L}}} := \sqrt{\frac{1}{2}}(|\overline{h}\rangle_{\overline{\text{L}}} - |\overline{t}\rangle_{\overline{\text{L}}})$, $|\overline{\text{fail}}\rangle_{\overline{\text{L}}} := \sqrt{\frac{1}{2}}(|\overline{h}\rangle_{\overline{\text{L}}} + |\overline{t}\rangle_{\overline{\text{L}}})$. After time $n = 30$, W measures lab L in the basis $\{|\text{ok}\rangle_{\text{L}}\,, |\text{fail}\rangle_{\text{L}}\}$, where $|\text{ok}\rangle_{\text{L}} := \sqrt{\frac{1}{2}}(|\downarrow, -1/2\rangle_{\text{L}} - |\uparrow, 1/2\rangle_{\text{L}})$, $|\text{fail}\rangle_{\text{L}} := \sqrt{\frac{1}{2}}(|\downarrow, -1/2\rangle_{\text{L}} + |\uparrow, 1/2\rangle_{\text{L}}$.

The final state of the labs (right hand side of Equation (C11)) in this measurement basis takes on the form

$$\sqrt{\frac{1}{12}}\left(|\overline{\text{fail}}\rangle_{\overline{\text{L}}}+|\overline{\text{ok}}\rangle_{\overline{\text{L}}}\right)\left(|\text{fail}\rangle_{\text{L}}+|\text{ok}\rangle_{\text{L}}\right)+\sqrt{\frac{1}{6}}\left(|\overline{\text{fail}}\rangle_{\overline{\text{L}}}-|\overline{\text{ok}}\rangle_{\overline{\text{L}}}\right)|\text{fail}\rangle_{\text{L}}\,. \tag{C12}$$

In Figure 5, we summarise this thought experiment using branching notation. If the outcomes of W and $\overline{\text{W}}$ are not ok, $\overline{\text{ok}}$ respectively, then the protocol is re-set and repeated. If their outcomes are ok, $\overline{\text{ok}}$, then the experiment is finished and F, $\overline{\text{F}}$, W and $\overline{\text{W}}$ reason about their ok, $\overline{\text{ok}}$ 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, $\overline{\text{ok}}$ by W and $\overline{\text{W}}$ 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 will 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 Figure 8 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, S explicitly, in addition to the assumptions U and D implicitly, as described in Appendix C 1.

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

*Theorem 1: 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 D: 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

$$\Omega=\{a_1,x_1,a_2,x_2,\ldots,a_N,x_N\}_{\{a_j,x_j\}_j}, \tag{D1}$$

where $a_j\in\perp$ for $x_j=0$ and $a_j\in\mathtt{0}_j$ 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 Figure 2) that if agents A_1, A_2, …, A_N were to perform measurements under settings $\vec{x}$ with outcomes corresponding to projectors $\pi^{\text{A}_1}_{a_1,x_1},\pi^{\text{A}_2}_{a_2,x_2},\ldots,\pi^{N}_{a_N,x_N}$ respectively, then the probability of these elementary events is

$$\begin{aligned}&P(a_1,a_2,\ldots,a_N,\vec{x})=P(a_1,a_2,\ldots,a_N|\vec{x})P(\vec{x})\\ =&\text{tr}\left[\mathcal{E}^N\left(\pi^{\text{A}_N}_{a_N,x_N}\mathcal{M}^{\text{A}_N}_{\text{unitary}}\left(\ldots\mathcal{E}^2\left(\pi^{\text{A}_2}_{a_2,x_2}\mathcal{M}^{\text{A}_2}_{\text{unitary}}\left(\mathcal{E}^1\left(\pi^{\text{A}_1}_{a_1,x_1}\mathcal{M}^{\text{A}_1}_{\text{unitary}}(\rho_0)\pi^{\text{A}_1}_{a_1,x_1}\right)\right)\pi^{\text{A}_2}_{a_2,x_2}\right)\ldots\right)\pi^{\text{A}_N}_{a_N,x_N}\right)\right]P(\vec{x}),\end{aligned} \tag{D2}$$

where ρ_0 is the initial "pre-selected" state, namely $\rho_{\text{S}_1,\ldots,\text{S}_m}\otimes|0\rangle\langle 0|^{\otimes N}$ and $P(\vec{x})$ the unconditional probability distribution over settings. The other term in Equation (D2) is the probability of outcomes $\{a_1,a_2,\ldots,a_N\}$ conditioned on setting $\vec{x}$.

The probability of the other events can be derived by marginalising over this distribution. In particular, we will be interested in the probability of events corresponding to agents $\{\text{A}_{j_1},\text{A}_{j_2},\ldots,\text{A}_{j_p}\}$ obtaining outcomes $\{a_{j_1},a_{j_2},\ldots,a_{j_p}\}$, conditioned of the setting being $\vec{x}$. This is given by

$$\begin{aligned}&P(a_{j_1},a_{j_2},\ldots,a_{j_p}|\,\vec{x})\\ &=\sum_{\substack{a_1\text{ if }x_1=1\,\&\,1\notin\mathtt{OUT}\\ a_2\text{ if }x_2=1\,\&\,2\notin\mathtt{OUT}\\ \vdots\\ a_N\text{ if }x_N=1\,\&\,N\notin\mathtt{OUT}}}\text{tr}\left[\mathcal{E}^N\left(\pi^{\text{A}_N}_{a_N,x_N}\ldots\mathcal{E}^2\left(\pi^{\text{A}_2}_{a_2,x_2}\mathcal{M}^{\text{A}_2}_{\text{unitary}}\left(\mathcal{E}^1\left(\pi^{\text{A}_1}_{a_1,x_1}\mathcal{M}^{\text{A}_1}_{\text{unitary}}(\rho_0)\pi^{\text{A}_1}_{a_1,x_1}\right)\right)\pi^{\text{A}_2}_{a_2,x_2}\right)\ldots\pi^{\text{A}_N}_{a_N,x_N}\right)\right],\end{aligned} \tag{D3}$$

$\mathtt{OUT}:=\{j_1,j_2,\ldots,j_p\}$. We have not summed over settings $x_j=0$ since the corresponding observer is modelled 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

$$\begin{aligned}\sum_j \mathrm{tr}_{\mathrm{A}}\big[\mathcal{E}(\pi_j\hat{A}\pi_j)\big] &= \sum_j \mathrm{tr}_{\mathrm{A}}\big[\mathrm{tr}_{\mathrm{B}}[U_{\mathrm{AB}}(\pi_j\hat{A}\pi_j)\otimes\rho_{\mathrm{B}}U_{\mathrm{AB}}^\dagger]\big]\\ &= \sum_j \mathrm{tr}_{\mathrm{AB}}\big[U_{\mathrm{AB}}(\pi_j\hat{A}\otimes\rho_{\mathrm{B}}\pi_j)U_{\mathrm{AB}}^\dagger\big]\\ &= \mathrm{tr}_{\mathrm{AB}}\big[(\sum_j \pi_j^2)\hat{A}\otimes\rho_{\mathrm{B}}\big] = \mathrm{tr}_{\mathrm{A}}[\hat{A}].\end{aligned} \tag{D4}$$

Applying this equality iteratively to Equation (D3) allows us to simplify it.

$$\begin{aligned}P(a_{j_1},a_{j_2},\ldots,a_{j_p}|\,\vec{x}) = &\sum_{\substack{a_1 \text{ if } x_1=1\,\&\,1\notin \mathtt{OUT}\\ a_2 \text{ if } x_2=1\,\&\,2\notin \mathtt{OUT}\\ \vdots\\ a_{K-1}\text{ if } x_{K-1}=1\,\&\,K-1\notin \mathtt{OUT}}} \mathrm{tr}\Big[**\Big],\\ **= \pi^{\mathrm{A}_K}_{a_K,1}\mathcal{E}^{K-1}&\left(\pi^{\mathrm{A}_{K-1}}_{a_{K-1},x_{K-1}}\ldots\mathcal{E}^2\left(\pi^{\mathrm{A}_2}_{a_2,x_2}\mathcal{M}^{\mathrm{A}_2}_{\mathrm{unitary}}\left(\mathcal{E}^1\left(\pi^{\mathrm{A}_1}_{a_1,x_1}\mathcal{M}^{\mathrm{A}_1}_{\mathrm{unitary}}(\rho_0)\pi^{\mathrm{A}_1}_{a_1,x_1}\right)\right)\pi^{\mathrm{A}_2}_{a_2,x_2}\right)\ldots\pi^{\mathrm{A}_{K-1}}_{a_{K-1},x_{K-1}}\right).\end{aligned} \tag{D5}$$

where $K := \max(\mathtt{OUT})$. 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 III.8), i.e. for the probability of a set of observers $\{\mathrm{A}_{j_1},\mathrm{A}_{j_2},\ldots,\mathrm{A}_{j_p}\}$ obtaining outcomes $\{a_{j_1},a_{j_2},\ldots,a_{j_p}\}$ given $\{\mathrm{A}_{l_1},\mathrm{A}_{l_2},\ldots,\mathrm{A}_{l_q}\}$ measurement outcomes $\{a_{l_1},a_{l_2},\ldots,a_{l_q}\}$ and setting $\vec{x}$. When using the definition of conditional probability, Equation (D3) and simplifying by means of Equation (D5), we obtain

$$P\left(a_{j_1},a_{j_2},\ldots,a_{j_p}\,\middle|\,a_{l_1},\ldots a_{l_q},\vec{x}\right) := \sum_{\substack{a_1 \text{ if } x_1=1\,\&\,1\notin \mathtt{OUT}\cup\mathtt{IN}\\ a_2 \text{ if } x_2=1\,\&\,2\notin \mathtt{OUT}\cup\mathtt{IN}\\ \vdots\\ a_{Q-1}\text{ if } x_{Q-1}=1\,\&\,Q-1\notin \mathtt{OUT}\cup\mathtt{IN}}} \frac{\text{Numerator}}{\sum_{\substack{a_1 \text{ if } x_1=1\\ a_2 \text{ if } x_2=1\\ \vdots\\ a_{L-1}\text{ if } x_{L-1}=1}} \text{Denominator}}, \tag{D6}$$

$$\text{Numerator} := \tag{D7}$$

$$\mathrm{tr}\left[\pi^{\mathrm{A}_Q}_{a_Q,1}\mathcal{M}^{\mathrm{A}_Q}_{\mathrm{unitary}}\left(\mathcal{E}^{Q-1}\left(\pi^{\mathrm{A}_{Q-1}}_{a_{Q-1},x_{Q-1}}\ldots\mathcal{E}^2\left(\pi^{\mathrm{A}_2}_{a_2,x_2}\mathcal{M}^{\mathrm{A}_2}_{\mathrm{unitary}}\left(\mathcal{E}^1\left(\pi^{\mathrm{A}_1}_{a_1,x_1}\mathcal{M}^{\mathrm{A}_1}_{\mathrm{unitary}}(\rho_0)\pi^{\mathrm{A}_1}_{a_1,x_1}\right)\right)\pi^{\mathrm{A}_2}_{a_2,x_2}\right)\ldots\pi^{\mathrm{A}_{Q-1}}_{a_{Q-1},x_{Q-1}}\right)\right)\right] \tag{D8}$$

$$\text{Denominator} := \tag{D9}$$

$$\mathrm{tr}\left[\pi^{\mathrm{A}_L}_{a_L,1}\mathcal{M}^{\mathrm{A}_L}_{\mathrm{unitary}}\left(\mathcal{E}^{L-1}\left(\pi^{\mathrm{A}_{L-1}}_{a_{L-1},x_{L-1}}\ldots\mathcal{E}^2\left(\pi^{\mathrm{A}_2}_{a_2,x_2}\mathcal{M}^{\mathrm{A}_2}_{\mathrm{unitary}}\left(\mathcal{E}^1\left(\pi^{\mathrm{A}_1}_{a_1,x_1}\mathcal{M}^{\mathrm{A}_1}_{\mathrm{unitary}}(\rho_0)\pi^{\mathrm{A}_1}_{a_1,x_1}\right)\right)\pi^{\mathrm{A}_2}_{a_2,x_2}\right)\ldots\pi^{\mathrm{A}_{L-1}}_{a_{L-1},x_{L-1}}\right)\right)\right] \tag{D10}$$

where $Q := \max(\mathtt{OUT}\cup\mathtt{IN})$, $\mathtt{IN} = \{l_1,l_2,\ldots,l_q\}$, $L := \max(\mathtt{IN})$. Furthermore, we have the constraint that $x_{j_1} = x_{j_2} = \ldots = x_{j_p} = x_{l_1} = x_{l_2} = \ldots = x_{l_q} = 1$, since we are reasoning about these measurement outcomes. Note that if we now sum $P\left(a_{j_1},a_{j_2},\ldots,a_{j_p}\,\middle|\,a_{l_1},\ldots a_{l_q},\vec{x}\right)$ over the elements in the set $\{a_{j_1},a_{j_2},\ldots,a_{j_p}\}$ we obtain one, and thus the distribution is normalised. 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 modelled unitarily. Also, as mentioned previously, were we have also pre-selected state ρ_0. The "post-selected state" is merely the post measurement state of the last measurement performed by observers A_j, $j \in \mathtt{OUT}$. We can easily derive an expression for our predictions (Definition III.2) using Equation (D6) and our prior $P(\vec{x})$:

$$P\left(a_{j_1},a_{j_2},\ldots,a_{j_p}\,\middle|\,a_{l_1},\ldots a_{l_q}\right) := \sum_{\vec{x}} P\left(a_{j_1},a_{j_2},\ldots,a_{j_p}\,\middle|\,a_{l_1},\ldots a_{l_q},\vec{x}\right)P(\vec{x}), \tag{D11}$$

where for consistency we have defined $P\left(a_{j_1},a_{j_2},\ldots,a_{j_p}\,\middle|\,a_{l_1},\ldots,a_{l_q},\vec{x}\right) = 0$ if there exists $k \in \{j_1,...,j_p,l_1,...,l_q\}$ s.t. $a_k \neq \perp$, & $x_k = 0$.

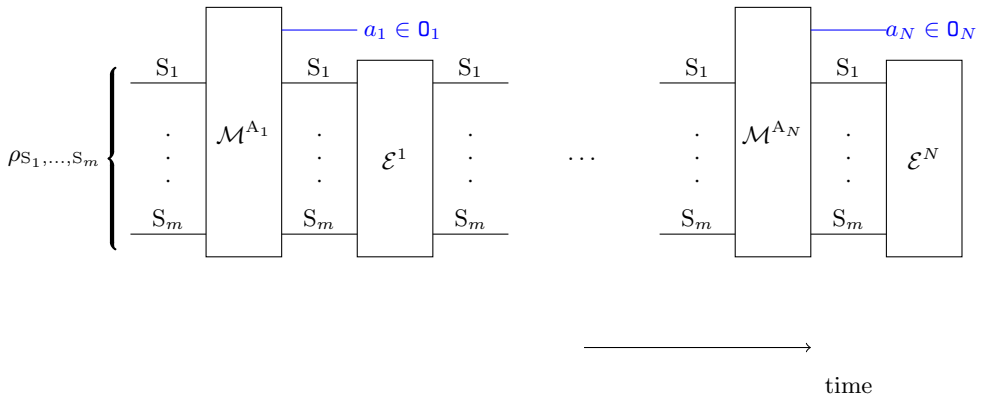

FIG. 6: A $\mathcal{C}^{sys}$-form standard quantum circuit. Here $\mathtt{O}_i$ is the set of possible non-trivial values of the outcome a_i. Each $\mathcal{M}^{\mathrm{A}_i} := \{\pi^{\mathtt{S}_i}_{a_i}\}_{a_i}$ implements a projective measurement of the subset $\mathtt{S}_i \subseteq \mathtt{S} := \{\mathrm{S}_1, ..., \mathrm{S}_m\}$ of systems.

The simplification coming from the iterative application of Equation (D3) to our conditional probability has important physical consequences, namely that the expressions Equations (D6) and (D11) readily do not depend on the channels, measurement schemes nor settings of agents in the future of when the agents in $\{\mathrm{A}_k | \, k \in \mathtt{OUT} \cup \mathtt{IN}\}$ perform their operations, i.e. in the future of the agents who measurement outcomes are being reasoned about.

