


Addressing some common objections to generalized noncontextuality

David Schmid^{1,*}, John H. Selby^{1,†} and Robert W. Spekkens^{2,‡}

¹*International Centre for Theory of Quantum Technologies, University of Gdańsk, 80-308 Gdańsk, Poland*

²*Perimeter Institute for Theoretical Physics, 31 Caroline Street North, Waterloo, Ontario, Canada N2L 2Y5*

 (Received 5 August 2023; revised 29 January 2024; accepted 31 January 2024; published 20 February 2024)

When should a given operational phenomenology be deemed to admit of a classical explanation? When it can be realized in a generalized-noncontextual ontological model. The case for answering the question in this fashion has been made in many previous works and motivates research on the notion of generalized noncontextuality. Many criticisms and concerns have been raised, however, regarding the definition of this notion and of the possibility of testing it experimentally. In this work, we respond to some of the most common of these objections. One such objection is that the existence of a classical record of which laboratory procedure was actually performed in each run of an experiment implies that the operational equivalence relations that are a necessary ingredient of any proof of the failure of noncontextuality do not hold, and consequently that conclusions of nonclassicality based on these equivalences are mistaken. We explain why this concern is unfounded. Our response affords the opportunity for us to clarify certain facts about generalized noncontextuality, such as the possibility of having proofs of its failure based on a consideration of the subsystem structure of composite systems. Similarly, through our responses to each of the other objections, we elucidate some under-appreciated facts about the notion of generalized noncontextuality and experimental tests thereof.

DOI: [10.1103/PhysRevA.109.022228](https://doi.org/10.1103/PhysRevA.109.022228)

I. INTRODUCTION

The notion of generalized noncontextuality was introduced in Ref. [1] as an extension of Kochen-Specker noncontextuality [2]. Realizability by a generalized-noncontextual ontological model provides a notion of classical explainability for operational phenomena. Consequently, demonstrating that a given experiment *cannot* be explained within any generalized-noncontextual ontological model constitutes a rigorous proof of nonclassicality. Many previous works have provided arguments for why generalized noncontextuality is a gold standard notion of classical explainability; see, for instance, Ref. [3] or the introductions of Refs. [4,5]. We touch on some of these arguments in passing in this work.

Our aim here, however, is to collect and respond to a number of *objections* that have been raised against the notion of generalized noncontextuality, including challenges to its motivations, its consistency, and its experimental testability. We also elaborate on a number of other conceptual points that have the potential to be misunderstood, or points that are known to some experts but which we feel deserve wider recognition.

Arguably, the most interesting analysis provided in this paper is the one in Sec. III, where we address the claim that proofs of contextuality are undermined by the existence of classical records of which operational procedures were carried out in each run of an experiment. We show that, contrary

to this claim, one can correctly assess the noncontextual-realizability of the operational statistics whether or not such records exist. Along the way, we demonstrate new possibilities for proving the failure of noncontextuality in scenarios with composite systems.

Some topics that are *not* part of the scope of this article include (i) providing an introduction to noncontextuality and the methods for testing or characterizing it, (ii) providing an account of the arguments in favor of defining classical explainability of operational statistics in terms of realizability by a generalized-noncontextual ontological model, and (iii) discussing arguments concerning the relative merit of generalized noncontextuality and Kochen-Specker noncontextuality. We refer the reader to earlier works for these topics.

This paper assumes basic familiarity with the notions of operational theories, ontological models, and generalized noncontextuality. Where possible, we focus on the simpler case of prepare-measure scenarios, although most of what is said can be generalized to scenarios with more general compositional structure [6]. Henceforth, the term “noncontextual” will be taken to refer to the notion of generalized noncontextuality introduced in Ref. [1].

II. PRELIMINARIES

It is useful to distinguish two perspectives on witnessing the failure of generalized noncontextuality, which we refer to as *algebraic* and *geometric*. They provide two different ways of conceptualizing the constraints on the ontological model implied by operational identities under the assumption of noncontextuality. In the algebraic approach, one seeks to determine whether one can represent each operational state

*david Schmid10@gmail.com

†john.h.selby@gmail.com

‡rspekkens@perimeterinstitute.ca

by a probability distribution on the ontic state space, and each operational effect by a response function on the ontic state space, while respecting the identities that hold among these. In the geometric approach, by contrast, one conceptualizes the identities holding among the operational states as stipulating the geometric shape of the convex hull of the states and similarly for the operational effects, and the question of ontological representability is expressed as a particular embedding for these geometric shapes.

The distinction should be understood in roughly the same way as the distinction between algebraic and geometric proofs of the Kochen-Specker theorem,¹ although in the case of Kochen-Specker, the arena for the geometric conditions is Hilbert space whereas for generalized noncontextuality it is the vector space of Hermitian operators [or, more generally, the vector space of generalized probabilistic theory (GPT) states and effects]. Just as any algebraic proof of the Kochen-Specker theorem can be translated into a geometric proof and vice versa, so too can any algebraic approach to witnessing the failure of generalized noncontextuality be translated into the geometric approach and vice versa.

Although the difference between the approaches is a cosmetic one, sometimes one perspective or the other is more insightful or simple, so we recap both approaches in the next section. Of particular relevance to this work is the fact that, as we will see, the geometric perspective (which is the newer of the two) will often be useful for making *especially* clear how some past concerns about generalized noncontextuality are unfounded. Indeed, it seems likely to us that if this perspective had been adopted first, then many of these objections would never have arisen in the first place.

We here focus on tests of generalized noncontextuality in prepare-measure scenarios. The generalization to arbitrary compositional scenarios can be found in Ref. [6].

For a comprehensive introduction to noncontextuality (according to both perspectives we discuss), we refer the reader to the series of three lectures at [7–9] (and references therein).

A. Algebraic approach to witnessing the failure of generalized noncontextuality

The algebraic approach was the first to be adopted for witnessing the failure of generalized noncontextuality [1] and so is the more widely known of the two. In this approach, the relevant input data to the analysis are operational identities—typically, linear constraints among the states and among the

measurements. One uses these to derive noncontextuality inequalities, whose violation demonstrates that the operational predictions of the scenario cannot be reproduced by a noncontextual ontological model.

In the simplest experiment of interest, one implements a set of preparation procedures and a set of measurement procedures and one records the outcome statistics observed for each pairing. An *operational state* is an operational equivalence class of preparation procedures, where two preparation procedures are defined to be operationally equivalent if they generate the same statistics for all possible measurements. An *operational effect* is an operational equivalence class of measurement-outcome pairs, where two such pairs are operationally equivalent if they are assigned the same probability by all preparation procedures. The operational states generally satisfy nontrivial identities, termed *operational identities*, as do the operational effects. A common form of such an identity for operational states is a linear dependence relation:

$$\sum_{x \in X} \alpha_x \mathbf{s}_x = 0, \quad (1)$$

where $\alpha_x \in \mathbb{R}$ and \mathbf{s}_x is an operational state, represented as a vector in a GPT [10–12]. In the case of quantum theory, these are simply representations of density operators in the real vector space of Hermitian operators, such as the Bloch vectors representing the density operators of a qubit. We henceforth make frequent use of the GPT representation, and so we often refer to operational states as *GPT states* and operational effects as *GPT effects*.

An example of a circumstance implying a relation of the form of Eq. (1) is when a convex mixture of two GPT states is equal to a third GPT state. Operational identities also hold among the GPT effects. These identities can often be inferred by how a given state or effect is implemented (e.g., as a convex mixture of two others). They can also be inferred from a tomographic characterization of the GPT states and GPT effects. Finally, they can additionally be inferred from principles, such as no-signaling, or the absence of retrocausation. Demanding that these identities are also respected by the ontological representations of the states and effects implies constraints on the outcome statistics, typically in the form of inequalities known as noncontextuality inequalities.

B. Geometric approach to witnessing the failure of generalized noncontextuality

The second perspective on noncontextuality is relatively recent. In this approach, the relevant input data to the analysis are a set of states and a set of measurement effects, as represented in some generalized probabilistic theory [10–12]. One then tests whether these can be embedded in a simplex and its dual (such that the probabilistic predictions are preserved); if such an embedding does not exist, this demonstrates that the operational statistics for that scenario cannot be reproduced within a noncontextual ontological model.

More precisely, this approach relies on the fact that operational theories that are noncontextual are associated with generalized probabilistic theories that are simplex-embeddable [6,13]. A generalized probabilistic theory, or GPT, is simplex-embeddable if its state space linearly embeds

¹Algebraic proofs of the Kochen-Specker theorem proceed by considering a set of Hermitian operators (observables) and demonstrating that the functional relations that these satisfy cannot be satisfied by a set of classical variables when the value assigned to the variable representing a given observable is independent of what other observables are measured together with it. Geometric proofs of the Kochen-Specker theorem, on the other hand, consider the orthogonality relations holding among rays in Hilbert space describing outcomes of a set of rank-1 projective measurements, and demonstrate that these rays cannot be assigned values 0 or 1 in such a way that a single element of every orthogonal set is assigned value 1, when the value assigned to a ray must be assigned independently of which orthogonal basis it is considered a part.

in a simplex and its effect space linearly embeds in the dual to that simplex, in such a way that the probabilities it predicts are unchanged. One can apply this approach to the study of particular scenarios and experiments as well. One simply obtains a characterization of the GPT states and effects realized in the experiment, termed an *accessible GPT fragment* [14]. These characterizations provide inner bounds on the full GPT state space and the full GPT effect space respectively, such that if the accessible GPT fragment realized in the experiment is not simplex-embeddable, then one can conclude that the GPT describing the system is not simplex-embeddable either.

Thus, in this approach, one determines if a theory or experiment is consistent with the principle of noncontextuality by testing whether the GPT representation of that theory or experiment satisfies a geometric criterion (simplex-embeddability). In this way, one need not consider operational identities as algebraic equations that in turn imply specific noncontextuality inequalities which one tests. Rather, one can think of the operational identities as constraints on the geometry of the state space. For instance, the full set of operational identities holding among a set of states is simply a description of the geometry of the convex hull of those states.

As noted above, although one *could* try to translate an analysis in one perspective to the other, one or the other approach will sometimes be more insightful.

C. Operational identities involving subsystems

Although the most commonly studied operational identities are of the form of Eq. (1), there are other types that can be leveraged for proving the failure of noncontextuality. This was clearly stated even in the first paper on generalized noncontextuality, which noted (as just one other example) that distinct ways of purifying a given quantum state correspond to distinct but operationally equivalent preparation procedures in that state’s operational equivalence class [1]. For example, suppose that two GPT states on system A are defined as

$$\begin{aligned} \mathbf{s}_A^{(1)} &= \text{tr}_B[\mathbf{s}_{AB}^{(1)}], \\ \mathbf{s}_A^{(2)} &= \text{tr}_B[\mathbf{s}_{AB}^{(2)}], \end{aligned} \quad (2)$$

where tr_B is shorthand for the transformation $\mathbf{s}_{AB} \mapsto \mathbf{u}_B \cdot \mathbf{s}_{AB}$, with \mathbf{u}_B the unit effect for the system B . If it is the case that

$$\mathbf{s}_A^{(1)} = \mathbf{s}_A^{(2)}, \quad (3)$$

then this describes a valid operational identity, around which one could construct proofs of noncontextuality. One can also consider more general operational identities that involve both linear combinations and partial traces, e.g.,

$$\sum_{x \in X} \alpha_x \text{tr}_B[\mathbf{s}_{AB}^{(x)}] = \sum_{y \in Y} \beta_y \text{tr}_B[\mathbf{s}_{AB}^{(y)}]. \quad (4)$$

We will not attempt to give a completely general algebraic description of the scope of operational identities one can consider; however, one can find a completely general diagrammatic description in Refs. [6,15].