Appendix E: Reduction of the augmented circuit in standard quantum scenarios

In Section VI we defined the subclass of EWFSs which we call standard quantum scenarios (Definition VI.3), 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 E.1 (Standard quantum circuit representations)**.** *Consider a quantum protocol involving N agents $\mathtt{A} := \{\mathrm{A}_1, ..., \mathrm{A}_N\}$ and m systems $\mathtt{S} := \{\mathrm{S}_1, .., \mathrm{S}_m\}$ where each agent A_i performs a projective measurement $\mathcal{M}^{\mathrm{A}_i} := \{\pi^{\mathtt{S}_i}_{a_i}\}_{a_i \in \mathtt{O}_i}$ at time t_i that acts non-trivially on a subset $\mathtt{S}_i \subseteq \mathtt{S}$ of the systems, obtaining an outcome a_i that can take values in a set $\mathtt{O}_i$, followed by a channel $\mathcal{E}_i$ that may act on all $\mathtt{S}$. 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 modelled as projectors on the systems $\mathtt{S}$ alone and in another case, all measurements can be equivalently purified to unitaries on the systems and some ancillas.*

1. **$\mathcal{C}^{sys}$-form quantum circuit (Figure 6)** *A quantum circuit acting on $\mathtt{S}$ which is defined through the composition $\mathcal{E}_N \circ \mathcal{M}^{\mathrm{A}_N} \circ ... \circ \mathcal{E}_1 \circ \mathcal{M}^{\mathrm{A}_1}$, where each operation is defined over all of $\mathtt{S}$ but it is given that each $\mathcal{M}^{\mathrm{A}_i}$ acts non-trivially on some subset $\mathtt{S}_i \subseteq \mathtt{S}$. Outcome probabilities are calculated by applying the Born rule to the circuit, wit the projective measurements $\mathcal{M}^{\mathrm{A}_i} := \{\pi^{\mathtt{S}_i}_{a_i}\}_{a_i \in \mathtt{O}_i}$ on $\mathtt{S}_i$.*

2. **$\mathcal{C}^{sys+anc}$-form quantum circuit (Figure 7)** *A quantum circuit acting on $\mathtt{S} \cup \{\mathrm{M}_1, ..., \mathrm{M}_N\}$ where M_i denotes an ancillary quantum system corresponding to the measurement $\mathcal{M}^{\mathrm{A}_i}$ whose state space is isomorphic to that of the systems $\mathtt{S}_i \subseteq \mathtt{S}$ on which the measurement acts non-trivially. It is defined through the composition $\mathcal{E}_N \circ \mathcal{M}^{\mathrm{A}_N}_{unitary} \circ ... \circ \mathcal{E}_1 \circ \mathcal{M}^{\mathrm{A}_1}_{unitary}$ where each measurement $\mathcal{M}^{\mathrm{A}_i}$ is purified to a unitary interaction $\mathcal{M}^{\mathrm{A}_i}_{unitary}$ acting on $\mathtt{S} \cup \mathrm{M}_i$ (and non-trivially on $\mathtt{S}_i \cup M_i$), where $\mathcal{M}^{\mathrm{A}_i}_{unitary}$ corresponds to the unitary that implements a coherent copy from $\mathtt{S}_i$ to M_i in the orthonormal basis given by the measurement projectors (as defined in Equation (3)). Each $\mathcal{E}_i$ acts only on $\mathtt{S}$. Outcome probabilities are calculated by measuring the ancillas M_i (using isomorphic projectors $\{\pi^{\mathrm{M}_i}_{a_i} := |a_i\rangle \langle a_i|_{\mathrm{M}_i}\}_{a_i \in \mathtt{O}_i}$) 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 Figure 6 and Figure 7.

Theorem E.1 (Recovering standard quantum circuits)**.** *If an EWFS corresponds to a standard quantum scenario (Definition VI.3), then its augmented circuit can be equivalently reduced to a standard quantum circuit, such that the same (non-trivial) predictions are obtained from the original augmented circuit, the $\mathcal{C}^{sys}$-form standard circuit or the $\mathcal{C}^{sys+anc}$-form standard circuit. Explicitly, for any disjoint sets $\vec{a}_j = (a_{j_1}, ..., a_{j_p})$ and $\vec{a}_l = (a_{l_1}, ..., a_{l_q})$ of outcomes, and any choice of settings $\vec{x} = \vec{\xi}$ such that $x_i = 1$ for all $i \in \{j_1, ..., j_p, l_1, ..., l_q\}$, we have*

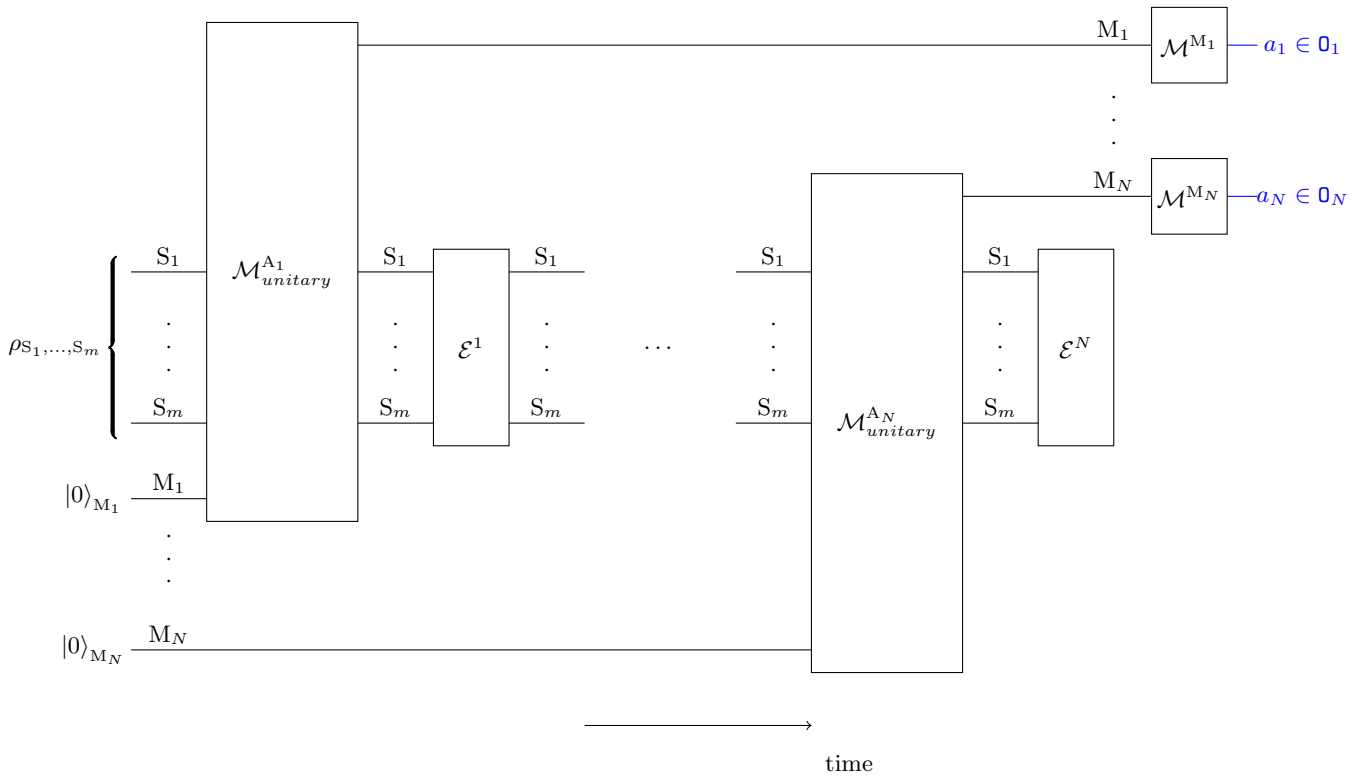

FIG. 7: A $\mathcal{C}^{sys+anc}$-form standard quantum circuit where each measurement in Figure 6 is purified to a unitary using an ancilla. Here $\mathtt{O}_i$ is the set of possible non-trivial values of the outcome a_i. The measurements $\mathcal{M}^{\mathrm{M}_i} := \{\pi^{\mathrm{M}_i}_{a_i}\}_{a_i}$ applied in the global future, implement the isomorphic projective measurement of the ancilla M_i, as do the measurements $\mathcal{M}^{\mathrm{A}_i}$ (from Figure 6) on the subset $\mathtt{S}_i \subseteq \mathtt{S} := \{\mathrm{S}_1, ..., \mathrm{S}_m\}$ of systems. It is immediate to see (and well-known) that the present circuit and that of Figure 6 are operationally equivalent.

$$P_{aug}(\vec{a}_j = \vec{a}_j | \vec{a}_j = \vec{a}_j, \vec{x} = \vec{\xi}) = P_{std}(\vec{a}'_j = \vec{a}_j | \vec{a}'_j = \vec{a}_j), \tag{E1}$$

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 $\mathcal{C}^{sys}$-form or $\mathcal{C}^{sys+anc}$-form standard quantum circuit (where no settings are involved).

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

Appendix F: Detailed analysis of the FR experiment

1. Entanglement version of the FR experiment

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 C 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 C 2, i.e., those captured by Equation (C4)-Equation (C6). Additionally, this reasoning occurs only in a round where the super-agents observe the outcomes $u = w = \mathrm{ok}$, which is associated with a probability $\frac{1}{12}$ in FR's arguments as shown in Equation (C3). We start by reproducing this probability as an explicit setting conditioning prediction in our framework. The probability of Equation (C3) is equivalent to conventional prediction $P^{FR}_{conv}(u = w = \mathrm{ok})$ (Definition III.6) 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 $\mathcal{M}^{A}_{unitary}$ and $\mathcal{M}^{B}_{unitary}$ as seen in Equation (C3)). 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 Theorem IV.1). Therefore it follows immediately that

$$P(u = w = \mathrm{ok} | \, |\psi\rangle^{t=2}) := P^{FR}_{conv}(u = w = \mathrm{ok}) = P(u = w = \mathrm{ok} | x_A = x_B = 0) \tag{F1}$$

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

Now we proceed to analysing each of the statements that agents make when the above post-selection on $u = w = \text{ok}$ succeeds. Consider the statement obtained in Equation (C4), here Ursula, upon knowing that $u = w = ok$ 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 $P_{conv}(b = 1|u = w = \text{ok}) = 1$.

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 $M^{\text{A}}_{unitary} \otimes M^{\text{B}}_{unitary}$ 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 $\pi^{\text{B}}_{1,1} = |11\rangle\langle 11|_{\text{SB}}$ 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 ($M^{\text{A}}_{unitary}$) 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 of Figure 4 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 D.

$$P(b = 1|u = w = \text{ok}, (x_1, x_2) = (0, 1)) = \frac{P(b = 1, u = w = \text{ok}|(x_1, x_2) = (0, 1))}{P(u = w = \text{ok}|(x_1, x_2) = (0, 1))}. \tag{F2}$$

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.

$$\begin{aligned} &P(b = 1, u = w = \text{ok}|(x_1, x_2) = (0, 1)) = \frac{1}{12} \\ =&|\langle\text{ok}|_{\text{RA}} \otimes \langle\text{ok}|_{\text{SB}} \left(1_{RA} \otimes \pi^{\text{B}}_{x_2=1,b=1}\right) |\psi^{t=2}\rangle|^2 \\ =&|\langle\text{ok}|_{\text{RA}} \otimes \langle\text{ok}|_{\text{SB}} \left(1_{RA} \otimes |11\rangle\langle 11|_{\text{SB}}\right) |\psi^{t=2}\rangle|^2. \end{aligned} \tag{F3}$$

$$\begin{aligned} &P(u = w = \text{ok}|(x_1, x_2) = (0, 1)) = \frac{1}{12} \\ =&\sum_{b\in\{0,1\}} |\langle\text{ok}|_{\text{RA}} \otimes \langle\text{ok}|_{\text{SB}} \left(1_{RA} \otimes \pi^{\text{B}}_{x_2=1,b}\right) |\psi^{t=2}\rangle|^2 \\ =&\sum_{b\in\{0,1\}} |\langle\text{ok}|_{\text{RA}} \otimes \langle\text{ok}|_{\text{SB}} \left(1_{RA} \otimes |bb\rangle\langle bb|_{\text{SB}}\right) |\psi^{t=2}\rangle|^2. \end{aligned} \tag{F4}$$

Having formalised 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:

$$u = w = \text{ok} \wedge (x_1, x_2) = (0, 1) \Rightarrow b = 1. \tag{F5}$$

We now proceed to the next statement of the FR reasoning, given by Equation (C5), 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 $P_{conv}(a = 1|b = 1) = 1$.

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

$$P(a = 1|b = 1, (x_1, x_2) = (1, 1)) = \frac{P(a = 1, b = 1|(x_1, x_2) = (1, 1))}{P(b = 1|(x_1, x_2) = (1, 1))} \tag{F6}$$

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

$$\begin{aligned} P(a=1,b=1|(x_1,x_2)=(1,1)) &= \frac{1}{3} \\ =&|\Big(\pi^{\mathrm{A}}_{x_1=1,a=1}\otimes\pi^{\mathrm{B}}_{x_2=1,b=1}\Big)\,|\psi^{t=2}\rangle\,|^2 \\ =&|\Big(\,|11\rangle\langle 11|_{\mathrm{RA}}\otimes|11\rangle\langle 11|_{\mathrm{SB}}\Big)\,|\psi^{t=2}\rangle\,|^2. \end{aligned} \tag{F7}$$

$$\begin{aligned} P(b=1|(x_1,x_2)=(1,1)) &= \frac{1}{3} \\ =&|\Big(1_{\mathrm{RA}}\otimes\pi^{\mathrm{B}}_{x_2=1,b=1}\Big)\,|\psi^{t=2}\rangle\,|^2 \\ =&|\Big(1_{\mathrm{RA}}\otimes|11\rangle\langle 11|_{\mathrm{SB}}\Big)\,|\psi^{t=2}\rangle\,|^2. \end{aligned} \tag{F8}$$

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

$$b=1\wedge(x_1,x_2)=(1,1)\Rightarrow a=1. \tag{F9}$$

We now turn to the third statement of the FR reasoning given in Equation (C6), 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 $P_{conv}(w=\text{fail}|a=1)=1$.

Analysing how FR calculate this probability using the Born rule, we see that this involves applying the projector $\pi^{\mathrm{A}}_{1,1}=|11\rangle\langle 11|_{\mathrm{RA}}$ to the state $|\psi^{t=2}\rangle$ which gives us Alice's setting $x_1=1$. Moreover, Bob is modelled purely unitarily here, through $M^{\mathrm{B}}_{unitary}$ 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.

$$P(w=\text{fail}|a=1,(x_1,x_2)=(1,0))=\frac{P(w=\text{fail},a=1|(x_1,x_2)=(1,0))}{P(a=1|(x_1,x_2)=(1,0))} \tag{F10}$$

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

$$\begin{aligned} P(w=\text{fail},a=1|(x_1,x_2)=(1,0)) &= \frac{2}{3} \\ =&|1_{\mathrm{RA}}\otimes\langle\text{fail}|_{\mathrm{SB}}\Big(\pi^{\mathrm{A}}_{x_1=1,a=1}\otimes 1_{\mathrm{SB}}\Big)\,|\psi^{t=2}\rangle\,|^2 \\ =&|1_{\mathrm{RA}}\otimes\langle\text{fail}|_{\mathrm{SB}}\Big(\,|11\rangle\langle 11|_{\mathrm{RA}}\otimes 1_{\mathrm{SB}}\Big)\,|\psi^{t=2}\rangle\,|^2 \end{aligned} \tag{F11}$$

$$\begin{aligned} P(a=1|(x_1,x_2)=(1,0)) &= \frac{2}{3} \\ =&|\Big(\pi^{\mathrm{A}}_{x_1=1,a=1}\otimes 1_{\mathrm{SB}}\Big)\,|\psi^{t=2}\rangle\,|^2 \\ =&|\Big(\,|11\rangle\langle 11|_{\mathrm{RA}}\otimes 1_{\mathrm{SB}}\Big)\,|\psi^{t=2}\rangle\,|^2 \end{aligned} \tag{F12}$$

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

$$a=1\wedge(x_1,x_2)=(1,0)\Rightarrow w=\text{fail}. \tag{F13}$$

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.

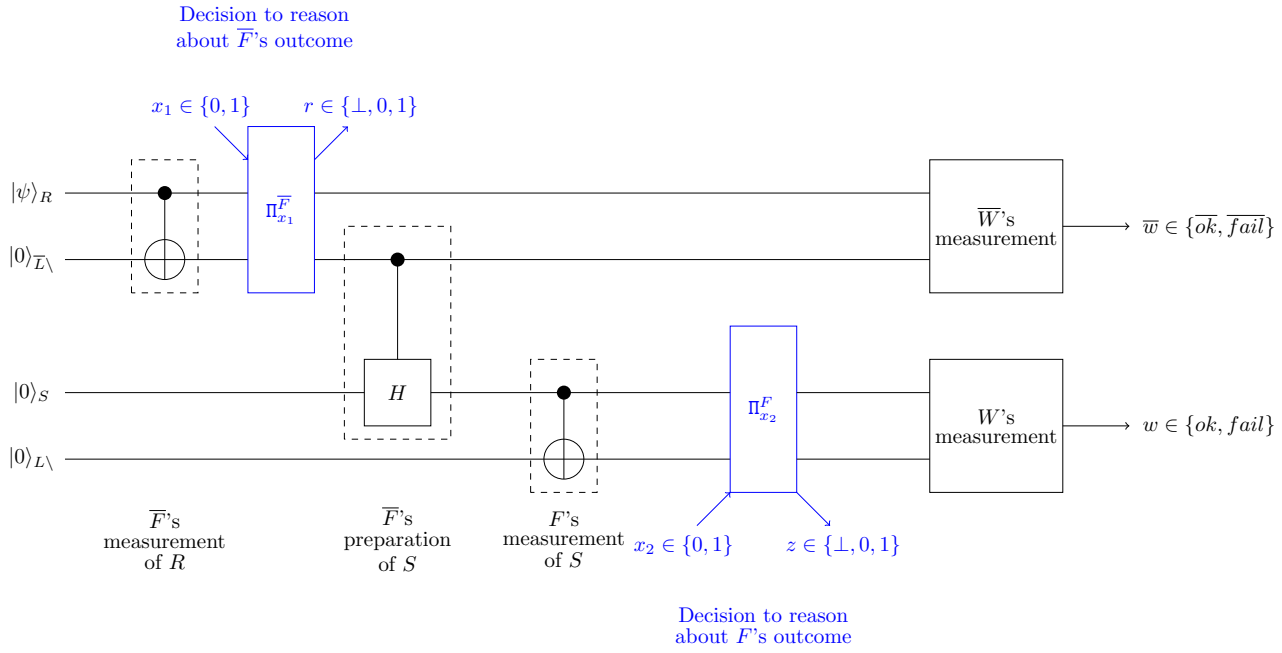

FIG. 8: Augmented circuit for the prepare and measure version of the FR scenario expressed in terms of the computational basis, as explained in the main text.

2. Prepare and measure version of the FR experiment

We have reviewed the original prepare and measure of the FR thought experiment in Appendix C 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. [2]. We will first show explicitly how original version of the FR protocol can also be modelled as an augmented circuit. Then, we will 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.

The augmented circuit The augmented circuit of the original prepare and measure version of the FR protocol is given in Figure 8. An equivalent version of this circuit is given in Figure 9, that makes the mapping to the entanglement version of the FR thought experiment more explicit.

The circuits encode the states and measurements of the original protocol in the computational basis as follows. The initial state $\sqrt{\frac{1}{3}}\,|\text{heads}\rangle_{\text{R}} + \sqrt{\frac{2}{3}}\,|\text{tails}\rangle_{\text{R}}$ of the coin in $\overline{F}$'s lab (Equation (C8)) is represented in the computational basis as $|\psi\rangle_R = \sqrt{\frac{1}{3}}\,|0\rangle_{\text{R}} + \sqrt{\frac{2}{3}}\,|1\rangle_{\text{R}}$, all other systems ($\overline{L}\setminus$, S and $L\setminus$) are initialised to $|0\rangle$. Then the measurement of $\overline{F}$ corresponds to a computational basis measurement, with the outcome $r = \textit{head}$ identified with $r = 0$ and $r = \text{tails}$ identified with $r = 1$. The preparation of S carried out by $\overline{F}$, based on the outcome of their measurement on R corresponds to a controlled Hadamard with the states $|{\downarrow}\rangle_{\text{S}}$, $|{\uparrow}\rangle_{\text{S}}$ of the FR scenarios represented as $|0\rangle_{\text{S}}$ and $|1\rangle_{\text{S}}$ 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 $\Pi^{\overline{F}}_{x_1}$ and $\Pi^{F}_{x_2}$ acting on the systems $\text{R}\overline{\text{L}}\setminus$ and $\text{SL}\setminus$ respectively are identical to the projectors $\Pi^{A}_{x_1}$ and $\Pi^{B}_{x_2}$ of Equation (14) acting on the systems RA and SB, and the final measurements of $\overline{\text{W}}$ and W are the same as the entanglement version of Appendix C 2, i.e., $\{|\text{ok}\rangle_{\text{SL}\setminus} = \frac{1}{\sqrt{2}}(|00\rangle - |11\rangle)_{\text{SL}\setminus}, |\text{fail}\rangle_{\text{SL}\setminus} := \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)_{\text{SL}\setminus}\}$ and similarly for $\{|\overline{\text{ok}}\rangle_{\text{R}\overline{\text{L}}\setminus}, |\overline{\text{fail}}\rangle_{\text{R}\overline{\text{L}}\setminus}\}$.

Statement-by-statement analysis We will now analyse their constructive proof step by step and point out what settings are required to reproduce their statements. We will 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:"k" where $G \in \{\text{W}, \overline{\text{W}}, \text{F}, \overline{\text{F}}\}$ denotes the agent G making statement k at time t. Here x_1 are for the measurements of $\overline{\text{F}}$ while x_2 are for F's. Analogously to in the entanglement scenario, we will not need to consider settings for W nor $\overline{\text{W}}$ since they are super observers. We will use italics when referring to the reasoning of the authors. What's more we will use the same notation as the authors to specify different statements, but with an additional

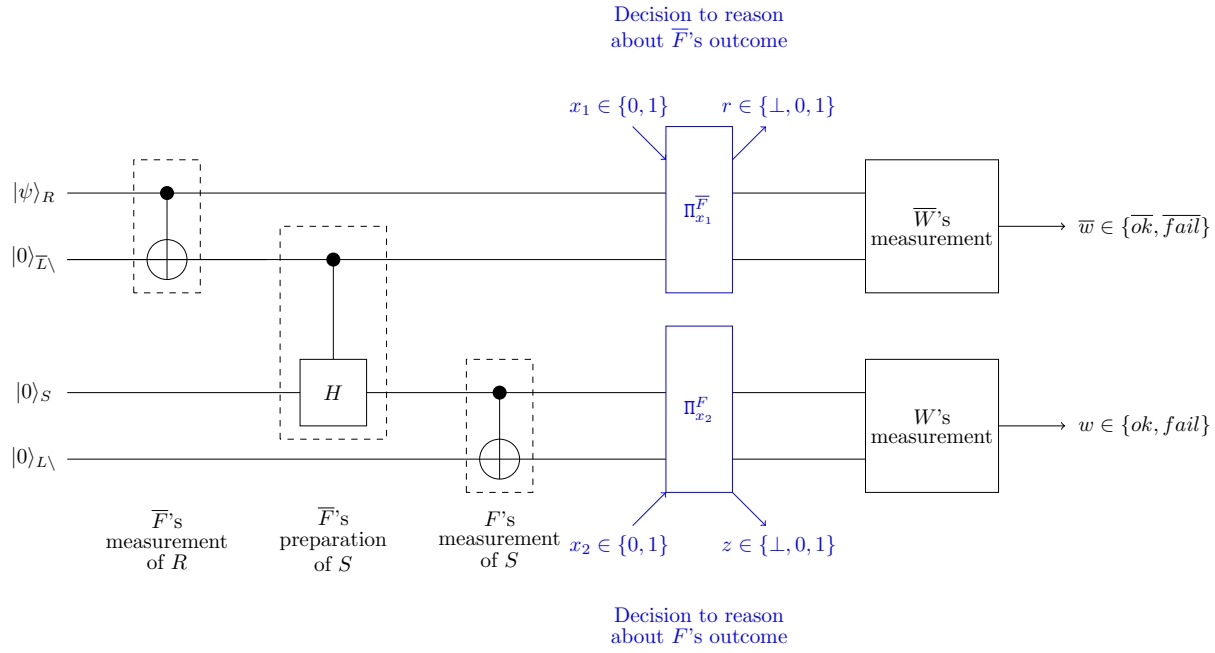

FIG. 9: Equivalent version of the augmented circuit (Figure 8) of the prepare and measure FR protocol. Note that the blue box corresponding to $\Pi^{\overline{F}}_{x_1}$ commutes with the controlled Hadamard gate since the control is on the same basis as the measurements associated with the box. Then, it is easy to verify that the joint state of $\mathrm{R}\overline{\mathrm{L}}\backslash\mathrm{SL}$ just before the blue boxes (i.e., just after F's measurement of S) is precisely the same state as $|\psi^{t=2}\rangle_{\mathrm{RASB}}$ (Equation (C2)) of the entanglement version of the FR scenario, with $\overline{\mathrm{L}}\backslash$ and $\mathrm{L}\backslash$ playing the role of A and B.

subscript indicating the (x_2, x_1) setting that said observer is using when making the statement in the corresponding augmented circuit: $G^t_{(x_1,x_2)}$:"k". The setting $(0,0)$ corresponds to the case where the agent is not reasoning about F nor $\overline{\mathrm{F}}$'s measurement outcome. Note that the reasoning of a particular observer only depends on their choice of (x_1, x_2) settings and not on those of the other observers. Furthermore, we will say $G^t_{(x_1,x_2)} = \texttt{correct}$ if G^t holds under settings (x_1, x_2), $G^t_{(x_1,x_2)} = \texttt{false}$ 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 $\overline{\mathrm{F}}$ at time $n:00$: $\overline{\mathrm{F}}^{n:00}$ *:"The value w is obtained by a measurement of L w.r.t. basis $\{\pi^{n:10}_{w=\mathrm{ok}}, \pi^{n:10}_{w=\mathrm{fail}}\}$, which is completed at time $n:31$".* Here $\pi^{n:10}_{w=\mathrm{ok}}$, $\pi^{n:10}_{w=\mathrm{fail}}$ are the projectors onto the $|\mathrm{ok}\rangle_{\mathrm{L}}$, $|fail\rangle_{\mathrm{L}}$ basis. This statement is correct according to the framework of this paper for all (x_1, x_2) i.e. $\overline{\mathrm{F}}^{n:00}_{(0,0)} = \overline{\mathrm{F}}^{n:00}_{(0,1)} = \overline{\mathrm{F}}^{n:00}_{(1,0)} = \overline{\mathrm{F}}^{n:00}_{(1,1)} = \texttt{correct}$.

Then, if $\overline{F}$ got $r =$ tails in her measurement of the coin flip, she would make the statement $\overline{\mathrm{F}}^{n:01}$*: "The spin S is in state $|\rightarrow\rangle_{\mathrm{S}}$ at time $n:10$".* In our formalism, this statement would also hold since $\overline{\mathrm{F}}$ is only reasoning about the measurement she made, we thus have $\overline{\mathrm{F}}^{n:01}_{(1,0)} = \overline{\mathrm{F}}^{n:01}_{(1,1)} = \texttt{correct}$ and $\overline{\mathrm{F}}^{n:01}_{(0,0)} = \overline{\mathrm{F}}^{n:01}_{(0,1)} = \emptyset$. Any further statements made by agent $\overline{\mathrm{F}}$ would have to be pre-selected on the spin being in state $|\rightarrow\rangle_{\mathrm{S}}$. In Figure 5 this would correspond to the selection of branch 1 and multiplying it by $\sqrt{3/2}$ to re-normalise the branch. The authors then go on to make the following claim: $\overline{\mathrm{F}}^{n:00}$ *and* $\overline{\mathrm{F}}^{n:01}$ *inserted into Q imply $w =$* fail (this is claim $\overline{\mathrm{F}}^{n:02}$ in table 3). This statement only holds if one does not take into account F's measurement at time $n:10$. In other words, $\overline{\mathrm{F}}^{n:02}_{(1,0)} = \texttt{correct}$, $\overline{\mathrm{F}}^{n:02}_{(1,1)} = \texttt{false}$. To see this, note from Figure 5 that their conclusion follows from noting the cancellation of the two blue $|\mathrm{ok}\rangle_{\mathrm{L}}$ kets when we pre-select on branch 2. However, branch 2 is further split into sub-branches 2.1.1. and 2.1.2. via F's measurement. This splitting causes the blue coloured kets to not cancel each other out from $\overline{\mathrm{F}}$'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 $\overline{\mathrm{F}}^{n:00}$ and $\overline{\mathrm{F}}^{n:01}$ imply that the probability that $w =$ fail is

$$
\begin{aligned}
&P\big(w=\text{fail}|r=\text{tails},(x_1,x_2)=(1,1)\big)\\
&=\sum_{z\in\{-1/2,1/2\}} P\big(w=\text{fail},z|r=\text{tails},(x_1,x_2)=(1,1)\big)\\
&=\frac{\sum_{z\in\{-1/2,1/2\}}\text{tr}\left[\pi^{\text{W}}_{\text{fail}}\pi^{\text{F}}_{z}\pi^{\overline{\text{F}}}_{\text{tail}}\rho_{\text{Fi}}\pi^{\overline{\text{F}}}_{\text{tail}}\pi^{\text{F}}_{z}\right]}{\sum_{\substack{z\in\{-1/2,1/2\}\\ w\in\{\text{ok,fail}\}}}\text{tr}\left[\pi^{\text{W}}_{w}\pi^{\text{F}}_{z}\pi^{\overline{\text{F}}}_{\text{tail}}\rho_{\text{Fi}}\pi^{\overline{\text{F}}}_{\text{tail}}\pi^{\text{F}}_{z}\right]}\\
&=\sum_{z\in\{-1/2,1/2\}}\frac{\text{tr}\left[\pi^{\text{W}}_{\text{fail}}\pi^{\text{F}}_{z}\pi^{\overline{\text{F}}}_{\text{tail}}\rho_{\text{Fi}}\pi^{\overline{\text{F}}}_{\text{tail}}\pi^{\text{F}}_{z}\right]}{\text{tr}\left[\pi^{\overline{\text{F}}}_{\text{tail}}\rho_{\text{Fi}}\right]}=\frac{1}{2},
\end{aligned}
\tag{F14}
$$

where $\rho_{\text{Fi}} := U_F U_{\text{spin}} U_{\overline{F}} \rho_0 (U_F U_{Spin} U_{\overline{F}})^\dagger$ with $\rho_0 := |\psi_0\rangle\langle\psi_0|$, $|\psi_0\rangle$ being the initial state (i.e. l.h.s. of Equation (C11)). The unitaries U_F, U_{spin}, $U_{\overline{F}}$ are those of the protocol (see Figure 5) 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 r.h.s. of Equation (C11). F's measurement outcome at time $n:00$ is denoted by r (with a denoting the corresponding random variable) and $\{\pi^{\text{F}}_z\}_{z\in\{-1/2,1/2\}}$ is the PVM of F's measurement at time $n{:}10$. We note that in the notation of the general framework of Section III, these projectors would be explicitly written as $\{\pi^{\text{F}}_{1,z}\}_{z\in\{-1/2,1/2\}}$ 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 Equation (F14) 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 of Figures 8 and 9.

Since this probability in Equation (F14) is less than one, we can conclude that $\overline{\text{F}}^{n:02}_{(1,1)} = \texttt{false}$. Now it is F's turn to make a statement. $\text{F}^{n:10}$*: "The value z is obtained by a measurement of spin S with respect to* $\{\pi^{\text{n:10}}_{z=-1/2}, \pi^{\text{n:10}}_{z=1/2}\}$*, 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: $\text{F}^{n:10}_{(0,1)} = \text{F}^{n:10}_{(1,1)} = \texttt{correct}$. They now further go on to state: *Suppose now that F observed* $z = 1/2$ *in round* n*. Since* $\langle\downarrow|\pi^{n:10}_{z=-1/2}|\downarrow\rangle = 1$*, it follows from Q that S was not in state* $|\downarrow\rangle_{\text{S}}$*, and hence that the random value* r *was not heads. Therefore* $\text{F}^{n:12}$*: "I am certain that* $\overline{F}$ *knows that* $r=$ *tails at time* $n:11$*".* Here F is reasoning about both her measurement outcome and that of $\overline{\text{F}}$. Therefore, by definition, this statement only makes sense when F chooses setting $(x_1, x_2) = (1, 1)$ since in the case $(x_1, x_2) = (0, 1)$, F cannot reason about $\overline{\text{F}}$'s measurement outcome since there is no classical outcome to assign to it. We therefore have $\text{F}^{n:12}_{(0,1)} = \text{F}^{n:12}_{(1,0)} = \emptyset$. Meanwhile, the following equation verifies that $\text{F}^{n:12}_{(1,1)} = \texttt{correct}$. Using Equation (D6), if we post-select on F getting $z = 1/2$ in round n, then we find that the probability that $\overline{\text{F}}$ got tails is one:

$$
P\big(r=\text{tails}|z=1/2,(x_1,x_2)=(1,1)\big)=\frac{\text{tr}[\pi^{\text{F}}_{1/2}\pi^{\overline{\text{F}}}_{\text{tail}}\rho_{\text{Fi}}\pi^{\overline{\text{F}}}_{\text{tail}}]}{\sum_{r\in\{\text{tails,heads}\}}\text{tr}[\pi^{\text{F}}_{1/2}\pi^{\overline{\text{F}}}_{r}\rho_{\text{Fi}}\pi^{\overline{\text{F}}}_{r}]}=1, \tag{F15}
$$

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

Similarly, the above equation can also be concluded from Figure 5 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 $\overline{\text{F}}$ which leads to this branch is $r =$ tails.