Most prior derivations of noncontextuality inequalities have relied on operational identities that are given by linear combinations like those in Eq. (1)—in particular, they did not make use of subsystem structure. The only instance of a

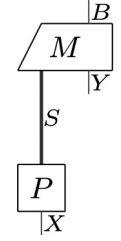


FIG. 1. The basic PM scenario.

more general operational identity that we are aware of is in Ref. [16] [see Eq. (8) and the surrounding discussion therein]. In Sec. III, we give a second example—a proof of contextuality that uses an operational identity of the form of Eq. (4). While our example is quite simple, it demonstrates that one is generally *forced* to consider operational identities of this more general form if one wishes to determine the full implications of noncontextuality.

D. Theory-agnostic tomography

A useful tool for determining the characterization of one’s experiment within a generalized probabilistic theory is theory-agnostic tomography, also known as GPT tomography [17,18]. While this tool is not required for understanding the definition of noncontextuality, theory-agnostic tomography is in many respects the ideal way of experimentally testing noncontextuality, and so it will be relevant to a number of the points we make herein. In theory-agnostic tomography, one carries out a large number of preparations and measurements on the given system, where these are chosen either randomly, or to roughly fill out an approximation of what one expects the true state and effect spaces to be. One does not assume anything *a priori* about the identity of each individual procedure (such as its GPT description), or about what GPT governs the experiment. Rather, one *extracts* (by an appropriate analysis) the GPT dimension and GPT descriptions of each state and effect in the experiment from the observed data. These realized GPT vectors then constitute inner approximations of the true GPT state and effect spaces. One can then use this information, for example, to assess whether the experiment is simplex-embeddable (classically explainable).

E. A standard proof of the failure of noncontextuality

Consider a prepare-measure scenario, depicted in Fig. 1, defined by a set of GPT states indexed by the set X , $\{\mathbf{s}_x\}_{x \in X}$, and a set of GPT measurements, indexed by the set Y and where for each $y \in Y$, the GPT measurement is described by the set of effects $\{\mathbf{e}_{b|y}\}_{b \in B}$. Imagine that the states satisfy some operational identities indexed by j ,

$$\forall j : \sum_{x \in X} \alpha_x^{(j)} \mathbf{s}_x = 0, \quad (5)$$

for $\alpha_x^{(j)} \in \mathbb{R}$, so that the assumption of generalized noncontextuality implies linear constraints of the same form on the associated epistemic states:

$$\forall j : \sum_{x \in X} \alpha_x^{(j)} \mu_x(\lambda) = 0 \quad \forall \lambda \in \Lambda, \quad (6)$$

where μ_x is the epistemic state associated with the GPT state \mathbf{s}_x . Similarly, we can imagine that the effects satisfy some operational identities indexed by k ,

$$\forall k : \sum_{b \in B, y \in Y} \beta_{b,y}^{(k)} \mathbf{e}_{b|y} = 0, \quad (7)$$

for $\beta_{b,y}^{(k)} \in \mathbb{R}$, so that the assumption of generalized noncontextuality implies a linear constraint of the same form on the associated response functions:

$$\forall k : \sum_{b \in B, y \in Y} \beta_{b,y}^{(k)} \xi_{b|y}(\lambda) = 0 \quad \forall \lambda \in \Lambda, \quad (8)$$

where $\xi_{b|y}$ is the response function associated with the GPT effect $\mathbf{e}_{b|y}$. Imagine moreover that one has derived noncontextuality inequalities from these operational identities, and that these have been violated by the observed statistics in the experiment. In this case, one has found a proof of the failure of noncontextuality in that prepare-measure scenario.

III. THE LABORATORY NOTEBOOK OBJECTION

A challenge that is sometimes made to the analysis given in the previous section is the following: The choice of preparation in the experiment is typically recorded in the experimenter's lab notebook. (Indeed, such a recording is *necessary* if the experimenter hopes to compute the statistics on which noncontextuality inequalities are tested.) In particular, if the experiment includes two preparation procedures that are distinct but operationally equivalent, then which of these is implemented in a given run of the experiment is indicated in the lab notebook. Consequently, there *does* exist a measurement that distinguishes the two, namely, the measurement that reveals the physical state of the lab notebook.² According to this argument, therefore, no two preparation procedures are ever found to be operationally equivalent. Because the assumption of generalized noncontextuality is an engine that turns operational equivalence relations among procedures into constraints on how they are represented in the ontological model, if there are no such equivalence relations, one obtains no constraints. Hence, there is no opportunity to derive noncontextuality inequalities and thus no opportunity to discover a failure of noncontextuality. We refer to this challenge as the “lab notebook objection” to generalized noncontextuality.

The first key fact that this objection misses is this: operational theories incorporate a notion of a physical system, which is treated as a primitive notion on which an individuating principle can be based. Preparation and measurement procedures are specific to a system, and consequently operational identities are evaluated *relative to a system*. Thus, for instance, two preparation procedures on a system S are deemed to be operationally equivalent if they yield the same

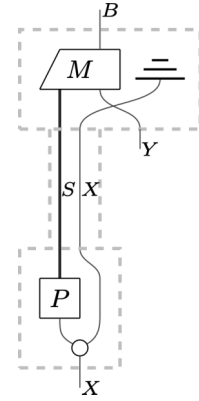


FIG. 2. The scenario with the lab notebook modeled as a physical system (denoted X) which is on the same footing as S .

statistics for all measurements *on* S . (For more general compositional scenarios, causal structure provides the individuating principle for procedures—see the discussion in Sec. VB.)

The system that is being prepared and measured in a given experiment can be conceptualized as the thing that acts as a causal intermediary between the preparation device and the measurement device. In an experiment wherein the preparations and measurements relate to the polarization of a photon, for instance, this degree of freedom constitutes S , the causal intermediary. The lab notebook in such an experiment is explicitly presumed *not* to act as such a causal intermediary. To imagine that the causal influence from the choice of preparation procedure to the outcome of the measurement is not mediated by the polarization degree of freedom of the photon, but rather by some physical records of how the preparation device was implemented is a radical and *a priori* rather implausible hypothesis about the causal structure of the experiment. To put it more strongly: as long as one grants that the lab notebook is an independent physical system from S , assessments of nonclassicality for system S alone (as opposed to assessments for the joint system comprised of *the lab notebook together with* S) are based on operational equivalences that are *defined* relative to measurements on S alone.

Still, a stubborn skeptic might remain concerned about the case where one rather chooses to study the nonclassicality of the joint system defined by the lab notebook together with system S . Indeed, it has sometimes been claimed that in this case, one reaches a different verdict regarding the nonclassicality of the system—one that is inconsistent with the verdict obtained when the system S alone is taken to be the system of interest.

To respond to this, it is useful to first recast the lab notebook objection into the language of GPT states and the operational identities that hold among them. Imagine that the choice of preparation is copied and viewed as a physical system X on equal footing with the system S , as shown in Fig. 2. System X plays the role of the lab notebook; its value constitutes the classical record of which preparation was performed. We denote a state of knowledge wherein one has certainty that X takes value x by δ_x . The GPT states on the composite system SX , therefore, are given by

$$\{\mathbf{s}_x \otimes \delta_x\}_{x \in X}. \quad (9)$$

²Even if the experimenter does not take care to record which procedure was implemented, the environmental degrees of freedom within the laboratory are likely to carry away information about which it was (e.g., in the precise pattern of light rays scattered off the laboratory apparatus), and therefore these are likely to encode a record of which it was.

These states are all linearly independent as GPT vectors, since $\sum_{x \in X} \gamma_x \mathbf{s}_x \otimes \delta_x = 0$ if and only if $\gamma_x = 0$ for all $x \in X$. The lab notebook objection is then expressible as follows: if we consider the operational states on the system S , then they satisfy nontrivial linear dependence relations of the form of Eq. (1), but if we include the lab notebook X in our analysis, then the operational states of the system and notebook are those of Eq. (9), which are *linearly independent*, and hence do not satisfy any nontrivial relation of the form of Eq. (1). In other words, including the notebook, the sceptic claims, leads to there being no nontrivial operational identities among the states. It is well known that one cannot prove the failure of noncontextuality in a prepare-measure scenario without making use of some nontrivial operational identities among the states [1]. Hence, the argument goes, one can always find a noncontextual model for the scenario, viewed as an experiment on SX .

Thus, proponents of the lab notebook argument claim that one reaches different verdicts for the exact same experiment, depending on whether one includes system X as a causal mediary in one's analysis.

However, this is not correct. When this scenario is correctly analyzed as an experiment on SX as depicted in Fig. 2, one gets the same answer as in the original analysis—the experiment *is* a proof of nonclassicality, even when conceptualized in this way.

The mistake arises from the belief that the linear independence of the states $\{\mathbf{s}_x \otimes \delta_x\}_{x \in X}$ implies that there are no operational identities among the GPT states on SX . As we noted in Sec. II C, not all operational identities take the form of bare linear dependence relations. (This realization came in part from discussions with Sainz, Wolfe, and Kunjwal.) Indeed, if the GPT states on S satisfy the operational identities in Eq. (5), then the GPT states on SX satisfy the operational identities

$$\forall j : \sum_{x \in X} \alpha_x^{(j)} \text{tr}_X(\mathbf{s}_x \otimes \delta_x) = 0. \quad (10)$$

On the basis of this operational identity, one can derive a noncontextuality inequality that is violated in this scenario—namely, the exact same inequality that one arrived at via the original analysis of the scenario (the one which did not treat X as a causal mediary on par with S).

Explicitly: an ontological model for the composite system SX posits³ an ontic state space $\Lambda_S \times \Lambda_X$ and represents each of the states $\mathbf{s}_x \otimes \delta_x$ by some probability distribution $\mu_x(\lambda_S, \lambda_X)$. The constraints implied by generalized noncontextuality together with the operational equivalence in Eq. (10)

is

$$\forall j : \sum_{x \in X} \alpha_x^{(j)} \sum_{\lambda_X \in \Lambda_X} \mu_x(\lambda_S, \lambda_X) = 0 \quad (11)$$

where we have made use of the fact that tr_X is represented in the ontological model by marginalization over Λ_X . But $\sum_{\lambda_X \in \Lambda_X} \mu_x(\lambda_S, \lambda_X) = \mu_x(\lambda_S)$, where $\mu_x(\lambda_S)$ is the distribution representing \mathbf{s}_x , so that Eq. (11) is simply

$$\forall j : \sum_{x \in X} \alpha_x^{(j)} \mu_x(\lambda_S) = 0, \quad (12)$$

which is simply Eq. (6), the constraint one obtained in the original analysis—which, by assumption, leads to a noncontextuality inequality that is violated by the observed statistics in the experiment.

In short, *whether or not* one chooses to treat the lab notebook as a dynamical system, one reaches the same verdict: the experiment in question does not admit of a noncontextual explanation. This was missed by proponents of the lab notebook objection because the full scope of possible operational identities was not recognized.