The authors of the FR paradox then claim: *Therefore from* $\text{F}^{n:12}$ *and invoking Q, we conclude* $\text{F}^{n:13}$*: "I am certain that* $\overline{F}$ *is certain that W will observe* $w=$ fail *at time* $n{:}31$*".* Statements $\{\text{F}^{n:13}_{(x_1,x_2)}\}_{x_1,x_2}$ do not correspond to a single statement in our framework since $\text{F}^{n:13}$ is a concatenation of two statements "upon observing $z = 1/2$, I know with certainty that $r =$ tails" made by F and "upon observing $r =$ tails, I know with certainty that $w =$ fail" made by $\overline{F}$. Note that these two statements are precisely FR's $\text{F}^{n:12}$ and $\overline{\text{F}}^{n:02}$ respectively and they are combined using the assumptions C and D to give $\text{F}^{n:13}$. Moreover, as we have seen that $\overline{\text{F}}^{n:02}_{(1,0)} = \texttt{correct}$, $\overline{\text{F}}^{n:02}_{(1,1)} = \texttt{false}$ and $\text{F}^{n:12}_{(0,1)} = \emptyset$ and $\text{F}^{n:12}_{(1,1)} = \texttt{correct}$, hence there is no common setting (x_1, x_2) for which both $\overline{\text{F}}^{n:02}$ and $\text{F}^{n:12}$ are correct. Hence there are no setting choice under which $\text{F}^{n:13}$ can be derived from $\text{F}^{n:12}$ and $\overline{\text{F}}^{n:02}$ as FR do. Agent F can always inherit both statements via the knowledge operator Equation (12), 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 $\mathrm{F}^{n:13}$, which tells us something about the outcome w =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= fail is only

$$\begin{aligned}
&P\big(w = \text{fail}|z = 1/2, (x_1, x_2) = (1, 1)\big) \\
&= \sum_{r\in\{\text{heads,tails}\}} P\big(r|z = 1/2, (x_1, x_2) = (1, 1)\big)P\big(w = \text{fail}|r, z = 1/2, (x_1, x_2) = (1, 1)\big) \\
&= P\big(w = \text{fail}|r = \text{tails}\,\&\, z = 1/2, (x_1, x_2) = (1, 1)\big) \\
&= \frac{\mathrm{tr}[\pi^{\mathrm{W}}_{\text{fail}}\pi^{\mathrm{F}}_{1/2}\pi^{\overline{\mathrm{F}}}_{\text{tails}}\rho_{\mathrm{Fi}}\pi^{\overline{\mathrm{F}}}_{\text{tails}}\pi^{\mathrm{F}}_{1/2}]}{\sum_{w\in\{\text{ok,fail}\}}\mathrm{tr}[\pi^{\mathrm{W}}_{w}\pi^{\mathrm{F}}_{1/2}\pi^{\overline{\mathrm{F}}}_{\text{tails}}\rho_{\mathrm{Fi}}\pi^{\overline{\mathrm{F}}}_{\text{tails}}\pi^{\mathrm{F}}_{1/2}]} \\
&= \frac{\mathrm{tr}[\pi^{\mathrm{W}}_{\text{fail}}\pi^{\mathrm{F}}_{1/2}\pi^{\overline{\mathrm{F}}}_{\text{tails}}\rho_{\mathrm{Fi}}\pi^{\overline{\mathrm{F}}}_{\text{tails}}\pi^{\mathrm{F}}_{1/2}]}{\mathrm{tr}[\pi^{\mathrm{F}}_{1/2}\pi^{\overline{\mathrm{F}}}_{\text{tails}}\rho_{\mathrm{Fi}}\pi^{\overline{\mathrm{F}}}_{\text{tails}}]} = \frac{1}{2} < 1,
\end{aligned} \tag{F16}$$

where we have used Equation (F15) 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 super-observers. The first statement is $\overline{\mathrm{W}}^{n:21}$*: "System R is initialised to* $|\text{init}\rangle_{\mathrm{R}}$ *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 $\overline{\mathrm{W}}^{n:21}_{(0,1)} = \overline{\mathrm{W}}^{n:21}_{(1,0)} = \overline{\mathrm{W}}^{n:21}_{(1,1)} =$ `correct`. The authors then point out that *the state* $U_{\overline{F}}|\text{init}, \phi_0, S_0\rangle$ *(r.h.s of Equation* (C10)*) is orthogonal to* $|\overline{\text{ok}}\rangle_{\overline{\mathrm{L}}}|{\downarrow}\rangle_{\mathrm{S}}$. This is indeed correct, as can be readily seen from Figure 5 by observing that $|{\downarrow}\rangle\langle{\downarrow}|_{\mathrm{S}}$ projects onto the superposition of branches 1.1.1. and 2.1.1. and that for these branches, the purple $|\overline{\text{ok}}\rangle_{\overline{\mathrm{L}}}$ terms cancel each other out. The authors then state this in the form of an expectation value of a projector, namely $\langle\text{init}|\pi^{n:00}_{(\overline{w},z)\neq(\overline{\text{ok}},-1/2)}|\text{init}\rangle = 1$. Here $\pi^{n:00}_{(\overline{w},z)\neq(\overline{\text{ok}},-1/2)} = 1 - \pi^{n:00}_{(\overline{w},z)=(\overline{\text{ok}},-1/2)}$, where $\pi^{n:00}_{(\overline{w},z)=(\overline{\text{ok}},-1/2)}$ is a Heisenberg picture projector that would first transform $|\text{init}\rangle_{\mathrm{R}}$ to $U_{\overline{F}}|\text{init}, \phi_0, S_0\rangle$ (through the appropriate isometry that appends $|\phi_0, S_0\rangle_{\overline{\mathrm{L}}\setminus\mathrm{S}}$ and performs $U_{\overline{F}}$) and then projects onto the outcomes $(\overline{w}, z) = (\overline{\text{ok}}, -1/2)$, as $|\overline{\text{ok}}\rangle\langle\overline{\text{ok}}|_{\overline{\mathrm{L}}} \otimes |{\downarrow}\rangle\langle{\downarrow}|_{\mathrm{S}} \otimes 1_{\mathrm{L}\setminus}$ i.e, $\pi^{n:00}_{(\overline{w},z)\neq(\overline{\text{ok}},-1/2)}$ is the Heisenberg projector onto the complement of outcomes $(\overline{w}, z) = (\overline{\text{ok}}, -1/2)$. They then claim *Agent* $\overline{W}$*, who uses Q, can hence be certain that* $(\overline{w}, z) \neq (\overline{\text{ok}}, -1/2)$ and that this implies (when $\overline{w}$=$\overline{\text{ok}}$) the statement $\overline{\mathrm{W}}^{n:22}$*: "I am certain that F knows that* $z = 1/2$ *at time* $n:11$*"*. Now the authors are allowing $\overline{\mathrm{W}}$ to take into account the measurement outcome of F in their reasoning but not the measurement of $\overline{\mathrm{F}}$. In other words, they are using settings $(0, 1)$ and thus assigning

$$\begin{aligned}
&P\big(z = 1/2|\overline{w} = \overline{\text{ok}}, (x_1, x_2) = (0, 1)\big) \\
&= \sum_{w\in\{\text{ok,fail}\}} P\big(z = 1/2, w|\overline{w} = \overline{\text{ok}}, (x_1, x_2) = (0, 1)\big) \\
&= \sum_{w\in\{\text{ok,fail}\}} \frac{\mathrm{tr}[\pi^{\mathrm{W}}_{w}\pi^{\overline{\mathrm{W}}}_{\overline{\text{ok}}}\pi^{\mathrm{F}}_{1/2}\rho_{\mathrm{Fi}}\pi^{\mathrm{F}}_{1/2}]}{\sum_{\substack{z\in\{-1/2,1/2\}\\ w'\in\{\text{ok,fail}\}}}\mathrm{tr}[\pi^{\mathrm{W}}_{w'}\pi^{\overline{\mathrm{W}}}_{\overline{\text{ok}}}\pi^{\mathrm{F}}_{z}\rho_{\mathrm{Fi}}\pi^{\mathrm{F}}_{z}]} \\
&= \frac{\mathrm{tr}[\pi^{\overline{\mathrm{W}}}_{\overline{\text{ok}}}\pi^{\mathrm{F}}_{1/2}\rho_{\mathrm{Fi}}\pi^{\mathrm{F}}_{1/2}]}{\sum_{z\in\{-1/2,1/2\}}\mathrm{tr}[\pi^{\overline{\mathrm{W}}}_{\overline{\text{ok}}}\pi^{\mathrm{F}}_{z}\rho_{\mathrm{Fi}}\pi^{\mathrm{F}}_{z}]} = 1,
\end{aligned} \tag{F17}$$

where the last line follows from observing that in Figure 5 we have that $\mathrm{tr}[\pi^{\overline{\mathrm{W}}}_{\overline{\text{ok}}}\pi^{\mathrm{F}}_{-1/2}\rho_{\mathrm{Fi}}\pi^{\mathrm{F}}_{-1/2}] = 0$ and thus the r.h.s. of above is one, in accordance with what the authors claim. We thus have $\overline{\mathrm{W}}^{n:22}_{(0,1)} =$ `correct`.

While they are taking into account F's measurement while reasoning, they are not taking into account $\overline{\mathrm{F}}$'s measurement when reasoning. We can check that the statement $\overline{\mathrm{W}}^{n:22}_{(1,1)} =$ `false` since the following probability,

which takes into account both $\overline{\mathrm{F}}$'s and F's measurements, is strictly less than one:

$$\begin{aligned}
&P\big(z=1/2|\overline{w}=\overline{\mathrm{ok}},(x_1,x_2)=(1,1)\big)\\
&=\sum_{\substack{w\in\{\mathrm{ok},\mathrm{fail}\}\\ r\in\{\mathrm{heads},\mathrm{tails}\}}} P\big(z=1/2,w,r|\overline{w}=\overline{\mathrm{ok}},(x_1,x_2)=(1,1)\big)\\
&=\frac{\sum_{\substack{w\in\{\mathrm{ok},\mathrm{fail}\}\\ r\in\{\mathrm{heads},\mathrm{tails}\}}}\mathrm{tr}[\pi^{\mathrm{W}}_{w}\pi^{\overline{\mathrm{W}}}_{\overline{\mathrm{ok}}}\pi^{\mathrm{F}}_{1/2}\pi^{\overline{\mathrm{F}}}_{r}\rho_{\mathrm{Fi}}\pi^{\overline{\mathrm{F}}}_{r}\pi^{\mathrm{F}}_{1/2}]}{\sum_{\substack{z\in\{-1/2,1/2\}\\ w\in\{\mathrm{ok},\mathrm{fail}\}\\ r\in\{\mathrm{heads},\mathrm{tails}\}}}\mathrm{tr}[\pi^{\mathrm{W}}_{w}\pi^{\overline{\mathrm{W}}}_{\overline{\mathrm{ok}}}\pi^{\mathrm{F}}_{z}\pi^{\overline{\mathrm{F}}}_{r}\rho_{\mathrm{Fi}}\pi^{\overline{\mathrm{F}}}_{r}\pi^{\mathrm{F}}_{z}]}\\
&=\frac{\sum_{r\in\{\mathrm{heads},\mathrm{tails}\}}\mathrm{tr}[\pi^{\overline{\mathrm{W}}}_{\overline{\mathrm{ok}}}\pi^{\mathrm{F}}_{1/2}\pi^{\overline{\mathrm{F}}}_{r}\rho_{\mathrm{Fi}}\pi^{\overline{\mathrm{F}}}_{r}\pi^{\mathrm{F}}_{1/2}]}{\sum_{\substack{z\in\{-1/2,1/2\}\\ r\in\{\mathrm{heads},\mathrm{tails}\}}}\mathrm{tr}[\pi^{\overline{\mathrm{W}}}_{\overline{\mathrm{ok}}}\pi^{\mathrm{F}}_{z}\pi^{\overline{\mathrm{F}}}_{r}\rho_{\mathrm{Fi}}\pi^{\overline{\mathrm{F}}}_{r}\pi^{\mathrm{F}}_{z}]}=\frac{1}{3}<1,
\end{aligned}\tag{F18}$$

Furthermore $\overline{\mathrm{W}}^{n:22}_{(1,0)}=\emptyset$ since the statement $\overline{\mathrm{W}}^{n:22}$ is about F's measurement outcome. Before we move on, observe that Equations (F17) and (F18) 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 $\overline{\mathrm{F}}$'s setting $\underline{x_1}$ given the knowledge that the post-selection on $\overline{w}$ succeeded. This means that the probability $P\big(z=1/2|\overline{w}=\overline{\mathrm{ok}}\big)$ 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 $\overline{W}$ announces $\overline{w}$, agent W can be certain about $\overline{W}$'s knowledge, which justifies statement* $\mathrm{W}^{n:26}$*, where* $\mathrm{W}^{n:26}$ *is (assuming $\overline{W}$ announces $\overline{w}=\overline{ok}$ at time $n:21$.) "I am certain that $\overline{W}$ knows that $\overline{w}=\overline{ok}$ at time $n:21$."*. We agree that W can be sure of any correct announcement made by $\overline{\mathrm{W}}$, irrespective of their choice of (x_1,x_2) settings. Thus $\mathrm{W}^{n:26}_{(0,1)}=\mathrm{W}^{n:26}_{(1,0)}=\mathrm{W}^{n:26}_{(1,1)}=\texttt{correct}$.

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

$$\langle\psi_{\mathrm{Fi}}|\,\pi_{(\overline{w},w)=(\overline{\mathrm{ok}},\mathrm{ok})}\,|\psi_{\mathrm{Fi}}\rangle=\frac{1}{12},\tag{F19}$$

where $|\psi_{\mathrm{Fi}}\rangle$ is the final state given on the r.h.s. of Equation (C11), $\pi_{(\overline{w},w)=(\overline{\mathrm{ok}},\mathrm{ok})}=|\overline{\mathrm{ok}}\rangle\langle\overline{\mathrm{ok}}|_{\overline{\mathrm{L}}}\otimes|\mathrm{ok}\rangle\langle\mathrm{ok}|_{\mathrm{L}}$. Note that this can also be seen as the case with $(x_1,x_2)=(0,0)$, as we will see later in Equation (F23). The authors then claim $\mathrm{W}^{n:00}$*: "I am certain that there exists a round n in which the halting condition at time $n:40$ is satisfied."* Noting that

$$\begin{aligned}
&P\big(\overline{w}=\overline{\mathrm{ok}},w=\mathrm{ok}|(x_1,x_2)=(1,1)\big)\\
&=\sum_{\substack{z\in\{-1/2,1/2\}\\ r\in\{\mathrm{heads},\mathrm{tails}\}}}\mathrm{tr}[\pi^{\mathrm{W}}_{\mathrm{ok}}\pi^{\overline{\mathrm{W}}}_{\overline{\mathrm{ok}}}\pi^{\mathrm{F}}_{z}\pi^{\overline{\mathrm{F}}}_{r}\rho_{\mathrm{Fi}}\pi^{\overline{\mathrm{F}}}_{r}\pi^{\mathrm{F}}_{z}]\\
&=\sum_{\substack{z\in\{-1/2,1/2\}\\ r\in\{\mathrm{heads},\mathrm{tails}\}}}\big|\langle\overline{\mathrm{ok}}|\langle\mathrm{ok}|\,\pi^{\mathrm{F}}_{z}U_{F}U_{Spin}\pi^{\overline{\mathrm{F}}}_{r}U_{\overline{F}}\,|\mathrm{init},\phi_0,S_0\rangle_{\overline{\mathrm{LS}}}\,|\phi_0\rangle_{\mathrm{L}}\big|=\frac{5}{6}
\end{aligned}\tag{F20}$$

is positive, we conclude that $\mathrm{W}^{n:00}_{(1,1)}=\texttt{correct}$. Similarly, we observe that

$$P\big(\overline{w}=\overline{\mathrm{ok}},w=\mathrm{ok}|(x_1,x_2)=(0,1)\big)=\sum_{z\in\{-1/2,1/2\}}\mathrm{tr}[\pi^{\mathrm{W}}_{\mathrm{ok}}\pi^{\overline{\mathrm{W}}}_{\overline{\mathrm{ok}}}\pi^{\mathrm{F}}_{z}\rho_{\mathrm{Fi}}\pi^{\mathrm{F}}_{z}]=\frac{1}{2}\tag{F21}$$

in positive, thus $\mathrm{W}^{n:00}_{(0,1)}=\texttt{correct}$. Likewise, there exits $r\in\{\mathrm{tails},\mathrm{heads}\}$ such that

$$P\big(\overline{w}=\overline{\mathrm{ok}},w=\mathrm{ok}|(x_1,x_2)=(1,0)\big)=\sum_{r\in\{\mathrm{heads},\mathrm{tails}\}}\mathrm{tr}[\pi^{\mathrm{W}}_{\mathrm{ok}}\pi^{\overline{\mathrm{W}}}_{\overline{\mathrm{ok}}}\pi^{\overline{\mathrm{F}}}_{r}\rho_{\mathrm{Fi}}\pi^{\overline{\mathrm{F}}}_{r}]=\frac{1}{2}\tag{F22}$$

is positive thus $\mathrm{W}^{n:00}_{(1,0)}=\texttt{correct}$. Finally,

$$P\big(\overline{w}=\overline{\mathrm{ok}},w=\mathrm{ok}|(x_1,x_2)=(0,0)\big)=\mathrm{tr}[\pi^{\mathrm{W}}_{\mathrm{ok}}\pi^{\overline{\mathrm{W}}}_{\overline{\mathrm{ok}}}\rho_{\mathrm{Fi}}]=\frac{1}{12},\tag{F23}$$

hence $\overline{\mathrm{W}}^{n:00}_{(0,0)} = \texttt{correct}$. Therefore we conclude from Equations (F20) to (F23) that while the statement $\overline{\mathrm{W}}^{n:00}_{(x_1,x_2)}$ 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* $\overline{\mathrm{F}}$*'s statement* $\overline{\mathrm{F}}^{n:02}$ *into* $\mathrm{F}^{n:12}$*, obtaining statement* $\mathrm{F}^{n:13}$. As we pointed out above, said statement only holds under certain settings. The authors then go on to say *By virtue of [assumption]* C*, she [agent F] may then conclude that statement* $\mathrm{F}^{n:14}$ *holds, too.* This statement is $\mathrm{F}^{n:14}$*: "I am certain that W will observe* $w=$ fail *at time* $n:31$*"*. This statement is merely the same as $\mathrm{F}^{n:13}$ but now with the difference that F *herself* is certain that W will observe $w=$ fail at time $n:31$ rather that F being merely certain that $\overline{\mathrm{F}}$ is certain. There are no settings F can select for which this statement is true since

$$P\big(w = \text{fail}|z = 1/2, (x_1, x_2) = (x_1, 1)\big) < 1 \tag{F24}$$

for all settings $x_1 \in \{0, 1\}$. Moreover, F cannot inherit the statement "I am certain that W will observe $w=$ fail at time $n:31$" via the transfer of knowledge operator since this statement does not make reference to agent $\overline{\mathrm{F}}$ and thus 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 $\mathrm{F}^{n:13}$).

The remaining statements made by W and $\overline{\mathrm{W}}$ 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 Figure 8 and an equivalent version of this circuit (that makes the mapping to the entanglement formulation of the scenario more evident) is given in Figure 9.

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 Theorem IV.1 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 experiment

It Appendix F 1, we mapped each FR probability ($P_{conv}(b = 1|u = w = \text{ok}) = 1$, $P_{conv}(a = 1|b = 1) = 1$ and $P_{conv}(w = \text{fail}|a = 1) = 1$) 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 not 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 F 1 that $P_{conv}(b = 1|u = w = \text{ok}) = P(b = 1|u = w = \text{ok}, (x_1, x_2) = (0, 1)) = 1$. 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 non-trivial projector on Bob's side in its evaluation.

$$\begin{aligned} & P(b = 1, u = w = \text{ok}|(x_1, x_2) = (1, 1)) = \frac{1}{12} \\ = & \sum_{a\in\{0,1\}} |\,\langle \text{ok}|_{\mathrm{RA}} \otimes \langle \text{ok}|_{\mathrm{SB}} \left(\pi^{\mathrm{A}}_{x_1=1,a} \otimes \pi^{\mathrm{B}}_{x_2=1,b=1}\right) |\psi^{t=2}\rangle\,|^2 \\ = & \sum_{a\in\{0,1\}} |\,\langle \text{ok}|_{\mathrm{RA}} \otimes \langle \text{ok}|_{\mathrm{SB}} \Big(|aa\rangle\langle aa|_{\mathrm{RA}} \otimes |11\rangle\langle 11|_{\mathrm{SB}} \Big) |\psi^{t=2}\rangle\,|^2 \end{aligned} \tag{F25}$$

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

TABLE II: Table 3 of [2] 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 summarised in the modal logic language in column 4. The corresponding explicit version of the same implications that take the additional implicit assumptions into account are 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 modelling 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 of Figure 8, these encode the additional implicit assumptions.

$$\begin{aligned} &P(u = w = \text{ok}|(x_1, x_2) = (1,1)) = \frac{1}{4} \\ &= \sum_{a,b \in \{0,1\}} |\left\langle \text{ok}\right|_{\text{RA}} \otimes \left\langle \text{ok}\right|_{\text{SB}} \left(\pi^{\text{A}}_{x_1=1,a} \otimes \pi^{\text{B}}_{x_2=1,b}\right) |\psi^{t=2}\rangle\,|^2 \\ &= \sum_{a.b \in \{0,1\}} |\left\langle \text{ok}\right|_{\text{RA}} \otimes \left\langle \text{ok}\right|_{\text{SB}} \Big(|aa\rangle \left\langle aa\right|_{\text{RA}} \otimes |bb\rangle \left\langle bb\right|_{\text{SB}} \Big) |\psi^{t=2}\rangle\,|^2 \end{aligned} \tag{F26}$$

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

$$P(b = 1|u = w = \text{ok}, (x_1, x_2) = (1,1)) = \frac{1}{3} \neq 1 \tag{F27}$$

The FR probability $P_{conv}(a = 1|b = 1) = 1$ 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 $P_{conv}(w = \text{fail}|a = 1) = 1$ 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.

$$\begin{aligned} &P(a = 1, w = \text{fail}|(x_1, x_2) = (1,1)) = \frac{1}{3} \\ &= \sum_{b \in \{0,1\}} |1_{\text{RA}} \otimes \left\langle \text{fail}\right|_{\text{SB}} \left(\pi^{\text{A}}_{x_1=1,a=1} \otimes \pi^{\text{B}}_{x_2=1,b}\right) |\psi^{t=2}\rangle\,|^2 \\ &= \sum_{b \in \{0,1\}} |1_{\text{RA}} \otimes \left\langle \text{fail}\right|_{\text{SB}} \Big(|11\rangle \left\langle 11\right|_{\text{RA}} \otimes |bb\rangle \left\langle bb\right|_{\text{SB}} \Big) |\psi^{t=2}\rangle\,|^2 \end{aligned} \tag{F28}$$

$$\begin{aligned} & P(a=1|(x_1,x_2)=(1,1))=\frac{2}{3} \\ &= \sum_{b\in\{0,1\}} |\Big(\pi^{\mathrm{A}}_{x_1=1,a=1}\otimes\pi^{\mathrm{B}}_{x_2=1,b}\Big)\,|\psi^{t=2}\rangle\,|^2 \\ &= \sum_{b\in\{0,1\}} |\Big(\,|11\rangle\langle 11|_{\mathrm{RA}}\otimes|bb\rangle\langle bb|_{\mathrm{SB}}\,\Big)\,|\psi^{t=2}\rangle\,|^2 \end{aligned} \tag{F29}$$

This gives us

$$P(w=\text{fail}|a=1,(x_1,x_2)=(1,1))=\frac{1}{2}\neq 1. \tag{F30}$$

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 IV.2, which shows that in any EWFS, any potential inconsistencies can only arise 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, 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 G: 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 signalling under post-selection. Further, 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 will thus refer to the **I** applied to the settings of this classical example as $\mathbf{I}_C$ to distinguish it from the version of the assumption applied to the settings of our augmented circuit that model Heisenberg cuts.

The augmented circuit Figure 4 of the entanglement version of the FR protocol can be represented as shown in Figure 10(a). Consider now the circuit of Figure 10(b), which has the same configuration as the circuit of Figure 10(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 non-signalling 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 signalling 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 signalling 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 $u=a\cdot k_U$, $a=k_U=1$. Further, 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 post-selections 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 [14] Healey pointed out that whether or not Bob's state is collapsed influences the probabilities of Alice's outcome in the FR scenario due to the post-selection. Denying this dependence is what [14] refers to as the assumption of *intervention insensitivity*. Healey argues that FR implicitly assume intervention insensitivity in order obtain the logical contradiction. While Healey's intervention sensitivity may seem similar in spirit to the setting-dependence in our framework, [14] appears to regard intervention sensitivity as a special non-local feature of quantum theory. In the FR case, the setting-dependence takes the particular form of enabling signalling

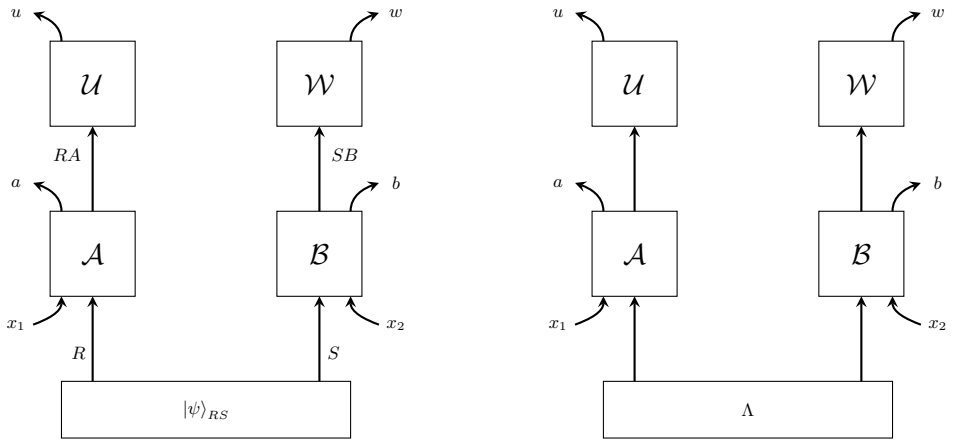

FIG. 10: Left: A concise representation of the augmented circuit (Figure 4) corresponding to the entanglement version of the FR protocol. Here the boxes $\mathcal{A}$ and $\mathcal{B}$ correspond to Alice and Bob's measurements modelled as unitaries $M^{\mathrm{A}}_{unitary}$ and $M^{\mathrm{B}}_{unitary}$ followed by the setting-dependent projectors $\{\pi^A_{x_1}\}$ and $\{\pi^B_{x_2}\}$. 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, 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-signalling) become correlated when conditioning on the future outcome w, and similarly x_2 and a become correlated under post-selection on u.

under post-selection 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 [29] 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 signalling under post-selection also in the FR scenario. Note that our general resolution and results do not depend on this aspect of signalling under post-selection 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 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 $\mathbf{I}_C$ also leads to an apparent paradox here, as the predictions do depend on the settings (the scenario violates $\mathbf{I}_C$). 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 $\mathbf{I}_C$, 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 signalling 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 non-local (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 non-local hidden variable explanations of Bell correlations (see for instance [42]). The correspondence between the FR argument and Hardy's logical argument for contextuality of quantum correlations is discussed in the next section.

Appendix H: Relation to Hardy's logical proof of contextuality

In [43] Lucien Hardy proves the (Bell) non-locality, and hence the contextuality of the correlations arising from a set of bipartite quantum states and measurements, through a logical argument that does not rely on the violation of Bell-type inequalities. The extremal state and measurements of Hardy's argument, as well as the chain of logical

reasoning are in direct correspondence with those of the entanglement version of the FR scenario presented here. We explain the relationship here and also clarify how FR's construction can be seen as an alternative proof of Hardy's theorem, thereby establishing the contextuality of this scenario.

However FR's claimed no-go theorem regarding a logical reasoning paradox between agents is a stronger statement, and we have shown in the main-text that such a paradox can always be avoided within our framework. In other words, while the FR chain of reasoning cannot lead to a logical paradox using quantum theory and classical logic (for observed outcomes) once implicit assumptions about settings are accounted for, the same chain of reasoning can be used in the FR setup to prove the contextuality of the scenario as Hardy did. The relationship between the FR scenario and Hardy's proof has been noted several times before in the literature (for instance, [5, 13, 36, 44]), this section is to be considered as a concrete overview of this relationship along with additional insights provided by our framework.

Consider the bipartite Hardy state

$$|\psi_{Hardy}\rangle_{\mathrm{AB}} := \frac{1}{\sqrt{3}}(|00\rangle + |10\rangle + |11\rangle_{\mathrm{AB}})$$

shared between Alice and Bob and suppose that they perform a Bell type experiment on this state with the setting choices $x_1 \in \{0,1\}$ for Alice and $x_2 \in \{0,1\}$ for Bob and corresponding outcomes $a, b \in \{0,1\}$. Suppose the settings $x_1 = 0$, $x_2 = 0$ correspond to Hadamard basis ($\{|+\rangle, |-\rangle\}$) measurements on the A and B subsystems respectively, in which case we take $a, b = 0$ corresponding to the $+$ outcome and $a, b = 1$ corresponding to the $-$ outcome. And let the settings $x_1 = 1$, $x_2 = 1$ correspond to computational basis ($\{|0\rangle, |1\rangle\}$) measurements on the A and B subsystems. Hardy's argument establishes that the resulting distribution $P(a, b|x_1, x_2)$ is (Bell) non-local (and consequently, contextual), through a logical argument that does not involve a consideration of (Bell-like) inequalities.

This is established as follows. If the correlations were Bell local, then we could simultaneously assign values to the outcomes of all the measurements. Explicitly, let $a_{x_1=0}$ and $a_{x_1=1}$ denote the outcome a of Alice corresponding to the setting choice $x_1 = 0$ and $x_1 = 1$ respectively, and similarly let $b_{x_2=0}$ and $b_{x_2=1}$ denote the outcomes of Bob when his setting $x_2 = 0$ and $x_2 = 1$ respectively. Consider a run of the experiment where the settings $x_1 = x_2 = 0$ are chosen and the outcomes $a_{x_1=0} = b_{x_2=0} = 1$ are obtained. Now using the state $|\psi_{Hardy}\rangle_{\mathrm{AB}}$, one can argue that whenever $a_{x_1=0} = 1$, we must have $b_{x_2=1} = 1$. This is because $a_{x_1=0} = 1$ implies that the post-measurement state on A is $|-\rangle_{\mathrm{A}}$, and $\langle -|_{\mathrm{A}} \langle 0|_{\mathrm{B}} |\psi_{Hardy}\rangle_{\mathrm{AB}} = 0$. We can also see that whenever $b_{x_2=1} = 1$, we must have $a_{x_1=1} = 1$ since $\langle 0|_{\mathrm{A}} \langle 1|_{\mathrm{B}} |\psi_{Hardy}\rangle_{\mathrm{AB}} = 0$. Finally, whenever $a_{x_1=1} = 1$, we must have $b_{x_2=0} = 0$ since $\langle 1|_{\mathrm{A}} \langle -|_{\mathrm{B}} |\psi_{Hardy}\rangle_{\mathrm{AB}} = 0$. This contradicts the fact that $a_{x_1=0} = b_{x_2=0} = 1$ was obtained in the said experimental run, establishing that in such an experimental run one cannot jointly assign values to the outcomes of all measurements. Such correlations are said to exhibit logical contextuality, and we refer the reader to [45] for further details on the same.

We now return to the FR scenario and explain how these directly correspond to the above correlations. For this, note that the $\{|00\rangle, |11\rangle\}$ basis measurements on the RA and SB subsystems of $|\psi^{t=2}\rangle_{\mathrm{RASB}} = \frac{1}{\sqrt{3}}(|0000\rangle + |1100\rangle + |1111\rangle)_{\mathrm{RASB}}$ are operationally equivalent to the computational basis $\{|0\rangle, |1\rangle\}$ measurements on the A and B subsystems of the bipartite (Hardy) state $|\psi_{Hardy}\rangle_{\mathrm{AB}} := \frac{1}{\sqrt{3}}(|00\rangle + |10\rangle + |11\rangle_{\mathrm{AB}})$. Further, the $\{|\mathrm{ok}\rangle, |\mathrm{fail}\rangle\}$ basis measurements on the RA and SB subsystems of $|\psi^{t=2}\rangle_{\mathrm{RASB}}$ are operationally equivalent to the Hadamard basis $\{|+\rangle, |-\rangle\}$ measurements on the A and B subsystems of $|\psi_{Hardy}\rangle_{\mathrm{AB}}$. In the FR scenario, $x_1 = 0$ ensures that the $\{|\mathrm{ok}\rangle_{\mathrm{RA}}, |\mathrm{fail}\rangle_{\mathrm{RA}}\}$ basis measurement is performed directly on $|\psi^{t=2}\rangle_{\mathrm{RASB}}$ giving the outcome u, while $x_1 = 1$ ensures that the $\{|00\rangle_{\mathrm{RA}}, |11\rangle_{\mathrm{RA}}\}$ basis measurement is performed on $|\psi^{t=2}\rangle_{\mathrm{RASB}}$, giving the outcome a, similarly on Bob's side. Thus if we generate a new outcome a' locally on Alice's side such that $a' = 0$ when $x_1 = 0$ and $u = \mathrm{fail}$, $a' = 1$ when $x_1 = 0$ and $u = \mathrm{ok}$, and $a' = a$ when $x = 1$, and similarly the outcome b' on Bob's side such that $b' = 0$ when $x_2 = 0$ and $w = \mathrm{fail}$, $b' = 1$ when $x_2 = 0$ and $w = \mathrm{ok}$ and $b' = b$ when $x_2 = 1$, the distribution $P(a', b'|x_1, x_2)$ is the same as that of the Hardy construction. This distribution is non-local and therefore cannot be explained by a local hidden variable model.