As a final clarifying remark, we note that the assumptions underlying the use of operational identities in noncontextuality arguments are exactly the same as the assumptions underlying the use of the Bloch sphere as a representation of a qubit. Consider the case of a single qubit, as represented by the Bloch ball. By definition, the points in the Bloch ball describe operational equivalence classes of preparation procedures, where each point contains all and only the information needed to predict the statistics of all measurements on the qubit. As a concrete example, the center point of the Bloch ball represents many different ways to prepare the maximally mixed state, such as taking an equal mixture of $|0\rangle$ and $|1\rangle$ or an equal mixture of $|+\rangle$ and $|-\rangle$. In any real experiment where one prepares states of the qubit, there will exist some records of what preparation was performed on the qubit. (In our example, this would be a record of whether Z or X eigenstates were prepared in a given run.)

If one chooses to represent the joint state of the qubit together with these records, the density matrices one so obtains will be linearly independent, and consequently will form a simplex rather than a Bloch ball. This in no way undermines the fact that the preparations of the qubit satisfy operational equivalences, nor does it undermine the validity of the Bloch representation. When one computes the operational states of the system alone, after tracing out the lab notebook, one recovers the Bloch ball.

To make sense of the Bloch sphere representation—just as is needed to make sense of generalized noncontextuality—one must assume that one can meaningfully single out a specific degree of freedom, and perform measurements on it (and it alone). Quantum physicists (both theorists and experimentalists) know how to study single systems in isolation and how to characterize the GPT governing some such system S ; recall, for example, the discussion of theory-agnostic tomography in Sec. II D. The existence of any number of records or copies of this information, or of details about how this information was obtained, is irrelevant to this fact. And once one has a characterization of the GPT governing the system of interest, determining whether the system is classically explainable or

³The fact that we take the ontic state space to be a Cartesian product of the ontic state spaces for S and for X could be viewed as a consequence of diagram preservation [6]. It also follows immediately from the causal structure assumed in the lab notebook argument—that system X is a system whose role is to encode perfect classical information about which preparation was performed. (In fact, one can moreover conclude from the causal structure that Λ_X is isomorphic to the set of possible values of X , and that $\mu_x(\lambda_S, \lambda_X) = \mu_x(\lambda_S) \otimes \delta_{\lambda_X, x}$, but the argument does not need this specificity.)

not is simply a matter of testing simplex embeddability on it (or deriving and testing noncontextuality inequalities for it).

IV. OTHER COMMON OBJECTIONS

We now reply to a number of other objections to the notion of generalized noncontextuality.

A. The physical-mixtures objection

Another common objection (which is close in spirit to the lab notebook objection) is that the existence of classical records about what procedure was implemented precludes the possibility of defining or physically implementing mixtures of different laboratory procedures. Only if this record is somehow fundamentally erased from existence, the argument goes, could one hope to have implemented a true mixture of procedures.

Perhaps the simplest response to this objection is to note that some proofs of noncontextuality—for example, those using the simplex-embedding approach—make no explicit reference to mixtures of preparations (or indeed even to mixed states). Similarly, experimental tests of noncontextuality within this approach (e.g., using theory-agnostic tomography) do not require one to implement any particular mixtures of given states. One can determine whether a given theory (or experiment) is classical or nonclassical based solely on the geometry of the state and effect spaces.

However, there are insights to be learned by providing a more thorough analysis of this objection. It arises from a misunderstanding regarding the notion of a GPT state vector (or of a density operator, in quantum theory). Indeed, this is the same misunderstanding that sometimes leads to the claim (discussed in Sec. III) that one must include the lab notebook X as a physical system in one's analysis. A GPT state vector is an equivalence class of preparation procedures, where the equivalence relation is defined relative to all measurements *on a given system*. And one can define and experimentally characterize GPT state vectors, regardless of the existence of any number of records or copies of information pertaining to which laboratory procedures were used to generate them. (See also our comments at the end of Sec. III.)

In addition, this objection misses the fact that mixtures appearing in noncontextuality arguments can be (and should be) viewed as inferential rather than physical [15,19]. That is, they need only describe the knowledge of agents who are reasoning about the system. One need not imagine a dice-rolling procedure implemented physically to justify the applicability of a probabilistic mixture. Based on whatever actual procedures one happens to have implemented, one can always leverage classical probability theory to reason about any hypothetical ensemble of procedures, where each of the actual procedures appears with some particular relative frequency in the ensemble. Mixed states need not arise in any other capacity in noncontextuality scenarios.

A final related confusion concerns the distinction between proper and improper mixtures. (Recall that a proper mixture is

defined as a state of classical uncertainty about what quantum state describes a given system, whereas an improper mixture is defined as the marginal of some entangled bipartite pure state.) It is sometimes suggested that for a given mixed state, noncontextuality arguments presume that it is realized as a proper mixture and that this is somehow problematic in the sense that noncontextuality arguments are silent about improper mixtures. But neither of these is the case: rather, the details of how one prepares a given mixed state are irrelevant because all such preparation procedures are operationally equivalent. In other words, only the set of GPT states is relevant for questions of noncontextual realizability, and whether a given GPT state is realized as a proper or improper mixture is simply part of the preparation context and hence irrelevant to the ontological representation.

B. The device-dependence objection

Another frequent challenge to the notion of generalized noncontextuality rests on the fact that a classical computer can simulate any given set of prepare-measure statistics, even statistics that are not realizable within a noncontextual ontological model. Does the classical computer itself not then constitute a classical explanation of the statistics?

We first give a direct answer to this question, and an example to illustrate it. We then return to a deeper discussion of some key surrounding issues.

In short: whether a scenario is deemed noncontextual does not rest merely on the bare statistics, but also on the operational identities holding among the processes which generated those statistics. A classical computer simulation of an experiment fails to reproduce the operational identities that hold in the experiment, and so does not constitute a good classical explanation of the experimental statistics because it has failed to achieve a good explanation of the operational identities that are observed in the experiment.

This is best illustrated by a simple example (which was constructed in collaboration with Sainz, Wolfe, and Kunjwal). Consider an experiment with two binary inputs, the setting variables X and Y , and two binary outputs, the outcome variables A and B . Suppose the correlations between A and B conditioned on X and Y , denoted $P(AB|XY)$, achieve the maximum possible violation of a Clauser–Horne–Shimony–Holt (CHSH) inequality. That is, suppose that

$$P(AB|XY) = \frac{1}{2}([00] + [11])\delta_{XY,0} + \frac{1}{2}([01] + [10])\delta_{XY,1}, \quad (13)$$

where we have used the shorthand notation $[ab] := \delta_{A,a}\delta_{B,b}$. This is easily recognizable as the input-output correlation associated with a Popescu–Rohrlich box [20].

Now suppose that the experiment yielding these correlations is a bipartite Bell scenario, i.e., using measurements on a bipartite state. If the outcome at one wing is spacelike from the mechanism choosing the value of the setting variable at the opposite wing, then observing the correlations in Eq. (13) implies that the experiment cannot be explained by a locally causal ontological model. Even if the measurements are not spacelike separated, so that the experiment can be conceptualized as being of the prepare-measure variety, as depicted in

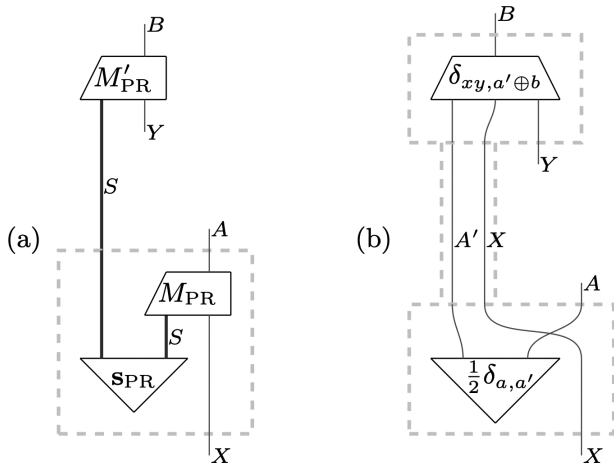


FIG. 3. Two experimental prepare-measure scenarios that achieve the correlations $P(AB|XY)$ of Eq. (13). (a) The causal mediary is a nonclassical GPT system, in which case the correlations are evidence of nonclassicality, and (b) the causal mediary is a classical system, in which case the correlations are *not* evidence of nonclassicality. The difference between experiments (a) and (b) is manifested in the operational identities that hold among the preparations of the system being transmitted. The specific GPT states and measurements in panel (a) live in the GPT known as Boxworld [11] and we follow the notation given in Eq. (8) of Ref. [21].

Fig. 3(a), then observing the correlations in Eq. (13) implies that the experiment cannot be explained by a noncontextual ontological model. In either of these scenarios, the correlations witness the impossibility of a certain type of classical explanation.

Now suppose that the experiment is again of the prepare-measure variety. This time, however, suppose that it is not a nonclassical system (i.e., a “boxworld” system) that is transmitted from the preparation to the measurement, but a classical system that encodes both X and A , as depicted in Fig. 3(b) (here A' denotes a copy of A). In this case, the experiment can still generate a conditional distribution $P(AB|XY)$ of the form of Eq. (13), but such correlations no longer witness the failure of a noncontextual ontological model and hence no longer witness the impossibility of a classical explanation. The reason that the realization of the correlations of Eq. (13) in the context of the experiment of Fig. 3(a) exhibits nonclassicality but the realization of the *same* correlations in the context of the experiment of Fig. 3(b) does not, is because the two realizations satisfy *different operational identities*. In the former case, the two effective GPT states on system S (one for each of the two possible values of X) that arise when one marginalizes over A are equal, and it is this operational identity that allows one to derive the noncontextuality inequality that is violated by the correlations of Eq. (13). In contrast, in the latter case, the two effective GPT states of the causal mediary $A'X$ are not the same (indeed, they are perfectly distinguishable, as X takes a different value in the two states), and so the noncontextuality inequality just mentioned is not a constraint on this scenario, and its violation cannot support any conclusions about noncontextuality.

We can now discuss some key surrounding issues. The objection above is primarily raised by researchers who favor the device-independent paradigm for demonstrating quantum-over-classical advantages in information processing. In a Bell scenario, the argument goes, one does not need to check any additional data to be sure that the observed statistics are nonclassical: one can check this from the observed correlations alone. It is often additionally claimed that this is a major advantage over tests of generalized noncontextuality.

However, it is not true that one does not need to check any additional data to be sure that the observed statistics are nonclassical in a Bell scenario. Rather, one must check that these statistics were generated in a particular causal structure: one where the outcomes are only connected by a common cause [22]. (This restriction on the classical simulation is typically formalized in terms of Bell’s notion of locality causality.) As such, the mere presence of Bell inequality violations in an experimental scenario is not by itself sufficient to witness nonclassicality. Similarly, the task of simulating noncontextuality inequality violations is only nontrivial if one takes into account additional empirical data: the operational identities among the operational states and effects in the experimental setup.

In Bell scenarios, this additional information (namely, whether the causal structure is one where the outcomes are only connected by a common cause) is often left implicit, which is why it is often said that one can decide if a given set of correlations is classical or nonclassical simply by examining the correlations themselves. The reason this information is typically neglected is the belief that it is quite independent of the system whose nonclassicality is being probed. Indeed, the typical way to justify such a claim—that the causal structure is one where the outcomes are only connected by a common cause—is to appeal to the theory of relativity, together with the experimental evidence that the choice of setting on each wing is spacelike separated from the outcomes at the other wing. This evidence comes in the form of distance and timing measurements, which are presumed to be quite independent from the measurements on the entangled quantum systems.