From the above construction, we can see that the choice of setting x_1 (or x_2) determines which of the two measurements (the one corresponding to the computational basis or that corresponding to the Hadamard basis) is performed directly on the subsystem RA (or SB) of the state $|\psi^{t=2}\rangle_{\mathrm{RASB}}$ isomorphic to $|\psi_{Hardy}\rangle_{\mathrm{AB}}$. These measurements are complimentary, and $(x_1, x_2) \in \{(0,0), (0,1), (1,0), (1,1)\}$ specify the four possible measurement contexts in this setup. The Hardy model, and therefore the FR model is logically contextual which means that we cannot jointly assign values to the outcomes of all these measurements [45]. The arguments for the logical proof of contextuality are also in direct correspondence with the logical reasoning steps leading to the paradoxical chain of Equation (C7), with $a_{x_1=0} = 0/1$, $b_{x_2=0} = 0/1$ in the Hardy case corresponding to $u = \mathrm{fail/ok}$, $w = \mathrm{fail/ok}$ in the FR case and $a_{x_1=1} = 0/1$, $b_{x_2=1} = 0/1$ corresponding to $a = 0/1$, $b = 0/1$.

In fact, the FR set-up can also be used to prove Hardy's theorem i.e., to provide a logical proof of contextuality of the underlying quantum states and measurements. The FR argument for the entanglement scenario establishes precisely this (by mirroring the Hardy argument as we have explained above). Therefore, even though the FR

and Hardy theorems have a one-to-one mapping between the states, measurements and statements involved in the proofs, Hardy's conclusion regarding the non-locality/contextuality of the scenario holds true while the validity of FR's claim that "Quantum theory cannot consistently justify the use of itself" (which is a stronger statement than Hardy's) does not generally hold true for quantum theory as our results show, but can be understood as holding only for a specific version of quantum theory that additionally assumes independence of statements from choices of Heisenberg cuts (i.e., when additional assuming **I**), as we have discussed in Section V.

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 categorised into three broad groups. The first category includes works (eg. [28, 46–49]) 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 [5, 10, 17] which (possibly after identifying implicit assumptions beyond Q, C, and S) conclude that there is indeed a paradox in the FR thought experiment. Some of these papers suggest possible resolutions by proposing additional reasoning rules.

The third category includes works such as [11, 13–16, 44, 50] 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.[16]

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 [2, 5]. 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.

Rejecting Q: In FR's original paper, the authors interpret the violation of Q as a violation of unitary quantum theory. Following [5], 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 U.

Rejecting Q means at least one prediction of the FR scenario (e.g., Equation (16) 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 Q. As noted in [51], certain versions of Bohmian mechanics, a non-local hidden variable interpretation, violate 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 III.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 III.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 **Q** formalised in our work, and they would not correspond to the setting-conditioned predictions of the augmented circuit (Definition III.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 non-trivial notion of such a cut. But such a fully classical theory could "emulate" a unitary measurement channel or setting $x_i = 0$ through non-local hidden mechanisms.

Moreover, by applying the general premise of our reasoning rules discussed in Section VII A, FR type paradoxes can also 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).

Rejecting U: The U assumption, first noted in [5], involves rejecting the idea that agents' labs evolve as unitary closed quantum systems. Interpretations involving objective mechanisms for "wavefunction collapse" reject

[16] The cited works are only representatives of the extensive research generated by the FR paper. For further references on previous responses to FR, see [9].

this, where unitary quantum theory breaks down at certain macroscopic scales. Such interpretations also violate the **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 formalisation of the U assumption as **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 **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 $\mathcal{M}^{\mathrm{A}_i}$ in our EWFS definition (Definition III.1) as a "measurement" according to this view point, one must restrict to standard quantum scenarios (Definition VI.3), where one agent cannot have complete quantum control over the labs of other agents. As shown by Corollary VI.1, in these scenarios, a joint distribution for all (non-trivial) 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 [28] makes an argument for objective decoherence based on the concept of pre-measurements. The notion of a pre-measurement coincides with the case of setting 0 in our framework i.e., the modelling of the measurement purely as a unitary evolution. They argue that pre-measurements 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 that no non-trivial measurement outcome can be assigned to a measurement modelled with setting 0, non-trivial outcomes require setting 1 (applying the projection postulate). The authors of [28] then suggest that this observation resolves the FR apparent paradox as FR's reasoning assigns outcome values to pre-measurements. However, [28] imposes an objective distinction between pre-measurements and measurements, considering the operations of the agents Alice and Bob only as pre measurements (setting 0) and those of the super-agents 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.

Rejecting C or D: The assumption C allows agents to inherit each other's knowledge. FR suggested that relational approaches, such as relational quantum mechanics [52] and QBism [53, 54], reject this assumption by allowing different subjective perspectives. However, since the original C assumption left unclear the formal modeling of agents' knowledge and choices of Heisenberg cuts, we have proposed a formalisation of this assumption in our framework, as **C**.

In subjective/relational approaches to EWFSs, agents can have different choices of Heisenberg cuts. Super-agents 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 Theorem IV.1 and Section 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 IV.2. Ignoring these setting choices leads to apparent breakdowns of these axioms, but, as argued in Section 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 [55], 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.

Rejecting S: Previous discussions [2, 5, 9] suggest that many-worlds interpretations [56] would tend to reject S because an outcome a_i could be 0 in one "branch" of the wavefunction 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 non-orthogonal state [16]. 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 are 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 wavefunction 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 wavefunction, 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.[17] Our formal version of each FR assumption is satisfied without contradictions. Nevertheless, it is insightful to interpret previous works' proposed reasoning rules.

Adding context labels At a more pedagogical level, [5] suggests an additional reasoning rule that could incorporated to avoid the paradox, which involves tagging all statements made by agents with certain "context labels" and only combining statements with matching labels. However no particular model for these labels or a rigorous formalisation 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 Theorem IV.1 and Theorem VI.1). We have also addressed the efficiency and causality issues.

Parsing rule for measurements Another recent work [10] proposes a parsing rule for quantum theory to determine when a unitary/isometry can be considered a measurement with a classical outcome. This rule deems an operation a measurement if no future non-commuting 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 [17]. Unlike our framework, which uses settings to describe measurements, [10] maintains ambiguity in measurement modeling, but the additional parsing rule clarifies when the ambiguity can be safely ignored in agents' reasoning.

If an operation $\mathcal{M}^{\mathrm{A}_i}$ in the EWFS description is deemed a measurement according to this parsing rule, then it would imply that there are no super-agents to the agent A_i in the scenario (also according to our definition of the non-superagent structure). As shown in Theorem E.1, 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 formalised without reference to commutativity of operations. In particular, a non-commuting operation acting solely on the memory M_i might fail the parsing rule of [10] 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 super-agents 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 [10] and another work [12] 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.

Consistent histories interpretation Another notable approach is the consistent histories (CH) interpretation of quantum theory [26, 27]. 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 beenn applied to explain a number of quantum paradoxes arising in standard (non-Wigner's Friend like) quantum scenarios, such as contextuality paradoxes, pre and post-selection paradoxes

[17] Although additional steps can be employed to simplify the reasoning by dropping redundant settings and parameters, as is also the case in classical multi-agent scenarios (cf. Section VII A).

etc. In the context of Wigner's Friend scenarios, [11] 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 IV.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 formalised 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 in 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 non-commuting projectors central to the CH approach are a non-classical aspect, we have discussed how our resolution of the paradox shares strong similarities with how analogous multi-agent inconsistencies are resolved in purely classical theories (see also the previous paragraph on parsing rules, for further discussion on the link between non-commutativity 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 [9]. In particular, [9] 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 FR's, there is no single objective history of events that is realised.

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 (modelling the Heisenberg cuts) in accordance with universal validity of unitary quantum theory and without assuming the existence of 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 [2, 5]) which are violated [11, 13–16, 50]. Some specific examples include Scott Aaronson's blog post that refers to an additional "unformalised" assumption of FR, Healey's assumption of intervention insensitivity [14], and the assumption regarding collapse/no-collapse that Araujo points out [15]. 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) non-locality 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 non-locality, 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, this correspondence is explained in Appendix H where we discuss the relation between FR's protocol and Hardy's proof regarding (Bell) non-locality without inequalities.

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 formalise Araujo's assumption. Araujo also noted issues with FR's treatment of post-selection, which are accounted for in our analysis by explicitly computing probabilities conditioned on the post-selection and avoiding collider bias. We have also seen that in the FR scenario (Appendix F 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 Section V C, 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 fuelled an intriguing research program on EWFSs in quantum foundations inspiring other no-go theorems (such as [3, 4]) based on a similar set-up where the underlying assumptions are formalised more rigorously.

Appendix J: Proofs of all results

1. Proofs of results from the main text

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

$$S \in \Sigma \quad \Rightarrow \quad \neg S \cap \Sigma = \emptyset, \tag{6}$$

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 $k = \hbar$, then the outcomes $\vec{a}_l$ take values $\vec{a}_l$ with a probability $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, k = \hbar)$." 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 $k = \hbar$, then the outcomes $\vec{a}_l$ **do not** take values $\vec{a}_l$ with a probability $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, k = \hbar)$." More precisely, this is a set of statements:

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

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

Corollary IV.1. *If any two agents use the same choice of settings $\vec{x}$ for all measurements $\{\mathcal{M}^{\mathrm{A}_i}\}_i$ 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 of Figure 1 and therefore the joint state of all systems and memories $\mathrm{S}_1, \ldots \mathrm{S}_m, \mathrm{M}_1, \ldots, \mathrm{M}_N$ 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 III.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 D 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. □

Theorem IV.1.

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 $P_{conv}(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l)$ 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 $i \in \{j_1, ..., j_p, l_1, ..., l_q\}$ and $x_i = 0$ for all $i \notin \{j_1, ..., j_p, l_1, ..., l_q\}$.*

2. *Consistency: For any EWFS, the set of all statements Σ^{aug} obtained in the corresponding augmented circuit (Definition III.10) are consistent according to Definition III.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 $\mathrm{A}_i \not\prec \mathrm{A}_k$ for all $k \in \{j_1, ..., j_p, l_1, ..., l_q\}$, the prediction is independent of the setting x_i. That is, for all such i, we have the following, where we denote $P(a = a)$ as $P(a)$ for short and note that $\vec{x} = (x_1, ..., x_N)$.*

$$\begin{aligned} \forall \xi_i, \xi_i', \quad & P(\vec{a}_j|\vec{a}_l, (\xi_1, ...\xi_i, ..., \xi_N)) = \\ & P(\vec{a}_j|\vec{a}_l, (\xi_1, ...\xi_i', ..., \xi_N)). \end{aligned} \tag{10}$$

Proof. 1. *Completeness:* Recall that a conventional prediction $P_{conv}(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l)$ in an EWFS (Definition III.6) is computed by modelling the measurements $\mathcal{M}^{\mathrm{A}_i}$ for all $i \in \{j_1, ..., j_p, l_1, ..., l_q\}$ as purely unitary evolutions of the agents' labs (Equation (3)). By construction of the augmented circuit for the EWFS, this corresponds to the case where $x_i = 0$ for all $i \in \{j_1, ..., j_p, l_1, ..., l_q\}$.

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

Now, following Section III B of the main text, consider an initial state $|\psi\rangle_{\mathtt{S}_i} \otimes |0\rangle_{\mathrm{M}_i}$ of the system and memory (on which the measurement $\mathcal{M}^{\mathrm{A}_i}$ acts) where $|\psi\rangle_{\mathtt{S}_i} = \sum_{a_i \in \mathtt{O}_i} c_{a_i} |a_i\rangle_{\mathtt{S}_i}$. That section of the main text shows that the following two procedures are operationally equivalent: (1) a projective measurement $\{\pi_{a_i}^{\mathtt{S}_i} = |a_i\rangle\langle a_i|_{\mathtt{S}_i}\}_{a_i \in \mathtt{O}_i}$ is applied on the system and then the unitary channel $\mathcal{M}_{unitary}^{\mathrm{A}_i}$ is applied on the system and memory (2) the unitary channel $\mathcal{M}_{unitary}^{\mathrm{A}_i}$ is applied on the system and memory, and then the projective measurement $\{\pi_{a_i}^{\mathtt{S}_i\mathrm{M}_i} = |a_i a_i\rangle\langle a_i a_i|_{\mathtt{S}_i\mathrm{M}_i}\}_{a_i \in \mathtt{O}_i}$ is performed on the system and memory. The operational equivalence of (1) and (2) 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 (1) and (2) extends to all initial states $\rho_{\mathtt{S}_i} \otimes |0\rangle\langle 0|_{\mathrm{M}_i}$ of the system and memory. (1) is the procedure used for dealing with a measurement $\mathcal{M}^{\mathrm{A}_i}$ when computing conventional predictions involving the outcome a_i (Definition III.6) while (2) is the procedure used in the augmented EWFS for calculating the setting conditioned predictions (Definition III.8) involving the outcome a_i where by default we will 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 $\vec{x} = \vec{\xi^*}$ defined as: $x_i = 1$ for $i \in \{j_1, ..., j_p, l_1, ..., l_q\}$ and $x_i = 0$ otherwise. The only difference being that the corresponding prediction in the augmented circuit explicitly conditions on the setting choice $\vec{x} = \xi^*$, and we have $P_{conv}(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l) = P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi^*})$.

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. *C*onsistency: The fact that the set Σ^{aug} of statements associated with an EWFS in our framework satisfies the consistency condition of Definition III.5 immediately follows from the procedure through which these statements are derived.

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

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

3. *C*ausality: 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

[29–31, 57]. 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 $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi})$, and the condition (C) $\mathrm{A}_i \not\prec \mathrm{A}_k$ for all $k \in \{j_1, ..., j_p, l_1, ..., l_q\}$. Notice that an A_i satisfies this condition, in particular when $t_i > \max(t_{j_1}, ..., t_{j_p}, t_{l_1}, ..., t_{l_q})$. In Appendix D, we have established that any setting-conditioned prediction $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi})$ in an augmented EWFS can be simplified to the form of Equation (D6), where the components x_i of the setting vector $\vec{x}$ which are associated with a time $t_i > \max(j_1, ..., j_p, l_1, ..., l_q)$ do not feature in the probability expression, which implies the required independence Equation (10) for all such settings x_i.

For cases where we have $t_i < t_k$ for some $k \in \{j_1, ..., j_p, l_1, ..., l_q\}$ but $\mathrm{A}_i \not\prec \mathrm{A}_k$, by construction, this would only happen when the circuit contains no directed path of wires from the measurement $\mathcal{M}^{\mathrm{A}_i}$ at t_i to the measurement $\mathcal{M}^{\mathrm{A}_k}$ 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 non-signalling 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.[18]

□

Corollary IV.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 5 assumptions are satisfied by statements Σ^{aug} derived in our augmented circuit framework, and the consistency of our framework shown in Theorem IV.1 guarantees that all 5 can be consistently applied to reason without any contradictions.

Q, **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 modelling). Moreover, all our results apply to general EWFS (Definition III.1) which allows agents to have full quantum control over the labs of others, in the precise sense described in the formalisation 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 normalised probabilities (see Appendix D) and the set of statements Σ^{aug} is consistent as proven in Theorem IV.1. No single valid normalised 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 $S \in \Sigma^{aug}$, then A_i can directly compute the setting-conditioned prediction $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi})$ that defines the statement S and thereby inherit the knowledge. Consistency as shown in Theorem IV.1 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 $\vec{a}_j = \vec{a}_j$, $\vec{a}_l = \vec{a}_l$.[19]

Finally the distributive axiom **D** is a rather basic axiom of logical inference, that it holds in our framework can be seen as follows. Firstly, since $\Sigma_L^{aug} \subseteq \Sigma^{aug}$, consistency of the superset as shown in Theorem IV.1 implies consistency of the subset. Then consider a logical statement $S_1 \in \Sigma_L^{aug}$ associated with a logical setting-conditioned prediction $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi}) \in \{0, 1\}$, it can be of the following form where the set $\vec{a}_l$ of outcomes can be empty

$$\begin{aligned} &\vec{a}_l = \vec{a}_l \wedge \vec{x} = \vec{\xi} \Rightarrow \neg(\vec{a}_j = \vec{a}_j), \\ &\vec{a}_l = \vec{a}_l \wedge \vec{x} = \vec{\xi} \Rightarrow \vec{a}_j = \vec{a}_j. \end{aligned} \tag{J1}$$

[18] 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 space-like separation, then the above transformation can be seen as transforming to another reference frame where the time order relative to the co-ordinate time is reversed.

[19] 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.

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.,

$$\begin{aligned} &\vec{a}_n = \vec{a}_n \wedge \vec{x} = \vec{\xi}' \Rightarrow \neg(\vec{a}_m = \vec{a}_m), \\ &\vec{a}_n = \vec{a}_n \wedge \vec{x} = \vec{\xi}' \Rightarrow \vec{a}_m = \vec{a}_m. \end{aligned} \tag{J2}$$

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 $\vec{a}_m$, $\vec{a}_n$ 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 III.5 and Theorem IV.1). This shows that $S_2 \in \Sigma_L^{aug}$. As this holds for all agents using our framework to reason, it follows that Equation (13) holds.