In current schemes for experimentally testing noncontextuality, by contrast, the additional information one gathers to assess classicality comes from additional preparations and measurements on the degree of freedom whose nonclassicality is being probed. In particular, one finds the best-fit GPT representations of the preparations and measurements (from which one can extract operational equivalences if one desires).

This apparent contrast might seem at first to vindicate the claim that contextuality tests have a different status than Bell tests. However, we now seek to show that the contrast is largely illusory and likely to diminish further as better tests of noncontextuality are devised.

While it is true that *current* tests of noncontextuality require a preliminary step of finding the best-fit GPT representations of one’s preparations and measurements, it is possible that a future test of noncontextuality might be found where all of the operational identities that are used can be justified on grounds that are distinct from the experimental statistics gathered for preparations and measurements on the system in question. This would parallel more closely the type of empirical evidence used to justify the applicability of local causality

in a Bell test that closes the locality loophole. Whether this possibility is realized is an important open question for researchers studying noncontextuality.⁴

Furthermore, spacelike separation of the wings of the experiment is not the only sort of evidence one can leverage to support a conclusion of nonclassicality in a Bell experiment. In other words, one can have strong evidence of nonclassicality in Bell experiments even when the locality loophole is *not* closed. For example, the statistical data accumulated in the experiment can provide evidence for nonclassicality because the classical explanations involving cause-effect relations between the wings *overfit* the data relative to explanations involving a quantum common-cause [23]. The latter sort of demonstration of nonclassicality in a Bell scenario has a close analog in contextuality scenarios.

The claim that noncontextuality tests are different from Bell tests is also undermined by the fact that the widespread claim that nonclassicality in Bell tests can be inferred *from the observed correlations alone* is not accurate. Specifically, we argue that one cannot, strictly speaking, implement a Bell test by simply taking the finite-run relative frequencies seen in the experiment and plugging these into the left-hand side of a Bell inequality.

All real-world experiments are finite-run and all finite-run statistics include fluctuations. For this reason, no real-world experiment yields the true probabilities, i.e., the relative frequencies that would be observed in an idealized limit of infinitely many samples. Nonetheless, there are certain constraints on the true probabilities that an experimentalist might know to hold. For instance, in a Bell test, if an experimenter is confident that there is spacelike separation between the wings and they are confident in the correctness of relativity theory, then they can assume, as a constraint on their estimate of the true probabilities, that it must satisfy the no-signaling condition. The finite-run relative frequencies, however, will generally *violate* the no-signaling condition simply because of statistical fluctuations.⁵ Therefore, to find an estimate of the true probabilities that respects the no-signaling condition, one cannot use the naïve procedure of taking the relative frequencies as estimates of the true probabilities.

A more methodologically sound analysis technique for experimental tests of Bell's notion of local causality (in the sense of the methodology of statistical model selection) estimates the true probabilities through a fitting procedure. For instance,

⁴If the possibility is realized, then one could test noncontextuality without first obtaining the best-fit GPT representations of the preparations and measurements. Rather, one could simply do a hypothesis test on the possibility of a noncontextual model by looking for representations of the preparations and measurements as classical distributions and response functions over the ontic state space respectively and demanding that these satisfy the constraints implied by the operational identities. If no such representations can be found that yield a good fit to the data, one has ruled out the hypothesis of a noncontextual model.

⁵Specifically, for any finite-run statistics, the relative frequencies of outcome values at Alice's wing will generally show slight differences for different settings values at Bob's wing simply because of statistical fluctuations.

one can adopt a statistical model for the hidden variable source (while assuming, without loss of generality, that the measurements respond deterministically in a prescribed manner [24]) and one can implement an optimization algorithm to find the best-fitting such model where the quality of fit is given by a measure of distance between the probability distribution that is predicted by the model and the relative frequencies that are observed in the experiment. Such an analysis of a Bell experiment was implemented in Ref. [23] and used to rule out a locally causal model via a hypothesis test (though this was not the focus of that article). Similar techniques for implementing a hypothesis test of local causality have been used to contend with the memory loophole [25] in Bell tests, as described in Refs. [26,27]. Such techniques use the raw frequencies to find an estimate of the true probabilities while satisfying certain constraints, and then evaluate the Bell inequalities on these best-fit probabilities rather than on the raw relative frequencies.⁶ This undermines the claim that such fitting procedures are unique to tests of noncontextuality and hence also the claim that they constitute a way in which noncontextuality tests are different in kind from Bell tests.

Finally, we dispute the claim that Bell tests and noncontextuality tests are contrasting because the former are theory-independent while the latter are not. In fact, tests of generalized noncontextuality, just like Bell tests, do not need to make any prior assumption of the correctness of quantum theory, nor do they need to make any prior assumption about the identity of each state or measurement used in the experiment. This is most evident in tests of generalized noncontextuality based on theory-agnostic tomography, as discussed in Sec. II D. In these tests, one extracts from the data (rather than assumes) both the dimension of the GPT vector space needed to model the system and the precise characterizations of the GPT state vectors and effect vectors that best fit the realized preparations and measurements.

Indeed, when one takes the trouble to implement the greatest possible diversity of laboratory procedures on a given system, theory-agnostic tomography can provide evidence for tomographic completeness of the realized set (i.e., that the realized GPT state and effect vectors span the true state and effect spaces, respectively) in the sense that one has the opportunity in such an experiment to *falsify* the hypothesis that some set of procedures are tomographically complete. It is this possibility of falsification that makes it clear that one is not merely *assuming* tomographic completeness but gathering evidence for it. In such a case, one can argue that evidence of nonclassicality can be reached directly from the observed data alone, with no extra assumptions. Whether or not this evidence is compelling depends on the extent to which the

⁶In short, *both* Bell tests and noncontextuality tests must engage in finding best-fit *classical* representations of preparations and measurements satisfying certain constraints. This fitting procedure is distinct from the one that arises in current tests of noncontextuality wherein one must find best-fit *GPT* representations of the preparations and measurements. This step is what defines the constraints that the classical fit must satisfy. Although the GPT-fitting step is currently unique to tests of noncontextuality, it may be possible to circumvent it, as we noted above.

experiment really had an opportunity to falsify the hypothesis and hence on one's confidence that the laboratory procedures in the experiment do in fact span the true state and effect space of the system being probed (or a valid GPT subsystem thereof, in the sense defined in Ref. [28]). This point is discussed at greater length in the introduction of Ref. [17].

In our view, the possibility that the procedures one has experimentally implemented fail to span the true state and effect space of the system or subsystem being probed is the most significant loophole for tests of noncontextuality. No matter how hard one has tried and failed to falsify the hypothesis of tomographic completeness of a given set of procedures, it may be that at some future date, a novel experiment succeeds in achieving the falsification. But there is parallel kind of loophole in a Bell test. The claim that the two laboratories in a Bell experiment are spacelike separated is also one that is based on empirical data, and no matter how much data one has accumulated in favor of this assessment, it might be falsified. This is clear if one thinks about the problem of verifying spacelike separation as a two-party cryptographic task in the presence of an adversary.⁷

Both Bell tests and contextuality tests are also theory-laden in another sense. Imagine that one is seeking to establishing spacelike separation of a pair of events that are separated by a distance d . This requires that one has timing precision of order d/c , where c is the speed of light. A skeptic may then wonder on what grounds one is confident that one's clock in fact has this kind of precision. Generally, the grounds for such confidence always refer to our understanding of how the clock works according to our best physical theories.

The point is this: the sort of evidence one can have for characterizing the causal structure and representation of experimental procedures in a noncontextuality test is not dramatically different in kind from the sort of evidence one can have for characterizing the causal structure and representation of experimental procedures in a Bell test.

C. The efficient-simulability objection

In certain circles, it is common to assume the following desideratum for a good notion of classical explainability: that a given computational process (for instance, a quantum computation) should count as classically explainable if and only if it can be efficiently simulated on a classical computer. But

⁷Suppose Alice and Bob seek to confirm that given events in their laboratories are indeed spacelike separated, while an adversary seeks to fool them into thinking these events are spacelike separated when in fact they are not. Suppose, for instance, that Alice and Bob try to synchronize their clocks by a procedure wherein they transfer light signals to one another. If the adversary adds delays to these signals, they can cause Alice and Bob to have false beliefs about what clock readings correspond to synchronization, and hence false beliefs about what events are spacelike separated. Similarly, whatever protocol Alice and Bob use for seeking to estimate the distance between their laboratories, the adversary could seek to interfere with that protocol as well. It would be interesting to try and devise a protocol that could provide a guarantee of spacelike separation (relative to some set of background assumptions) in the presence of an adversary. As far as we are aware, no proposals for such a protocol have been made to date.

it is well known that there are subtheories of quantum mechanics that are efficiently simulable on a classical computer but that still exhibit contextuality. For example, the stabilizer subtheory for qubits is efficiently simulable due to the Gottesman-Knill theorem [29], and yet is contextual, due to the possibility of realizing the Greenberger-Horne-Zeilinger (GHZ) or Mermin-square proofs of contextuality within it. (The result can be generalized to stabilizer subtheories in any even dimension [30].) Consequently those who endorse the desideratum see this as a deficiency of generalized noncontextuality as a notion of classical explainability.

However, the idea that a notion of classical explainability must reproduce the divide between efficient and inefficient classical simulability is, in our view, unmotivated. Quantum computation forms only a small subset of the scope of all physical phenomena, and there is no reason to expect that every manifestation of nonclassicality must be useful for the specific task of universal computation.

For example, consider the kind of nonclassicality arising in Bell scenarios (which we take to be nonclassicality of the common cause [31–33]). It is generally thought that this is a meaningful and interesting notion of nonclassicality. And yet, as noted in the previous section, it is only when one imposes constraints on the simulation (specifically, a constraint on its causal structure) that there is any challenge to simulating Bell inequality violations on a classical computer. In the case of a prepare-measure scenario, if one adopts generalized noncontextuality as one's notion of classical explainability, then the question of interest is whether one can simulate the experiment while respecting specific identities on the classical representations of states and specific identities on the classical representation of effects, namely, those that mirror the identities that hold among the states and effects themselves.

Different notions of classical explainability, we believe, correspond to different assumptions about what *constraints* a classical model of some operational phenomena ought to satisfy. To evaluate the merit of a given notion is to evaluate the motivations for the constraints it proposes. Conceiving of the classical model as a classical simulation, in the sense of computational complexity theory, does not alleviate the need to make such an assessment. For instance, there are many different computational complexity classes for which one can define a classical and a quantum version. What differs between these classes is *which constraints* are imposed, for instance, the spatial and temporal resources that the computation is permitted to use.

As an example, Anders and Browne [34] consider a model of computation that is a version of measurement-based quantum computation, but where the classical processor which acts on the setting and outcome variables of the measurements can only make use of gates whose Boolean output is a *linear* function of the Boolean inputs (e.g., it can implement XOR and NOT gates, but it cannot implement an AND gate). They showed that in this model of computation, if one supplements the classical linear processor with a bipartite state and local measurements that are able to achieve the algebraically maximal violation of the CHSH inequalities [i.e., a Popescu-Rohrlich (PR) box [20]], then one can implement an AND gate on the Boolean inputs to the circuit. The PR box correlations have promoted the computational power of this model from

the parity-L class to universal classical computation. Because there are ways of implementing AND gates that do not require access to Bell-inequality-violating correlations, the nonclassicality of the state and measurement resources is not witnessed by the ability of the circuit to go beyond universal classical computation, but rather by its ability to go beyond the parity-L class. What this example suggests is that a given computational architecture can be judged to witness nonclassicality if the operational statistics it generates cannot be explained by a classical model *that respects the causal structure of that architecture*. This may happen even though the computational task it achieves (such as implementing an AND gate) is only difficult to achieve classically *relative to this causal structure*. This example has been discussed in greater detail in Ref. [35]. It has also been shown that in a measurement-based model of computation where the classical processor is linear, the power of the model can be increased by correlations that exhibit *contextuality* [36].