□

Theorem V.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 Section IV and Corollary IV.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 F 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

$$\begin{aligned} &P(b = 1 | u = w = \text{ok}, (x_1, x_2) = (0, 1)) \neq P(b = 1 | u = w = \text{ok}, (x_1, x_2) = (1, 1)), \\ &P(w = \text{fail} | a = 1, (x_1, x_2) = (1, 0)) \neq P(w = \text{fail} | a = 1, (x_1, x_2) = (1, 1)) \end{aligned} \tag{J3}$$

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

$$\begin{aligned} &P(w = \text{fail} | r = \text{tails}, (x_1, x_2) = (1, 0)) \neq P(w = \text{fail} | r = \text{tails}, (x_1, x_2) = (1, 1)), \\ &P(z = +\frac{1}{2} | \overline{w} = \overline{\text{ok}}, (x_1, x_2) = (0, 1)) \neq P(z = +\frac{1}{2} | \overline{w} = \overline{\text{ok}}, (x_1, x_2) = (1, 1)). \end{aligned} \tag{J4}$$

These precisely correspond to the three statements $\overline{\text{F}}^{n:02}$ and $\overline{\text{W}}^{n:22}$ 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 IV.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 Table I and Table II that this recovers the original statements of FR (in both versions) and therefore the apparent paradox. □

Theorem VI.1 (Non-action on memory and setting-independence)**.** *Consider an EWFS and a subset* $\text{A}_\mathcal{K}$ *of agents therein. Suppose that* $\text{A}_i \notin \text{A}_\mathcal{K}$ *is another agent in the EWFS such that no agent in* $\text{A}_\mathcal{K}$ *acts as a superagent to* A_i *i.e.,* $(\text{A}_i, \text{A}_k) \in n\mathcal{SA}\ \forall A_k \in A_\mathcal{K}$*. Then for every partition* $A_\mathcal{K} = \{\text{A}_{j_1}, ..., \text{A}_{j_p}\} \cup \{\text{A}_{l_1}, ..., \text{A}_{l_q}\}$ *of* $\text{A}_\mathcal{K}$*, 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,*

$$\begin{aligned} &P(\vec{a}_j | \vec{a}_l, (\xi_1, ..., \xi_N)) = \\ &P(\vec{a}_j | \vec{a}_l, (\xi_1, ...\xi_{i-1}, \xi_{i+1}..., \xi_N)). \end{aligned} \tag{22}$$

Recall that this expression is equivalent to the conditional independence given in Equation (10)*.*

Proof. Given any set $\mathrm{A}_{\mathcal{K}}$ of agents and another agent A_i such that $(\mathrm{A}_i, \mathrm{A}_k) \in n\mathcal{SA}\ \forall \mathrm{A}_k \in \mathrm{A}_{\mathcal{K}}$, we will first show that the joint probability of the outcomes of all agents in $\mathrm{A}_{\mathcal{K}}$ (these outcomes will be denoted using the vector $\vec{a}_{\mathcal{K}}$) is independent of the setting x_i i.e.,

$$P(\vec{a}_{\mathcal{K}}|(x_1, ..., x_N)) = P(\vec{a}_{\mathcal{K}}|(x_1, ..., x_{i-1}, x_{i+1}, ...x_N)). \tag{J5}$$

Here we have shortened the usual notion using value assignments $a = a$ in probabilities to just the variable a, since the meaning of the equation is clear from context. From this, the desired result will follow easily through the rules of conditional probability.

To establish Equation (J5), we consider the operationally equivalent augmented circuit (involving agents $\{\mathrm{A}'_1, ..., \mathrm{A}'_N\}$) to the original augmented circuit (involving agents $\{\mathrm{A}_1, ..., \mathrm{A}_N\}$), whose causal structure more explicitly reflects non-superagent structure $n\mathcal{SA}$ of the original EWFS. Recall that such an equivalent circuit is guaranteed to exist by Definition VI.2. The equivalence implies in particular, the equivalence of predictions in the two circuits, that is, for all disjoints sets of outcomes $\vec{a}_j$ and $\vec{a}_l$ in the original EWFS,

$$P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi}) = P(\vec{a}'_j = \vec{a}_j|\vec{a}'_l = \vec{a}_l, \vec{x}' = \vec{\xi}). \tag{J6}$$

We establish Equation (J5) 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 VI.2, $(\mathrm{A}_i, \mathrm{A}_k) \in n\mathcal{SA}$ (along with the fact that we have $i \neq k$) implies that in the equivalent EWFS, one of the following cases must hold

- **Case 1:** $\mathrm{A}'_i \not\prec \mathrm{A}'_k$
- **Case 2:** $\mathrm{A}'_i \prec \mathrm{A}'_k$, and $\mathcal{E}'_i$ acts trivially on M'_i and $\mathrm{A}_i \not\prec^{\mathrm{M}'_i} \mathrm{A}_k$

Case 1 In this case Equation (J5) immediately follows from the causality result of Theorem IV.1.

Case 2 In this case we must have the following property: for any agent A'_j with $\mathrm{A}'_i \prec^{\mathrm{M}'_i} \mathrm{A}'_j$ we must have $\mathrm{A}'_j \not\prec \mathrm{A}'_k$ for all $k \in \mathcal{K}$. Otherwise $\mathrm{A}'_i \prec^{\mathrm{M}'_i} \mathrm{A}'_j$ along with the directed path $\mathrm{A}'_j \prec \mathrm{A}'_k$ would immediately imply by Definition IV.1 that $\mathrm{A}'_i \prec^{\mathrm{M}'_i} \mathrm{A}'_k$, which would contradict the condition of Case 2 and hence the fact that $(\mathrm{A}_i, \mathrm{A}_k) \in n\mathcal{SA}$ Definition VI.2.

Now, the prediction of interest from Equation (J5) can be calculated explicitly by applying Equation (D3) as detailed in Appendix D. The above property guarantees that for all A'_j such that $\mathrm{A}'_i \prec^{\mathrm{M}'_i} \mathrm{A}'_j$, the operations of A'_j and those of any agent $\mathrm{A}'_k \in \mathrm{A}'_{\mathcal{K}}$ must act of disjoint sets of systems Definition IV.1. Note that $\mathrm{A}'_i \prec^{\mathrm{M}'_i} \mathrm{A}'_j$ implies that A'_j acts after A'_i in the augmented circuit. Then by the non-signalling property of quantum theory, and as we can explicitly see from Equation (D3), the trace in the probability calculation will act on all the outputs of A'_j's operation. Since A'_j is not in the set $\mathrm{A}'_{\mathcal{K}}$, 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 $\mathrm{tr} \circ \mathcal{E} = \mathrm{tr}$.

The above implies that all the operations of the agents A'_j such that $\mathrm{A}'_i \xrightarrow{\mathrm{M}'_i} \mathrm{A}'_j$ will drop out of the probability expression for Equation (J5). This probability expression will then have the trace over M'_i acting directly on the operation $\mathcal{E}'_i \circ \mathcal{M}'^{\mathrm{A}_i}$ 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 will act directly on the measurement $\mathcal{M}^{\mathrm{A}_i}$.

It is easy to see that $\mathrm{tr}_{\mathrm{M}'_i} \circ \mathcal{M}^{\mathrm{A}_i}$ 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 post-measurement 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 Theorem E.1. This is sufficient to establish the required Equation (J5) for this case.

Having established Equation (J5) for all sets $\mathrm{A}_{\mathcal{K}}$ of agents of the form required by the theorem statement, it is immediate that the same setting independence of Equation (J5) also holds for all subsets of agents $\mathrm{A}_{\mathcal{K}}$ (this can be seen by computing the relevant marginals of Equation (J5)). Now, for any partition $A_{\mathcal{K}} = \{\mathrm{A}_{j_1}, ..., \mathrm{A}_{j_p}\} \cup \{\mathrm{A}_{l_1}, ..., \mathrm{A}_{l_q}\}$ of $\mathrm{A}_{\mathcal{K}}$, we can compute prediction $P(\vec{a}_j|\vec{a}_l, \vec{x})$ by applying the conditional probability rule

$$P(\vec{a}_j|\vec{a}_l, \vec{x}) = \frac{P(\vec{a}_j, \vec{a}_l|\vec{x})}{P(\vec{a}_l|\vec{x})}. \tag{J7}$$

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 Equation (J5). This establishes the theorem.

□

2. Proofs of results from the Appendix

Theorem E.1 (Recovering standard quantum circuits)**.** *If an EWFS corresponds to a standard quantum scenario (Definition VI.3), then its augmented circuit can be equivalently reduced to a standard quantum circuit, such that the same (non-trivial) predictions are obtained from the original augmented circuit, the $\mathcal{C}^{sys}$-form standard circuit or the $\mathcal{C}^{sys+anc}$-form standard circuit. Explicitly, for any disjoint sets $\vec{a}_j = (a_{j_1}, ..., a_{j_p})$ and $\vec{a}_l = (a_{l_1}, ..., a_{l_q})$ of outcomes, and any choice of settings $\vec{x} = \vec{\xi}$ such that $x_i = 1$ for all $i \in \{j_1, ..., j_p, l_1, ..., l_q\}$, we have*

$$P_{aug}(\vec{a}_j = \vec{a}_j | \vec{a}_j = \vec{a}_j, \vec{x} = \vec{\xi}) = P_{std}(\vec{a}'_j = \vec{a}_j | \vec{a}'_j = \vec{a}_j), \tag{E1}$$

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 $\mathcal{C}^{sys}$-form or $\mathcal{C}^{sys+anc}$-form standard quantum circuit (where no settings are involved).

Proof. In a standard quantum scenario, by Definition VI.3, we must have $(\mathrm{A}_i, \mathrm{A}_j) \in n\mathcal{SA}$ for all $i, j \in \{1, ..., N\}$. This means that the augmented circuit of the given EWFS over agents $\{\mathrm{A}_1, ..., \mathrm{A}_N\}$ can be reduced to an operationally equivalent augmented circuit over corresponding agents $\{\mathrm{A}'_1, ..., \mathrm{A}'_N\}$ such that for any A'_i and A'_j we have: either A'_j acts before A'_i in time $t'_i > t'_j$ or in the alternative cases where $i = j$ or $t'_j > t'_i$, no operation that acts after the measurement $\mathcal{M}^{\mathrm{A}'_i}$ in the circuit (including the operation $\mathcal{E}'_i$) acts non-trivially on the memory M'_i. Here we have used the fact that $\mathrm{A}'_i \prec \mathrm{A}'_j$ according to the operational causal structure (Definition IV.1) implies $t'_i < t'_j$.

Applying this argument to every pair of agents in the scenario, this implies that for each measurement $\mathcal{M}^{\mathrm{A}'_i}$, the corresponding set of systems $\mathtt{S}'_i$ of systems (excluding the memory M'_i) on which it acts non-trivially is a subset of the systems $\mathtt{S}'$ and does not include any of the memories $\mathrm{M}'_j \in \mathtt{M}'$ of other agents (in contrast to a general EWFS of Definition III.1 where the "system" for one agent's measurement may include the " memories" of other agents). All the following arguments will 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 $a_i \in \mathtt{O}_i$ as that of the original scenario w.l.o.g.[20] Then A'_i's measurement on the systems $\mathtt{S}'_i$ is associated with the projectors $\{|a_i\rangle \langle a_i|_{\mathtt{S}'_i}\}_{a_i \in \mathtt{O}_i}$. Let $\rho_{\mathtt{S}'_i}$ be an arbitrary state, which corresponds to the state of the system $\mathtt{S}'_i$ just before A'_i's measurement. We can express this state is the basis of the measurement as follows, for some coefficients $c_{a_i, \bar{a}_i}$.

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

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

$$\mathcal{M}^{\mathrm{A}'_i}_{unitary}(\rho_{\mathtt{S}'_i} \otimes |0\rangle \langle 0|_{\mathrm{M}'_i}) \mathcal{M}^{\mathrm{A}'_i, \dagger}_{unitary} = \sum_{a_i, \bar{a}_i} c_{a_i, \bar{a}_i} |a_i a_i\rangle \langle \bar{a}_i \bar{a}_i|_{\mathtt{S}'_i \mathrm{M}'_i}. \tag{J8}$$

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

$$\begin{aligned} &(1_{\mathtt{S}'_i} \otimes \pi^{\mathrm{M}'_i}_{a_i}) \big(\mathcal{M}^{\mathrm{A}'_i}_{unitary}(\rho_{\mathtt{S}'_i} \otimes |0\rangle \langle 0|_{\mathrm{M}'_i}) \mathcal{M}^{\mathrm{A}'_i, \dagger}_{unitary} \big) (1_{\mathtt{S}'_i} \otimes \pi^{\mathrm{M}'_i}_{a_i}) \\ =&\pi^{\mathtt{S}'_i \mathrm{M}'_i}_{a_i} \mathcal{M}^{\mathrm{A}'_i}_{unitary}(\rho_{\mathtt{S}'_i} \otimes |0\rangle \langle 0|_{\mathrm{M}'_i}) \mathcal{M}^{\mathrm{A}'_i, \dagger}_{unitary} \pi^{\mathtt{S}'_i \mathrm{M}'_i}_{a_i} \end{aligned} \tag{J9}$$

We now show that the post-measurement on the system $\mathtt{S}'_i$ alone is independent of the setting x'_i, and this indeed the post-measurement state one would get by directly applying the projective measurement of $\mathcal{C}^{sys}$ on $\mathtt{S}'_i$. This will allows us to reduce our augmented circuit to a $\mathcal{C}^{sys}$-form standard quantum circuit as required. For this,

[20] Since the scenarios are operationally equivalent with a one-one-one correspondence between the outcome sets, we can use the same labels without loss of generality.

first consider this state under the setting $x_i' = 0$. This is given as $\mathrm{tr}_{\mathrm{M}_i'}\big[\mathcal{M}^{\mathrm{A}_i'}_{unitary}(\rho_{\mathtt{S}_i'} \otimes |0\rangle\langle 0|_{\mathrm{M}_i'})\mathcal{M}^{\mathrm{A}_i',\dagger}_{unitary}\big]$. Noting that $\sum_{a_i \in \mathtt{O}_i} \pi^{\mathrm{M}_i'}_{a_i} = 1_{\mathrm{M}_i'}$, we can expand this as follows

$$
\begin{aligned}
&\mathrm{tr}_{\mathrm{M}_i'}\big[\mathcal{M}^{\mathrm{A}_i'}_{unitary}(\rho_{\mathtt{S}_i'} \otimes |0\rangle\langle 0|_{\mathrm{M}_i'})\mathcal{M}^{\mathrm{A}_i',\dagger}_{unitary}\big] \\
&= \sum_{\bar{a}_i \in \mathtt{O}_i} (1_{\mathtt{S}_i'} \otimes \langle \bar{a}_i|_{\mathrm{M}_i'})\big(\mathcal{M}^{\mathrm{A}_i'}_{unitary}(\rho_{\mathtt{S}_i'} \otimes |0\rangle\langle 0|_{\mathrm{M}_i'})\mathcal{M}^{\mathrm{A}_i',\dagger}_{unitary}\big)(1_{\mathtt{S}_i'} \otimes |\bar{a}_i\rangle_{\mathrm{M}_i'}) \\
&= \sum_{\bar{a}_i \in \mathtt{O}_i} (1_{\mathtt{S}_i'} \otimes \langle \bar{a}_i|_{\mathrm{M}_i'})(1_{\mathtt{S}_i'} \otimes \sum_{a_i \in \mathtt{O}_i} \pi^{\mathrm{M}_i'}_{a_i})\big(\mathcal{M}'^{\mathrm{A}_i'}_{unitary}(\rho_{\mathtt{S}_i'} \otimes |0\rangle\langle 0|_{\mathrm{M}_i'})\mathcal{M}^{\mathrm{A}_i',\dagger}_{unitary}\big)(1_{\mathtt{S}_i'} \otimes \sum_{a_i \in \mathtt{O}_i} \pi^{\mathrm{M}_i'}_{a_i})(1_{\mathtt{S}_i'} \otimes |\bar{a}_i\rangle_{\mathrm{M}_i'}) \\
&= \sum_{\bar{a}_i \in \mathtt{O}_i} \sum_{a_i \in \mathtt{O}_i} (1_{\mathtt{S}_i'} \otimes \langle \bar{a}_i|_{\mathrm{M}_i'})(1_{\mathtt{S}_i'} \otimes \pi^{\mathrm{M}_i'}_{a_i})\big(\mathcal{M}^{\mathrm{A}_i'}_{unitary}(\rho_{\mathtt{S}_i'} \otimes |0\rangle\langle 0|_{\mathrm{M}_i'})\mathcal{M}^{\mathrm{A}_i',\dagger}_{unitary}\big)(1_{\mathtt{S}_i'} \otimes \pi^{\mathrm{M}_i'}_{a_i})(1_{\mathtt{S}_i'} \otimes |\bar{a}_i\rangle_{\mathrm{M}_i'}) \\
&= \mathrm{tr}_{\mathrm{M}_i'}\big[\sum_{a_i \in \mathtt{O}_i} (1_{\mathtt{S}_i'} \otimes \pi^{\mathrm{M}_i'}_{a_i})\big(\mathcal{M}^{\mathrm{A}_i'}_{unitary}(\rho_{\mathtt{S}_i'} \otimes |0\rangle\langle 0|_{\mathrm{M}_i'})\mathcal{M}^{\mathrm{A}_i',\dagger}_{unitary}\big)(1_{\mathtt{S}_i'} \otimes \pi^{\mathrm{M}_i'}_{a_i})\big].
\end{aligned}
\tag{J10}
$$