It is also worth noting that, *a priori*, there is no reason to think that a notion of nonclassicality that was entirely motivated by questions about *computational complexity* would be able to explain advantages for other information-processing tasks, such as communication and cryptography, in particular, the known advantage that generalized contextuality implies for certain types of random access codes [37,38].

D. The parochial-equivalences objection

Another objection one often hears is that whether a pair of preparations are deemed to be operationally equivalent or not depends on what measurements one has made in a given experiment, or on what measurements can be made using current technology. If this were true, it would completely undermine generalized noncontextuality as a foundational notion of nonclassicality, since verdicts of classicality would be determined more by current technology and choices of what experiments to carry out than it would be by fundamental physics.

However, the notion of operational equivalence for preparations on a system, as defined in Ref. [1], is equivalence of statistics for all measurements that are possible *by the lights of the operational theory one is assuming*.

Since a *tomographically complete* set of measurements is one such that its statistics are sufficient to infer the statistics of any other measurement, one can define operational equivalence of preparations on a system in terms of equivalence of the statistics of *all* measurements in a set that is tomographically complete by the lights of the operational theory. In short, operational equivalence is a notion that is only defined relative to an operational theory. If one assumes the correctness of quantum theory, then whether two preparations are operationally equivalent or not is assessed relative to a set of measurements that is tomographically complete *by the lights of quantum theory*. By contrast, if one assumes that a system is governed by some other GPT, distinct from quantum theory, then operational equivalences of preparations must be assessed relative to a set of measurements that is tomographically complete by the lights of that GPT.

It is helpful to consider a thermodynamic example that is sometimes put forward to elucidate the objection, and to see

in what way it misunderstands the definition of operational equivalence.

Consider an ideal gas of particles assumed to be governed by classical Newtonian mechanics, and consider a box with two compartments separated by a divider. Let us now define two different preparation procedures on the gas. For the first preparation, the gas is prepared at a specified temperature and pressure and such that it lies entirely in the left compartment (while the right compartment is empty); then, the divider is removed so that the gas expands into the entire box. The second preparation procedure is identical, but where the gas begins in the right compartment (while the left is empty) prior to removing the barrier. We have thereby described two distinct preparation procedures, in each of which the gas ends up distributed throughout the whole box. These two preparation procedures lead to the same macroscopic thermodynamic properties (in particular, temperature and pressure) for the gas, but they correspond to different microstates. (This follows from the reversibility of Newtonian dynamics and the fact that the microstates at the initial time are different.) It follows that the two preparations are indistinguishable by any measurement of macroscopic thermodynamic properties (and perhaps even indistinguishable by any practically realizable measurement given current technology), but they are nonetheless associated with different ontological states.

Therefore, *if* this type of indistinguishability of preparations was sufficient to infer their operational equivalence, then the two preparations would be operationally equivalent but represented by distinct distributions over the ontic states, and hence we would have described an example of preparation contextuality. If this were the case, then the example would undermine the notion of generalized noncontextuality insofar as contextuality is being proposed as a notion of nonclassicality and a system governed by Newtonian mechanics ought not to be assessed as nonclassical.

But the type of indistinguishability described here is not sufficient for inferring operational equivalence, and so the thought experiment does not constitute an example of preparation contextuality.

In other words, the thought experiment only gives the appearance of undermining the notion of noncontextuality *if* one forgets that operational equivalences are defined relative to the set of all measurements that are possible in principle *by the lights of the physical theory one is assuming*. In Newtonian mechanics, the pair of preparations in the thought experiment are not, in fact, operationally equivalent. This is because, by the lights of Newtonian mechanics, there is nothing forbidding a measurement that determines the exact microstate of the gas—the positions and momenta of each individual particle in the gas. One could then uniquely determine whether the microstate at the final time arose from time evolution of a microstate at the initial time wherein the gas started in the left compartment or from one wherein the gas started in the right compartment. Such a measurement is obviously an incredible technical challenge, but it is not ruled out by the lights of the physical theory being assumed. Indistinguishability relative to macroscopic thermodynamic properties or relative to measurements that are technologically feasible at the present day is *simply not relevant* to assessments of operational equivalence. Rather, all that matters is operational equivalence

relative to the measurements that are possible in principle by the lights of the physical theory under consideration.

In short, the objection considered here—that the notion of generalized noncontextuality is undermined by the fact that technological capabilities dictate whether in practice two laboratory procedures are distinguishable or not—simply misunderstands the notion of operational equivalence. For further discussion of this point, see Ref. [3] and Sec. II of Ref. [39].

Of course, if one wishes to assess directly from experimental data whether *nature* admits of a noncontextual model or not, then one cannot assume the correctness of any particular operational theory.

One way around this problem is to seek to experimentally determine the set of operational theories that are consistent with the experimental data, using theory-agnostic tomography, and then to assess the possibility of a noncontextual model relative to this set of operational theories. This is the approach taken in Ref. [17].

Theory-agnostic tomography requires that the set of preparations and the set measurements that are implemented on a system are *tomographically complete*. (Recall that a tomographically complete set of preparations is one such that its statistics are sufficient to infer the statistics of any other preparation, and a tomographically complete set of measurements is one such that its statistics are sufficient to infer the statistics of any other measurement.) But without prior knowledge of the operational theory governing a system, there is no way to know *a priori* whether the set of procedures that has been implemented is, in fact, tomographically complete. This is not, however, a deficiency in the definition of an operational theory. Rather, it is simply indicative of the fact that assessments of tomographic completeness, and hence of the operational theory that governs some system, are fallible. This in turn implies that assessments of noncontextuality are also fallible. This point is discussed further in Sec. V A.

There is a second way to try and directly test noncontextuality without presuming the correctness of any particular operational theory. Given that operational equivalence is indistinguishability relative to all measurements and given that a tomographically complete set of measurements is, by definition, one such that indistinguishability relative to it implies indistinguishability relative to all measurements, it follows that to assess whether two preparations are operationally equivalent, it is sufficient to assess whether they give the same statistics for all measurements in a *tomographically complete set*. Similarly, one can assess the operational equivalence of two measurements relative to a tomographically complete set of preparations. Thus, one can simply seek to assess operational equivalences in an experiment relative to a tomographically complete set of procedures (without seeking to meet the higher bar of determining all of the details about the GPT governing the system). Of course, assessments of tomographic completeness are fallible, but this merely implies that assessments of noncontextuality are also fallible, as we already noted above.

E. The imperfect-equivalences objection

Another concern which is sometimes raised about tests of generalized noncontextuality is that it is unclear how to ensure

that any given operational equivalence holds *exactly* between the procedures in any real experiment.

Imagine that one is interested in a particular operational identity between some target states $\{\mathbf{s}_x\}_{x \in \{1,2,3,4\}}$, say $\frac{1}{2}\mathbf{s}_1 + \frac{1}{2}\mathbf{s}_2 = \frac{1}{2}\mathbf{s}_3 + \frac{1}{2}\mathbf{s}_4$, and imagine that one has derived a noncontextuality inequality from this operational identity. In a real experiment, one can never succeed at preparing any of these target states exactly, but rather one generally ends up preparing some alternative nearby state, which we denote $\{\bar{\mathbf{s}}_x\}_{x=1,2,3,4}$. The latter states will generally not satisfy the operational identity $\frac{1}{2}\bar{\mathbf{s}}_1 + \frac{1}{2}\bar{\mathbf{s}}_2 = \frac{1}{2}\bar{\mathbf{s}}_3 + \frac{1}{2}\bar{\mathbf{s}}_4$ that one was targeting. Consequently, the noncontextuality inequality one wished to test is strictly not relevant to the experiment one actually performed, and it seems one is blocked from ever getting noise-robust tests of noncontextuality.

However, there are (at least) three different ways of circumventing this problem.

The first way to avoid this problem was introduced in Refs. [40,41]. Basically, the proposed resolution is to recognize that if one has experimentally determined the operational statistics generated by some set of GPT states (or GPT effects), then one can *logically infer* the statistics that would be generated by any convex mixture of these. So, one simply identifies a set of so-called *secondary states* and *secondary effects* that lie within the convex hull of those that were actually realized in the experiment, and that moreover satisfy exactly the desired operational identities. Although these states (effects) do not characterize any of the procedures that were actually implemented, they are known to be part of the operational theory governing the experiment, as they correspond to mixtures of procedures that were in fact realized and every operational theory is closed under mixing. One then tests the noncontextuality inequalities on the statistics described by these secondary states and effects (which can easily be computed from the states and effects). If one finds that the inequalities are violated, then one can be certain that there is no noncontextual model of the experimental data. The downside of this approach (noted in Ref. [41]) is that although one can always find secondary states and effects that satisfy the desired operational equivalences, these are always noisier than the realized ones. Since every noise-robust noncontextuality inequality has a threshold of noise beyond which it cannot be violated, it can happen that the transition from the primary to the secondary states and effects adds sufficient noise that one crosses the threshold and is unable to violate any noncontextuality inequality.

The second approach is more direct, and does not require introducing any secondary states or effects. Rather than deciding beforehand which noncontextual inequality is to be tested and consequently which operational identities are to be targeted in the experiment, instead one simply characterizes the GPT states and GPT effects that are actually realized in the experiment, then one determines the operational identities that happen to hold among these, and one derives noncontextuality inequalities based on *these* operational identities.

In our example above, for instance, the four realized states were denoted $\bar{\mathbf{s}}_1$, $\bar{\mathbf{s}}_2$, $\bar{\mathbf{s}}_3$, and $\bar{\mathbf{s}}_4$, and it was noted that they in general will not satisfy the simple operational identity $\frac{1}{2}\bar{\mathbf{s}}_1 + \frac{1}{2}\bar{\mathbf{s}}_2 = \frac{1}{2}\bar{\mathbf{s}}_3 + \frac{1}{2}\bar{\mathbf{s}}_4$. Nonetheless, if these states are confined to

a two-dimensional state space,⁸ then there will always exist real values $\{\alpha_x\}_{x \in \{1,2,3,4\}}$ for which $\alpha_1 \bar{s}_1 + \alpha_2 \bar{s}_2 = \alpha_3 \bar{s}_3 + \alpha_4 \bar{s}_4$, and these values can be inferred from the experimental characterization of the states. It then suffices to determine what noncontextuality inequalities follow from this operational identity. As it turns out, computing the noncontextuality inequalities that follow from an arbitrary set of operational identities can be achieved using a linear program [42].

The final approach circumvents the direct consideration of operational identities and noncontextuality inequalities altogether. One simply follows the procedure of theory-agnostic tomography (outlined in Sec. IID and discussed in detail in Refs. [17,18]) to experimentally determine the set of GPTs that are consistent with the experimental data. One then tests whether all of these GPTs are simplex-embeddable. Testing for simplex-embeddability is also achievable using a linear program [5].