In the second equality above, we have summed over the same indices $\sum_{a_i \in \mathtt{O}_i}$ in both occurences of the projectors $\pi^{\mathrm{M}_i'}_{a_i}$ as the cross terms disappear due to the trace over M_i'. Using this and Equation (J9), we immediately obtain the following

$$
\begin{aligned}
&\mathrm{tr}_{\mathrm{M}_i'}\big[\mathcal{M}^{\mathrm{A}_i'}_{unitary}(\rho_{\mathtt{S}_i'} \otimes |0\rangle\langle 0|_{\mathrm{M}_i'})\mathcal{M}^{\mathrm{A}_i',\dagger}_{unitary}\big] \\
&= \mathrm{tr}_{\mathrm{M}_i'}\big[\sum_{a_i \in \mathtt{O}_i} \pi^{\mathtt{S}_i'\mathrm{M}_i'}_{a_i}\mathcal{M}^{\mathrm{A}_i'}_{unitary}(\rho_{\mathtt{S}_i'} \otimes |0\rangle\langle 0|_{\mathrm{M}_i'})\mathcal{M}^{\mathrm{A}_i',\dagger}_{unitary}\pi^{\mathtt{S}_i'\mathrm{M}_i'}_{a_i}\big] \\
&= \sum_{a_i \in \mathtt{O}_i} c_{a_i,a_i}\,|a_i\rangle\langle a_i|_{\mathtt{S}_i'}\,.
\end{aligned}
\tag{J11}
$$

The right hand side is precisely the expression for the post-measurement state of $\mathtt{S}_i'$ under the setting $x_i = 1$ when using the trace preserving form of the evolution for this setting (c.f. Equation (9)). Thus we have shown that A_i''s measurement implements the following map on an arbitrary input state $\rho_{\mathtt{S}_i'}$ of the system $\mathtt{S}_i'$ irrespective of the setting x_i'.

$$
\rho_{\mathtt{S}_i'} = \sum_{a_i,\bar{a}_i \in \mathtt{O}_i} c_{\bar{a}_i,\bar{a}_i}\,|\bar{a}_i\rangle\langle \bar{a}_i|_{\mathtt{S}_i'} \mapsto \sum_{a_i \in \mathtt{O}_i} c_{\bar{a}_i,a_i}\,|a_i\rangle\langle a_i|_{\mathtt{S}_i'}
\tag{J12}
$$

This map can equivalently be described as

$$
\rho_{\mathtt{S}_i'} \mapsto \sum_{a_i \in \mathtt{O}_i} \pi^{\mathtt{S}_i'}_{a_i}(\rho_{\mathtt{S}_i'})\pi^{\mathtt{S}_i'}_{a_i}.
\tag{J13}
$$

Or, for the case of a particular outcome, we have the following trace non-increasing map from pre to post measurement state

$$
\rho_{\mathtt{S}_i'} \mapsto \frac{\pi^{\mathtt{S}_i'}_{a_i}(\rho_{\mathtt{S}_i'})\pi^{\mathtt{S}_i'}_{a_i}}{\mathrm{tr}\Big[\pi^{\mathtt{S}_i'}_{a_i}(\rho_{\mathtt{S}_i'})\pi^{\mathtt{S}_i'}_{a_i}\Big]}.
\tag{J14}
$$

These correspond precisely to the how the measurements $\mathcal{M}^{\mathrm{A}_i'} = \{\pi^{\mathtt{S}_i'}_{a_i}\}_{a_i \in \mathtt{O}_i'}$ of a $\mathcal{C}^{sys}$-form circuit (Definition E.1) over $\{A_1', ..., A_N'\}$ act.

To show the full reduction of the augmented circuit to a $\mathcal{C}^{sys}$-form circuit, let $\mathtt{S}_i'^{c} := \mathtt{S}' \backslash \mathtt{S}_i'$ denote the complement of $\mathtt{S}_i \subseteq \mathtt{S}'$ (the systems that $\mathcal{M}^{\mathrm{A}_i'}$ acts non-trivially on). The above argument, immediately generalises to the case where we consider $\rho_{\mathtt{S}_i'}$ to be a reduced state of a larger state $\rho_{\mathtt{S}_i'^{c}\mathtt{S}_i'}$ over all systems $\mathtt{S}'$ in the EWFS. Since under both settings, the measurement operation of A_i' only acts locally on $\mathtt{S}_i'$ by construction, the above argument implies that the joint state on $\mathtt{S}_i'^{c}\mathtt{S}_i'$ 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 $\{|a_i\rangle\langle a_i|_{\mathtt{S}_i'}\}_{a_i \in \mathtt{O}_i'}$ on the system $\mathtt{S}_i' \subseteq \mathtt{S}'$ alone while preserving all the predictions i.e., the augmented circuit reduces to an equivalent $\mathcal{C}^{sys}$-form standard circuit. Since the primed circuit is operationally equivalent to the original augmented EWFS, it follows that the $\mathcal{C}^{sys}$-form standard circuit obtained here is also operationally equivalent to the original augmented circuit.

To obtain an equivalent $\mathcal{C}^{sys+anc}$-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 Equation (J9) that this is equivalent to modelling each measurement in its unitary form $\mathcal{M}^{\mathrm{A}'_i}_{unitary}$ at the time t'_i of the measurement followed by a projective measurement $\{\pi^{\mathrm{M}'_i}_{a_i}\}_{a_i \in \mathtt{O}_i}$ 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 $\{\pi^{\mathrm{M}'_i}_{a_i}\}_{a_i \in \mathtt{O}_i}$ acting on the memory alone, can be equivalently performed at any time t_f after t'_N. This immediately yields $\mathcal{C}^{sys+anc}$-form standard circuit which is equivalent to the original augmented circuit, where the memories $\{\mathrm{M}'_1, ..., \mathrm{M}'_N\}$ act as the ancillas. This completes the proof of the current theorem.

□

[1] E. P. Wigner. *Remarks on the Mind-Body Question*, pages 171–184. Indiana University Press, 1967. https://link.springer.com/content/pdf/10.1007/978-3-642-78374-6_20.pdf.

[2] Daniela Frauchiger and Renato Renner. Quantum theory cannot consistently describe the use of itself. *Nature Communications*, 9:3711, 2018. https://doi.org/10.1038/s41467-018-05739-8.

[3] Časlav Brukner. A no-go theorem for observer-independent facts. *Entropy*, 20(5), 2018. https://www.mdpi.com/1099-4300/20/5/350.

[4] 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. *Nature Physics*, 16(12):1199–1205, 2020. https://doi.org/10.1038/s41567-020-0990-x.

[5] Nuriya Nurgalieva and Lídia del Rio. Inadequacy of modal logic in quantum settings. *In Proceedings QPL 2018, EPTCS 287*, pages pp. 267–297, 2019. https://arxiv.org/abs/1804.01106v2.

[6] Eric G. Cavalcanti and Howard M. Wiseman. Implications of local friendliness violation for quantum causality. *Entropy*, 23(8), 2021. https://www.mdpi.com/1099-4300/23/8/925.

[7] Yìlè Yīng, Marina Maciel Ansanelli, Andrea Di Biagio, Elie Wolfe, and Eric Gama Cavalcanti. Relating Wigner's Friend scenarios to nonclassical causal compatibility, monogamy relations, and fine tuning, 2023. https://arxiv.org/abs/2309.12987.

[8] Matthew Pusey. An inconsistent friend. *Nature Communications*, 2018. https://www.nature.com/articles/s41567-018-0293-7.

[9] Nuriya Nurgalieva and Renato Renner. Testing quantum theory with thought experiments. *Contemporary Physics*, 61(3):193–216, 2020. https://doi.org/10.1080/00107514.2021.1880075.

[10] Joseph M. Renes. Consistency in the description of quantum measurement: Quantum theory can consistently describe the use of itself, 2021. https://arxiv.org/abs/2107.02193.

[11] Marcelo Losada, Roberto Laura, and Olimpia Lombardi. Frauchiger-Renner argument and quantum histories. *Phys. Rev. A*, 100:052114, Nov 2019. https://link.aps.org/doi/10.1103/PhysRevA.100.052114.

[12] Alexios P. Polychronakos. Quantum mechanical rules for observed observers and the consistency of quantum theory, 2022. https://arxiv.org/abs/2202.04203.

[13] Scott Aaronson. It's hard to think when someone hadamards your brain. https://scottaaronson.blog/?p=3975, 2018.

[14] Richard Healey. Quantum theory and the limits of objectivity. *Foundations of Physics*, 48(11):1568–1589, 2018. https://doi.org/10.1007/s10701-018-0216-6.

[15] Mateus Araújo. The flaw in Frauchiger and Renner's argument. https://mateusaraujo.info/2018/10/24/the-flaw-in-frauchiger-and-renners-argument/, 2018.

[16] Anthony Sudbery. The hidden assumptions of Frauchiger and Renner, 2019. https://arxiv.org/abs/1905.13248.

[17] Varun Narasimhachar. Agents governed by quantum mechanics can use it intersubjectively and consistently, 2020. https://arxiv.org/abs/2010.01167.

[18] V. Vilasini and Mischa P. Woods. In preparation.

[19] Jonathan Barrett. Information processing in generalized probabilistic theories, 2006. https://arxiv.org/abs/quant-ph/0508211.

[20] Giulio Chiribella, Giacomo Mauro D'Ariano, and Paolo Perinotti. Probabilistic theories with purification. *Physical Review A*, 81(6), 2010. http://dx.doi.org/10.1103/PhysRevA.81.062348.

[21] Bob Coecke and Aleks Kissinger. Categorical quantum mechanics I: Causal quantum processes, 2016. https://arxiv.org/abs/1510.05468.

[22] David Deutsch. Quantum theory as a universal physical theory. *International Journal of Theoretical Physics*, 24(1):1–41, 1985. http://dx.doi.org/10.1007/BF00670071.

[23] V. Vilasini and Renato Renner. Embedding cyclic information-theoretic structures in acyclic space-times: No-go results for indefinite causality. *Phys. Rev. A*, 110:022227, 2024. https://link.aps.org/doi/10.1103/PhysRevA.110.022227.

[24] V. Vilasini and Renato Renner. Fundamental limits for realizing quantum processes in spacetime. *Phys. Rev. Lett.*, 133:080201, 2024. https://link.aps.org/doi/10.1103/PhysRevLett.133.080201.

[25] Lídia del Rio and Renato Renner. Reply to: Quantum mechanical rules for observed observers and the consistency of quantum theory. *Nature Communications*, 15(1):3024, 2024. https://doi.org/10.1038/s41467-024-47172-0.

[26] Robert B. Griffiths. Consistent histories and the interpretation of quantum mechanics. *Journal of Statistical Physics*, 36:219–272, 1984. https://doi.org/10.1007/BF01015734.

[27] Robert B. Griffiths. The Consistent Histories Approach to Quantum Mechanics. The Stanford Encyclopedia of Philosophy (Summer 2024 Edition), Edward N. Zalta and Uri Nodelman (eds.). https://plato.stanford.edu/entries/qm-consistent-histories/#Bib.

[28] Marek Żukowski and Marcin Markiewicz. Physics and metaphysics of Wigner's friends: Even performed premeasurements have no results. *Phys. Rev. Lett.*, 126:130402, 2021. https://link.aps.org/doi/10.1103/PhysRevLett.126.130402.

[29] Judea Pearl. Causality: Models, reasoning, and inference. *Second edition, Cambridge University Press*, 2009.

[30] Joe Henson, Raymond Lal, and Matthew F Pusey. Theory-independent limits on correlations from generalized bayesian networks. *New Journal of Physics*, 16(11):p. 113043, 2014. https://iopscience.iop.org/article/10.1088/1367-2630/16/11/113043.

[31] Jonathan Barrett, Robin Lorenz, and Ognyan Oreshkov. Quantum causal models, 2020. https://arxiv.org/abs/1906.10726.

[32] Lucien Hardy. Probability Theories with Dynamic Causal Structure: A New Framework for Quantum Gravity, 2005. http://arxiv.org/abs/gr-qc/0509120.

[33] Ognyan Oreshkov, Fabio Costa, and Časlav Brukner. Quantum correlations with no causal order. *Nature Communications*, 3:1092, 2012. https://www.nature.com/articles/ncomms2076.

[34] Giulio Chiribella, Giacomo Mauro D'Ariano, Paolo Perinotti, and Benoit Valiron. Quantum computations without definite causal structure. *Physical Review A*, 88(2):022318, 2013. https://link.aps.org/doi/10.1103/PhysRevA.88.022318.

[35] Jonathan Barrett, Robin Lorenz, and Ognyan Oreshkov. Cyclic quantum causal models. *Nature Communications*, 12(1), 2021. http://dx.doi.org/10.1038/s41467-020-20456-x.

[36] V Vilasini, Nuriya Nurgalieva, and Lídia del Rio. Multi-agent paradoxes beyond quantum theory. *New Journal of Physics*, 21(11):113028, 2019. https://doi.org/10.1088/1367-2630/ab4fc4.

[37] S. Popescu and D. Rohrlich. Quantum nonlocality as an axiom. *Foundations of Physics*, 24:379–385, 1994. https://doi.org/10.1007/BF02058098.

[38] Nick Ormrod, V. Vilasini, and Jonathan Barrett. Which theories have a measurement problem?, 2023. https://arxiv.org/abs/2303.03353.

[39] Nuriya Nurgalieva and V. Vilasini. Any theory that admits an extended Wigner's Friend-type paradox is logically contextual. In preparation based on talk at Quantum Physics and Logic 2023, and unpublished results included in NN's PhD Thesis https://www.research-collection.ethz.ch/handle/20.500.11850/649851.

[40] Laurens Walleghem, Rafael Wagner, Yìlè Yīng, and David Schmid. Extended Wigner's friend paradoxes do not require nonlocal correlations, 2024. https://arxiv.org/abs/2310.06976.

[41] Asher Peres. *Quantum Theory: Concepts and Methods*, volume 72. Kluwer Academic Publishers, New York, 1993. Fundamental Theories of Physics.

[42] 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 Journal of Physics*, 17(3):p. 33002, 2015. https://iopscience.iop.org/article/10.1088/1367-2630/17/3/033002.

[43] Lucien Hardy. Nonlocality for two particles without inequalities for almost all entangled states. *Phys. Rev. Lett.*, 71:1665–1668, 1993. https://link.aps.org/doi/10.1103/PhysRevLett.71.1665.

[44] 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', 2018. https://arxiv.org/abs/1810.10917.

[45] Samson Abramsky, Rui Soares Barbosa, Kohei Kishida, Raymond Lal, and Shane Mansfield. Contextuality, Cohomology and Paradox. *24th EACSL Annual Conference on Computer Science Logic (CSL 2015)*, 41:Ed. Stephan Kreutzer, pp. 211–228, 2015. https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.CSL.2015.211.

[46] Armando Relaño. Decoherence allows quantum theory to describe the use of itself, 2018. https://arxiv.org/abs/1810.07065.

[47] Armando Relaño. Decoherence framework for Wigner's-friend experiments. *Phys. Rev. A*, 101:032107, 2020. https://link.aps.org/doi/10.1103/PhysRevA.101.032107.

[48] R. E. Kastner. Unitary-only quantum theory cannot consistently describe the use of itself: On the Frauchiger–Renner paradox. *Foundations of Physics*, 50(5):441–456, 2020. https://doi.org/10.1007/s10701-020-00336-6.

[49] Andrea Di Biagio and Carlo Rovelli. Stable facts, relative facts. *Foundations of Physics*, 51(1):30, 2021. https://doi.org/10.1007/s10701-021-00429-w.

[50] Sebastian Fortin and Olimpia Lombardi. Wigner and his many friends: A new no-go result? 2019. https://arxiv.org/abs/1904.07412.

[51] David Schmid, Yìlè Yīng, and Matthew Leifer. A review and analysis of six extended Wigner's friend arguments. https://arxiv.org/abs/2308.16220, 2023.

[52] Carlo Rovelli. Relational quantum mechanics. *International Journal of Theoretical Physics*, 35(8):1637–1678, 1996. https://doi.org/10.1007/BF02302261.

[53] Carlton M. Caves, Christopher A. Fuchs, and Rüdiger Schack. Quantum probabilities as Bayesian probabilities. *Phys. Rev. A*, 65:022305, 2002. https://link.aps.org/doi/10.1103/PhysRevA.65.022305.

[54] Christopher A. Fuchs, N. David Mermin, and Rüdiger Schack. An introduction to qbism with an application to the locality of quantum mechanics. *American Journal of Physics*, 82(8):749–754, 2014.

[55] Stuart Samuel. The Frauchiger-Renner gedanken experiment: an interesting laboratory for exploring some topics in quantum mechanics, 2022. https://arxiv.org/abs/2208.00060.

[56] Hugh Everett. "Relative state" formulation of quantum mechanics. *Rev. Mod. Phys.*, 29:454–462, 1957. https://link.aps.org/doi/10.1103/RevModPhys.29.454.

[57] Robert R. Tucci. Factorization of quantum density matrices according to Bayesian and Markov networks, 2007. https://arxiv.org/abs/quant-ph/0701201.