F. The Kochen-Specker-were-naïve objection

Another objection that we have heard (in particular from philosophers of physics) is that assumptions of noncontextuality are naïve and unmotivated, and are studied today only because Kochen and Specker oversold their eponymous theorem. Recall that both Kochen and Specker [2] and Bell [43] independently arrived at no-go theorems from an assumption of noncontextuality (which we here term *KS-noncontextuality*, in order to distinguish it from the assumption of generalized noncontextuality). However—the argument goes—Kochen and Specker did not emphasize the role of this assumption when summarizing their no-go result, stating simply that [2] “The main aim of this paper is to give a proof of the nonexistence of hidden variables.” Bell, by contrast, was more circumspect [43]:

That so much follows from such apparently innocent assumptions leads us to question their innocence. Are the requirements imposed, which are satisfied by quantum mechanical states, reasonable requirements on the dispersion free states? Indeed they are not [...]. It was tacitly assumed that measurement of an observable must yield the same value independently of what other measurements may be made simultaneously [...]. These different possibilities require different experimental arrangements; there is no *a priori* reason to believe that the results [...] should be the same. The result of an observation may reasonably depend not only on the state of the system (including hidden variables) but also on the complete disposition of the apparatus.

Indeed, it is well known that one can construct explicit hidden variable models that do not satisfy the assumption of KS-noncontextuality and that reproduce all of quantum theory—Bohmian mechanics is one example. The critics of noncontextuality, particularly those who find Bohmian mechanics to be a satisfactory interpretation, take this fact as evidence that Kochen and Specker’s endorsement of the assumption as a natural one was naïve and that we should reject

the assumption of KS-noncontextuality as unreasonable, as Bell suggests in the above quote.

First, let us note that this is a rather uncharitable reading of Kochen and Specker’s work, as they certainly recognize the possibility of hidden variable theories that violate their assumption. Just prior to the quote that is cited by their critics, for instance, they state:

There are on the one hand purported proofs of the nonexistence of hidden variables, most notably von Neumann’s proof, and on the other, various attempts to introduce hidden variables such as de Broglie [44] and Bohm [45] and [46]. One of the difficulties in evaluating these contradictory results is that no exact mathematical criterion is given to enable one to judge the degree of success of these proposals.

Nonetheless, it seems to us fair to say that Kochen and Specker did not articulate any clear *a priori* motivation for their assumption of noncontextuality. Bell, by contrast, stated outright that he did not see any good argument in favor of such a principle of noncontextuality.

While it may be true that no good argument in favor of endorsing KS-noncontextuality had been given at the time of Bell and Kochen-Specker’s writings, such an argument *was* provided in subsequent work: one can motivate noncontextuality using a methodological principle for theory construction due to Leibniz (a version of his principle of the identity of indiscernibles) that has a long history of success in physics [3]. This principle motivates both KS-noncontextuality and also the notion of generalized noncontextuality introduced in Ref. [1].

Moreover, we consider the proof that one can characterize noncontextuality as simplex-embeddability within the framework of GPTs [13] to constitute another motivation for taking it as a good notion of classical explainability. For any simplex-embeddable GPT system, all the statistics that can be observed are compatible with the hypothesis that a *strictly classical* GPT gives the true description of one’s system. This is because one can never establish by empirical means that an apparent restriction on states and effects—i.e., a restriction to a state space that is a strict subset of the full simplex and/or to an effect space that is a strict subset of the full hypercube of effects—is fundamental as opposed to merely being due to a technological limitation that might be overcome in the future.⁹ In short, any experimental data that can be realized by a simplex-embeddable GPT can also be realized by a strictly classical GPT. But strictly classical GPTs have been motivated [11] (independently of any Leibnizian arguments) to be *the* GPT description of a system which is classical in the usual sense of being describable by a set of random variables (the different valuations of which define the possible ontic states

⁸If the states are not confined to a two-dimensional state space, then one simply requires more than four states to have a nontrivial linear dependence relation and hence a nontrivial operational identity.

⁹A concrete example helps to illustrate the point. If one performs theory-agnostic tomography on a system, and the state and effect spaces one realizes in the experiment are found to approximate those of the stabilizer states and measurements for a qubit (which are also the states and effects of the simplest system in the toy theory of Ref. [47]), then the range of GPTs that are consistent with this experimental data includes the strictly classical (i.e., simplicial) GPT of dimension four.

of the system) which can be measured perfectly. So simplex-embeddability is a natural notion of classical explainability.

Another set of motivations (which is less precise but arguably as compelling as those above) arises from inspection of the epistemically restricted classical theories in Refs. [47–50]. These theories are noncontextual and provide a compelling explanation of the operational phenomena that they reproduce. But more than this, it is the noncontextuality of the theories that *makes* the explanations compelling, and it is for this reason that we take such theories to provide further evidence of the naturalness of the principle of generalized noncontextuality. For example, a distinctive feature of quantum theory is that a given mixed state can be convexly decomposed into an ensemble of pure states in many different ways. Ontological models that are preparation-noncontextual explain the multiplicity of these different convex decompositions by modeling pure quantum states as nonpoint distributions over the ontic state space, and using the fact that many different mixtures of nonpoint distributions may yield the same distribution. (See, e.g., Sec. III.A.4 of Ref. [47].) Thus, noncontextuality provides a natural explanation of the multiplicity of convex decompositions of a mixed state in those subtheories of quantum theory that admit of a noncontextual ontological model.

For other motivations for taking generalized noncontextuality as a notion of classicality, see Refs. [3,15], or the introductions of Refs. [4,5].

In any case, if sceptics wish to criticize the *a priori* naturalness of assumptions of noncontextuality, it is obviously insufficient for them to base their criticisms only on one or two writings that are half a century old. They must also engage with all of the more recent motivations just discussed.

V. FURTHER DISCUSSION

In the following sections, we expand on some of the above points, or discuss related ideas that we think deserve wider recognition.

A. Fallibility of assessments of contextuality

If one assumes the correctness of some particular operational theory (e.g., quantum theory), then the question of whether the theory (or a fragment of it) admits of a noncontextual ontological model can be settled by a theoretical investigation. In particular, one can derive the relevant operational identities from the theory, and derive a no-go theorem based on these. No experiment needs to be performed in this case.

Consider now the question of whether a given set of *experimental procedures* that are realized in the lab admit of a noncontextual model. Here, one may or may not wish to assume the correctness of some particular operational theory, but one does *not* presume to know how each laboratory procedure is represented in the theory. If one does assume the correctness of, say, quantum theory, then the question one is answering is whether one's experiment lives inside a fragment of quantum theory that is classically explainable, or whether one has accessed a broad enough fragment of quantum theory to be provably nonclassical. (This can be useful

for the purposes of benchmarking experimental procedures that one hopes to use in a quantum-information-processing task.) The highest bar, however, is to test whether *nature itself* is noncontextual. To do this, one cannot assume *a priori* the correctness of any particular operational theory.

In either of these last two cases, one must deduce the characterizations of one's laboratory procedures from experimental data. Typically, one focusses on a particular type of system, and one considers a prepare-measure experiment on it. Assessments of the possibility of a noncontextual model for this experiment are based on the operational identities that are found therein, or equivalently, on the shape of the fragment of the space of GPT states and GPT effects that are realized in the experiment. Of course, if one is mistaken about the latter, then this will lead to mistaken conclusions about noncontextual-realizability and thus classical explainability. As with any inference from finite-run data to a scientific hypothesis, the inference one makes from the data of a contextuality experiment to the characterization of the GPT states and effects (and hence the operational identities among these) might be mistaken. However, one *can* build up evidence for or against a given hypothesis about operational identities. This evidence can be empirical, for instance, based on the best-fit states and effects arising from theory-agnostic tomography. But it can also come from physical principles, such as locality or the absence of retrocausation. It might also come from appealing to a particular physical theory (and the full body of evidence one has in support of that theory) and our knowledge of how that theory is applied to describe the particular laboratory procedures in question.

When the evidence in favor of operational identities among GPT states or among GPT effects comes from empirical data, the main way in which such assessments might ultimately prove to be incorrect is if the experimenter is mistaken about what constitutes a tomographically complete set of procedures. We refer the reader to Refs. [17,18] for a discussion of this issue. It was also discussed at the end of Sec. IV B.

It also important, however, to study precisely when mistaken assumptions about operational identities (i.e., the shapes of fragments of the GPT state and effect spaces) lead to mistaken assessments of noncontextuality.

A first important result in this vein was given in Ref. [51], which gave some sufficient conditions under which one can prove the failure of noncontextuality *even in cases where one is mistaken about or unsure of the operational identities*. A second important result in this vein follows from Lemma 11 of Ref. [52], concerning the question of whether contextuality proofs that assume the correctness of quantum theory are still valid if the world is in fact described by a postquantum GPT. In particular, the authors prove that such proofs continue to hold in the postquantum theory under some very reasonable assumptions about how quantum theory emerges from the postquantum theory via a decoherence-like process. An elaboration on this result and related matters is provided in forthcoming work [28]. For example, Ref. [28] shows that robust proofs of contextuality are possible using only a subsystem or a subspace of a larger physical system. So, for example, the mere existence of unprobed degrees of freedom—internal or otherwise—do not in and of themselves undermine contextuality proofs.

B. Noncontextuality is evaluated relative to a causal structure

In general, assessments of noncontextuality can only be made relative to a causal structure. This has been obscured by the fact that most research to date has focused on the simplest case of prepare-measure scenarios, where the causal structure was too simple to have merited any discussion.

Even in this simplest case, however, one must make assumptions of a causal nature—for example, that the experiment can be conceptualized as a preparation of a system followed by a measurement of that system, and that this system acts as the complete causal intermediary between the two stages. This is the system relative to which one evaluates operational identities: e.g., two preparations of a system are deemed operationally equivalent if they give the same statistics for all measurements *on that system*. In other words, the system delimits the scope of the universal quantifier in the definition of operational equivalence. The notion of system here is deemed to be a primitive notion, just as it is in the framework of GPTs, and indeed in most areas of physics. That is, we imagine that one has some individuating schema that allows for an identification of systems (equivalently, degrees of freedom) and an identification of the set of experimental procedures that pertain to these systems. In many physical contexts, especially those where experimentalists have a good deal of control, it is simple to identify systems and experimental procedures which act on them—there is little ambiguity in what is meant by the polarization degree of freedom of a photon, or by transformations on it. However, *formalizing* the schema by which physicists identify systems on the basis of operational statistics is a more subtle matter [17,18,53,54].

When one goes beyond prepare-measure scenarios, stronger causal assumptions are generally required. As we saw in Sec. III, one may also make assumptions about the subsystem structure of composite systems in order to obtain the strongest possible constraints from noncontextuality. Given that subsystem structure can be understood as a type of causal structure (see Appendix B of Ref. [15]), this is another type of causal assumption.

Let us now set up a more general example, where one imagines an experiment, perhaps as a subroutine of a quantum computation. We take Fig. 4(a) as our working example. The belief that the experiment is governed by this circuit is a *causal hypothesis*, and it contains a great deal more information than the operational statistics $p(DEF|ABC)$ on their own. The circuit diagram represents a commitment to the existence of a number of systems (S_1, S_2, S_3, S_4), represented as bold wires in the circuit, together with transformations (T_1, T_2, T_3, T_4), represented as gates in the circuit, and where the transformations may be chosen via classical setting variables (A, B, C) and may output classical outcome variables (D, E, F). The assumption that an experiment can be decomposed in this manner furthermore relies on the assumption that these transformations are autonomous in the sense that any one can be varied independently of the others. Note that an assumption of autonomy of causal mechanisms is also central to the framework for causal modeling used in the classical sphere [55,56] and allows for inferences about counterfactual questions, such as how the observed outcome statistics would have been different if one had modified the circuit in a

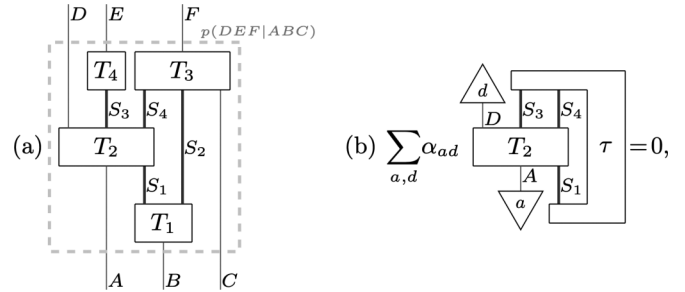


FIG. 4. To determine if the statistics $p(DEF|ABC)$ generated by some GPT circuit are classically explainable or not, one must look at the operational identities holding among different processes that could appear in any given gate within the circuit. Thus, these operational identities can only be defined relative to the circuit structure. For instance, an operational identity between the different possibilities for the gate taking S_1 to S_3, S_4 (indexed by setting A and outcome D) can be written as in panel (b).

particular manner. From the raw operational statistics $p(DEF|ABC)$ on their own, there is generally no way to verify that a given causal hypothesis is correct. However, constructing causal hypotheses is one of the central tasks of science, and one which is increasingly studied in a formal manner, both in classical and quantum contexts [55–59].

Consider now how noncontextuality arguments proceed under the assumption that Fig. 4(a) describes the causal structure. We do so from both of the two different perspectives on noncontextuality, described in Secs. II A and II B, respectively.

Consider first the schema of Sec. II A, wherein one begins by identifying operational identities holding among the processes generating the observed correlations. This requires one to consider each possible circuit element individually. Obviously, this requires knowing the input and output systems of each gate in the circuit, which is information contained in the causal hypothesis. One can see this graphically using the notion of a *tester*, e.g., the comb τ in Fig. 4(b); we refer the reader to Refs. [6,12] for details.

The second perspective is essentially an extension of the schema of Sec. II B from prepare-measure scenarios to arbitrary causal structures. It gives a holistic characterization of when an arbitrary GPT circuit admits of a classical explanation. As was shown in Ref. [6], a GPT circuit admits of a classical explanation (in the sense that the operational theory which it describes admits of a noncontextual model) if and only if one can find a linear, diagram-preserving map taking it into the process theory of substochastic matrices, while preserving the predicted correlations. This is equivalent to asking if a positive quasiprobability representation exists for the GPT in question [6,60].

This is illustrated schematically in Fig. 5. Clearly, one can only evaluate nonclassicality in this manner if one already has a circuit—a causal hypothesis—in mind.

Given a particular causal hypothesis, one always has the option to lump together processes to obtain a coarse-grained description; for example, in a prepare-transform-measure scenario on a single system, one may lump together the transformation and the measurement to reduce the scenario

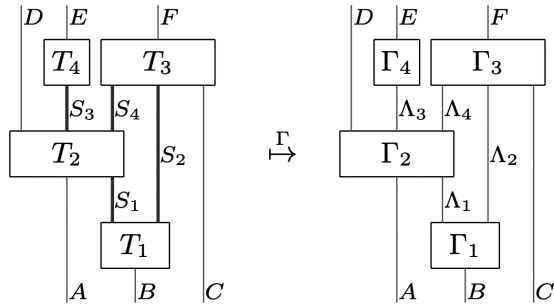


FIG. 5. As shown in Ref. [6], the statistics $p(DEF|ABC)$ generated by some GPT circuit are explainable within a noncontextual ontological model Γ if and only if there exists a linear map from the GPT circuit into a circuit of the same form but where all systems are classical random variables, and where all transformations are substochastic maps.

to an effective prepare-measure scenario. While this lumping of circuit elements can sometimes simplify one's analysis, it can also prevent one from deducing the *complete* consequences of noncontextuality relative to the causal hypothesis.

Take for example the stabilizer theory of a single qubit. In this operational theory, one can find proofs of contextual prepare-transform-measure scenarios, but one *cannot* find any such proofs in prepare-measure scenarios [61]. And yet, the composition of any transformation with any state (or measurement) in the theory yields another state (or measurement) in the theory. Thus, it follows that when one lumps together the transformation with either the state or the effect, the effective scenario that results is one which admits of a noncontextual model. This is simply a consequence of the choice to only carry out a coarse-grained analysis. In general, one must consider the ontological representation of each circuit element individually in order to determine necessary and sufficient conditions for noncontextuality, relative to the (fine-grained) causal hypothesis.

The considerations of this section raise the question of how one decides between different causal hypotheses, a question to which we turn in the next section.

C. When one should assume diagram preservation with respect to the standard quantum circuit

As mentioned earlier, constructing a causal hypothesis for some observed phenomena is a difficult but central scientific task. We now argue that—at least in various branches of experimental quantum information processing, where one has a good deal of control over the systems in question—it is typically straightforward to write down *the* quantum circuit associated with a given experiment or protocol. We argue that the structure of this circuit provides the natural causal hypothesis, and if one seeks realist explanations of the observed data, one should demand that the explanation respects this causal hypothesis.

On what grounds do physicists ever associate a system with some operational procedures carried out in a laboratory? We poke and we prod at the world until we identify meaningful loci of intervention. With enough experimentation, we

eventually distill out meaningful notions of systems (like electrons, photons, etc.), and we imagine that these systems have properties which are prepared, measured, and transformed by our interventions. These systems, then, are the most natural candidates for causal mediaries, and the standard quantum circuit describing the experiment is the natural candidate for the causal structure. In other words, this causal hypothesis is the most natural culmination of all the evidence gathered to date for how one can break up an experiment into localized systems and autonomous transformations on them. To postulate any other causal structure is a more radical move and requires special justification.

Consider for example the standard Bell scenario. The standard quantum circuit for this scenario invokes a common cause which sets up correlations between the local measurements performed by the two parties. The conservative causal hypothesis is that the causal structure has the same form at the ontological level, and it is precisely this assumption that leads to Bell inequalities.

The generalization of this line of reasoning to arbitrary causal scenarios is given by the rapidly growing field of causal compatibility inequalities [22,62–64], where one demonstrates nonclassicality by showing that quantum circuits of a given causal structure are capable of generating a broader set of correlations than classical ones. Such arguments *also* rely on the assumption that an experiment associated with a given quantum circuit is in fact a faithful realization of the causal structure described by the quantum circuit.

Of course, some physicists (such as proponents of Bohmian mechanics) do believe in superluminal causal influences, and so would advocate for a causal hypothesis which does *not* mirror the structure of the standard quantum circuit. So this is not to say that there is *no* sense in considering non-standard causal hypotheses; however, interpretations which make radical causal assumptions are made less compelling as a consequence. Moreover, these nonstandard causal hypotheses are typically endorsed by those who do not believe that there is any other way out of no-go theorems like Bell's, but, as we argue in the next section, we believe that there *are*, in fact, other ways out.

Another reason to demand that the ontological representation of an experiment respect the conservative causal structure is that representations using radical causal structures typically *overfit* the data. For example, explanations of Bell inequality violations which appeal to superluminal causation are often general enough to allow for signaling correlations, and overfitting is generally a consequence of this [23]. Alternatively, one can avoid this overfitting by imposing restrictions on the scope of causal mechanisms allowed in the model, but these restrictions generally lead to violations of noncontextuality [65].

A motivation for studying ontological representations is to the search for deeper explanations for our experiments and theories. The most natural explanations are those with the most conservative assumptions about the causal structure, which we have argued correspond to the standard quantum circuit representation. It follows that one should focus one's attention on ontological models that respect the structure of the quantum circuit. This desiderata is captured by demanding diagram-preservation relative to the standard quantum circuit.

D. What to do in light of the failure of noncontextuality in quantum theory

For experiments described by operational quantum theory, one cannot necessarily find ontological representations that respect the conservative causal structure and are noncontextual. There are by now many proofs of this fact, spanning a variety of physical scenarios [14,30,37,38,40,61,66–81]. What should one conclude when faced with these no-go theorems?

There are three natural possibilities. The first is to imagine that quantum theory is not the true theory of nature and that the operational equivalences that have been observed to date are not operational equivalences in the true theory, which might then still be consistent with noncontextuality. The second is to simply bite the bullet and grant that nature is described by a contextual ontological model. This is the route, for instance, that advocates of Bohmian mechanics endorse. The third is to relax some of the background assumptions going into proofs of noncontextuality, such as the assumptions built into the framework of ontological models, in a manner that allows one to maintain the spirit of noncontextuality.

The first response above is, in our view, unlikely to be the correct resolution to the problem. For one, it can be shown that the operational equivalences arising in quantum theory will *continue to hold* in a postquantum theory, if one grants a few reasonable physical assumptions regarding the sense in which quantum theory emerges from this postquantum theory via a decoherence-like process [28,52]. Also, in the case where the operational equivalences follow from the lack of signaling between spacelike separated regions, such as in a Bell experiment, this sort of response requires one to imagine that the true postquantum theory is one that allows for superluminal signaling and hence conflicts with relativity theory, a possibility that we take to be unlikely.

We find the second response to be unsatisfactory because contextual theories lack much of the explanatory power that are provided by noncontextual theories. For instance, if one considers the operational phenomenology of the odd-dimensional stabilizer subtheory of quantum theory, then the Bohmian account of this phenomenology is far more convoluted and counterintuitive than the description provided by the Spekkens toy theory [47], or equivalently, Gross's discrete Wigner representation [82].

This leaves the third possible response, that one must consider modifying the framework of ontological models in such a way that one can construct a realist description of quantum theory that salvages the spirit of noncontextuality. This is easier said than done, since the standard framework of ontological models is an extremely general and compelling framework for providing realist explanations, and it is unclear how to modify it while retaining these features. Nonetheless, we believe that this is the correct response, and first steps in this direction can be found in Ref. [15] (see also Ref. [19]). In particular, for the special class of contextuality experiments that are Bell experiments, while a standard response to Bell-inequality violations is to concede that nature allows superluminal causes (relativity be damned), the third type of response asks one to instead question the background assumptions going into Bell-like no-go theorems. If one modifies the framework of causal modeling that underlies these no-go theorems, then one can hope to find an intrinsically quantum notion of causation [57–59] that reproduces the quantum predictions in Bell scenarios while preserving the spirit of locality. More specifically, the aim of such works is to explain Bell violations as consequences of nonclassical common causes [31–33,58,83] rather than superluminal causes.

Thus, our preferred response to both noncontextuality no-go theorems and Bell-like no-go theorems is to devise a more general notion of *nonclassical realism* that allows us to give causal explanations of observed correlations in a manner that is consistent with the Leibnizian methodological principle [15,19].

ACKNOWLEDGMENTS

We thank Elie Wolfe, Ana Belén Sainz, and Ravi Kunjwal for useful discussions, especially regarding operational equivalences among composite systems and regarding the PR box example. We also thank Matt Pusey for useful discussions, and thank Lídia del Rio for feedback on Sec. III, and indeed for motivating us to write it (whether or not that was intentional!). J.H.S. was supported by the National Science Centre, Poland (Opus project, Categorical Foundations of the Non-Classicality of Nature, Project no. 2021/41/B/ST2/03149). D.S. was supported by the Foundation for Polish Science (IRAP project, ICTQT, contract no. MAB/2018/5, co-financed by EU within Smart Growth Operational Programme). All diagrams were prepared using TikZit.

-
- [1] R. W. Spekkens, *Phys. Rev. A* **71**, 052108 (2005).
 [2] S. Kochen and E. Specker, *J. Math. Mech.* **17**, 59 (1967).
 [3] R. W. Spekkens, [arXiv:1909.04628](https://arxiv.org/abs/1909.04628).
 [4] D. Schmid, Ph.D. thesis, University of Waterloo, 2021, <http://hdl.handle.net/10012/17136>.
 [5] J. H. Selby, E. Wolfe, D. Schmid, A. B. Sainz, and V. P. Rossi, *Phys. Rev. Lett.* **132**, 050202 (2024).
 [6] D. Schmid, J. H. Selby, M. F. Pusey, and R. W. Spekkens, [arXiv:2005.07161](https://arxiv.org/abs/2005.07161).
 [7] D. Schmid, *Noncontextuality, part 1 (Youtube)* (2022).

- [8] D. Schmid, *Noncontextuality, part 2 (Youtube)* (2022).
 [9] D. Schmid, *Noncontextuality, part 3 (Youtube)* (2022).
 [10] L. Hardy, [arXiv:quant-ph/0101012](https://arxiv.org/abs/quant-ph/0101012).
 [11] J. Barrett, *Phys. Rev. A* **75**, 032304 (2007).
 [12] G. Chiribella, G. M. D'Ariano, and P. Perinotti, *Phys. Rev. A* **81**, 062348 (2010).
 [13] D. Schmid, J. H. Selby, E. Wolfe, R. Kunjwal, and R. W. Spekkens, *PRX Quantum* **2**, 010331 (2021).
 [14] J. H. Selby, D. Schmid, E. Wolfe, A. B. Sainz, R. Kunjwal, and R. W. Spekkens, *Phys. Rev. A* **107**, 062203 (2023).

- [15] D. Schmid, J. H. Selby, and R. W. Spekkens, [arXiv:2009.03297](#).
- [16] R. W. Spekkens, [Found. Phys.](#) **44**, 1125 (2014).
- [17] M. D. Mazurek, M. F. Pusey, K. J. Resch, and R. W. Spekkens, [PRX Quantum](#) **2**, 020302 (2021).
- [18] M. Grabowecky, C. Pollack, A. Cameron, R. Spekkens, and K. Resch, [Phys. Rev. A](#) **105**, 032204 (2022).
- [19] D. Schmid, [PIRSA:19120030](#) 10.48660/19120030 (2019).
- [20] S. Popescu and D. Rohrlich, [Found. Phys.](#) **24**, 379 (1994).
- [21] P. J. Cavalcanti, J. H. Selby, J. Sikora, T. D. Galley, and A. B. Sainz, [npj Quantum Inf.](#) **8**, 76 (2022).
- [22] C. J. Wood and R. W. Spekkens, [New J. Phys.](#) **17**, 033002 (2015).
- [23] P. J. Daley, K. J. Resch, and R. W. Spekkens, [Phys. Rev. A](#) **105**, 042220 (2022).
- [24] A. Fine, [Phys. Rev. Lett.](#) **48**, 291 (1982).
- [25] J. Barrett, D. Collins, L. Hardy, A. Kent, and S. Popescu, [Phys. Rev. A](#) **66**, 042111 (2002).
- [26] P. Bierhorst, [J. Phys. A: Math. Theor.](#) **48**, 195302 (2015).
- [27] L. K. Shalm, E. Meyer-Scott, B. G. Christensen, P. Bierhorst, M. A. Wayne, M. J. Stevens, T. Gerrits, S. Glancy, D. R. Hamel, M. S. Allman *et al.*, [Phys. Rev. Lett.](#) **115**, 250402 (2015).
- [28] J. H. Selby, D. Schmid, V. Rossi, and A. B. Sainz, When failures of tomographic completeness are not problematic for noncontextuality (unpublished).
- [29] D. Gottesman, *Group22: Proceedings of the XXII International Colloquium on Group Theoretical Methods in Physics*, edited by S. P. Corney, R. Delbourgo, and P. D. Jarvis (International Press, Cambridge, 1999), pp. 32–43.
- [30] D. Schmid, H. Du, J. H. Selby, and M. F. Pusey, [Phys. Rev. Lett.](#) **129**, 120403 (2022).
- [31] E. Wolfe, D. Schmid, A. B. Sainz, R. Kunjwal, and R. W. Spekkens, [Quantum](#) **4**, 280 (2020).
- [32] D. Schmid, D. Rosset, and F. Buscemi, [Quantum](#) **4**, 262 (2020).
- [33] D. Schmid, T. C. Fraser, R. Kunjwal, A. B. Sainz, E. Wolfe, and R. W. Spekkens, [Quantum](#) **7**, 1194 (2023).
- [34] J. Anders and D. E. Browne, [Phys. Rev. Lett.](#) **102**, 050502 (2009).
- [35] R. W. Spekkens, [PIRSA:19110120](#) (2019).
- [36] R. Raussendorf, [Phys. Rev. A](#) **88**, 022322 (2013).
- [37] A. Chailloux, I. Kerenidis, S. Kundu, and J. Sikora, [New J. Phys.](#) **18**, 045003 (2016).
- [38] A. Ambainis, M. Banik, A. Chaturvedi, D. Kravchenko, and A. Rai, [Quantum Inf. Process.](#) **18**, 111 (2019).
- [39] L. Catani, M. Leifer, D. Schmid, and R. W. Spekkens, [arXiv:2207.11791](#).
- [40] M. F. Pusey, [Phys. Rev. A](#) **98**, 022112 (2018).
- [41] M. D. Mazurek, M. F. Pusey, R. Kunjwal, K. J. Resch, and R. W. Spekkens, [Nat. Commun.](#) **7**, 11780 (2016).
- [42] D. Schmid, R. W. Spekkens, and E. Wolfe, [Phys. Rev. A](#) **97**, 062103 (2018).
- [43] J. S. Bell, [Rev. Mod. Phys.](#) **38**, 447 (1966).
- [44] L. de Broglie, *Non-Linear Wave Mechanics* (Elsevier, Amsterdam, The Netherlands, 1960).
- [45] D. Bohm, [Phys. Rev.](#) **85**, 166 (1952).
- [46] D. Bohm, [Phys. Rev.](#) **85**, 180 (1952).
- [47] R. W. Spekkens, [Phys. Rev. A](#) **75**, 032110 (2007).
- [48] S. D. Bartlett, T. Rudolph, and R. W. Spekkens, [Phys. Rev. A](#) **86**, 012103 (2012).
- [49] L. Catani, M. Leifer, D. Schmid, and R. W. Spekkens, [Quantum](#) **7**, 1119 (2023).
- [50] R. W. Spekkens, Quasi-quantization: Classical statistical theories with an epistemic restriction, in *Quantum Theory: Informational Foundations and Foils*, edited by G. Chiribella and R. W. Spekkens (Springer Netherlands, Dordrecht, 2016), pp. 83–135.
- [51] M. F. Pusey, L. del Rio, and B. Meyer, [arXiv:1904.08699](#).
- [52] M. P. Mueller and A. J. Garner, [Phys. Rev. X](#) **13**, 041001 (2023).
- [53] G. Chiribella, [Entropy](#) **20**, 358 (2018).
- [54] L. Krämer and L. Del Rio, [Philos. Trans. R. Soc., A](#) **376**, 20170321 (2018).
- [55] J. Pearl, *Causality* (Cambridge University Press, Cambridge, 2009).
- [56] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, Prediction, and Search* (MIT Press, Massachusetts, 2000).
- [57] F. Costa and S. Shrapnel, [New J. Phys.](#) **18**, 063032 (2016).
- [58] J.-M. A. Allen, J. Barrett, D. C. Horsman, C. M. Lee, and R. W. Spekkens, [Phys. Rev. X](#) **7**, 031021 (2017).
- [59] J. Barrett, R. Lorenz, and O. Oreshkov, [arXiv:1906.10726](#).
- [60] R. W. Spekkens, [Phys. Rev. Lett.](#) **101**, 020401 (2008).
- [61] P. Lillystone, J. J. Wallman, and J. Emerson, [Phys. Rev. Lett.](#) **122**, 140405 (2019).
- [62] T. Fritz, [New J. Phys.](#) **14**, 103001 (2012).
- [63] A. Tavakoli, A. Pozas-Kerstjens, M.-O. Renou *et al.*, [Rep. Prog. Phys.](#) **85**, 056001 (2021).
- [64] R. Chaves, G. Moreno, E. Polino, D. Poderini, I. Agresti, A. Suprano, M. R. Barros, G. Carvacho, E. Wolfe, A. Canabarro, R. W. Spekkens, and F. Sciarrino, [PRX Quantum](#) **2**, 040323 (2021).
- [65] D. Schmid, J. Selby, and R. W. Spekkens (unpublished).
- [66] M. S. Leifer and R. W. Spekkens, [Phys. Rev. Lett.](#) **95**, 200405 (2005).
- [67] M. F. Pusey and M. S. Leifer, in *Proceedings of the 12th International Workshop on Quantum Physics and Logic*, edited by C. Heunen, P. Selinger, and J. Vicary, Electronic Proceedings in Theoretical Computer Science (Open Publishing Association, Oxford, UK, 2015), pp. 295–306.
- [68] R. W. Spekkens, D. H. Buzacott, A. J. Keehn, B. Toner, and G. J. Pryde, [Phys. Rev. Lett.](#) **102**, 010401 (2009).
- [69] Y.-C. Liang, R. W. Spekkens, and H. M. Wiseman, [Phys. Rep.](#) **506**, 1 (2011).
- [70] M. F. Pusey, [Phys. Rev. Lett.](#) **113**, 200401 (2014).
- [71] R. Kunjwal and R. W. Spekkens, [Phys. Rev. Lett.](#) **115**, 110403 (2015).
- [72] D. Schmid and R. W. Spekkens, [Phys. Rev. X](#) **8**, 011015 (2018).
- [73] R. Kunjwal, M. Lostaglio, and M. F. Pusey, [Phys. Rev. A](#) **100**, 042116 (2019).
- [74] D. Saha and A. Chaturvedi, [Phys. Rev. A](#) **100**, 022108 (2019).
- [75] M. Lostaglio, [Phys. Rev. Lett.](#) **125**, 230603 (2020).
- [76] M. Lostaglio and G. Senno, [Quantum](#) **4**, 258 (2020).
- [77] S. A. Yadavalli and R. Kunjwal, [Quantum](#) **6**, 839 (2022).
- [78] J. H. Selby, D. Schmid, E. Wolfe, A. B. Sainz, R. Kunjwal, and R. W. Spekkens, [Phys. Rev. Lett.](#) **130**, 230201 (2023).
- [79] C. Roch Carceller, K. Flatt, H. Lee, J. Bae, and J. B. Brask, [Phys. Rev. Lett.](#) **129**, 050501 (2022).

- [80] K. Flatt, H. Lee, C. R. I. Carceller, J. B. Brask, and J. Bae, [PRX Quantum](#) **3**, 030337 (2022).
- [81] D. Schmid, [Quantum](#) **8**, 1217 (2024).
- [82] D. Gross, [J. Math. Phys.](#) **47**, 122107 (2006).
- [83] E. G. Cavalcanti and R. Lal, [J. Phys. A: Math. Theor.](#) **47**, 424018 (2014).