Causal reappraisal of the quantum three-box paradox

Pawel Blasiak [*] and Ewa Borsuk [†]
Institute of Nuclear Physics Polish Academy of Sciences, PL-31342 Kraków, Poland

(Received 29 July 2021; accepted 23 November 2021; published 10 January 2022)

The quantum three-box paradox is a prototypical example of some bizarre predictions for intermediate measurements made on *pre- and postselected* systems. Although in principle those effects can be explained by measurement disturbance, it is not clear what mechanisms are required to fully account for the observed correlations. In this paper, this paradox is scrutinized from the causal point of view. We consider an array of potential causal structures behind the experiment, eliminating those without enough explanatory power. A distinction is made between the propagation in the system of just the *measurement outcome* and the information about the *full measurement context*. We also discuss the consequences of the *realist* position in which preexisting values are revealed by measurements. Interestingly, the answers depend crucially on whether the original version of the paradox is considered or its extension where the third box is allowed for inspection too. This illustrates the richness of the paradox which is better appreciated from the causal perspective.

DOI: 10.1103/PhysRevA.105.012207

I. INTRODUCTION

A paradox builds upon a conflict between observed facts and preconceptions that we hold about them, with a view to elicit a revision of the latter for a deeper understanding of a given problem or phenomenon. On the one hand, this may be just a warning about a superficial understanding of the mathematics, which if correctly applied does not lead to any contradictions. It is particularly true about the problems involving probability. On the other hand, a paradox can be an indication of a deeper misconception regarding the mechanisms at work, and thus potentially revealing something new about the nature itself. Therein lies the interest for the foundational issues of quantum theory and, in particular, the question about the causal nature behind its predictions. Notably, some remarkable results on quantum nonlocality [1–3] or contextuality [4–6] are prime examples with a paradox and causality at the background.

A *three-box paradox* [7] is a flagship example of the *pre- and postselection* (PPS) scenarios, in which some surprising predictions about intermediate measurements are made. It was originally proposed as an illustration of the so-called ABL rule [8] (after Aharonov, Bergmann, and Lebowitz). For the three-box paradox case it makes a strange prediction regarding the position of a particle which is always found where it is looked for. This has sparked controversy regarding the nature of the paradox and conclusions that can be drawn from this bizarre effect [9–15]. The first objection concerns the presence of postselection in the experiment, since the rejection of data is a potential source of noncausal correlations known as a *selection bias* [16]. The second problem stems from the possible role of *measurement disturbance* in the experiment, since in this case the disturbance can propagate in the system making the information about the intermediate measurements available at the moment of postselection. Those issues certainly affect interpretation of the paradox and thus need careful reassessment within a proper conceptual framework.

In this paper we tread the path eloquently expressed in Pearl's [17] conjecture that "*human intuition is organized around casual, not statistical, relations*." It suggests that in an attempt at resolving a paradox one should rather focus on causal mechanisms behind the observed correlations. Not only does this give a way to the bottom of the paradox by explicating the implicit assumptions that we make, but it sometimes may even offer something new about the causal mechanisms at work. Notably, the causal approach has recently gained a solid mathematical foundation in the works of Pearl and others [16,18,19] (which goes along similar lines as introduced by Bell [1]). Some remarkable results in the field of causal inference pertinent to the present paper include *d-separation rules* and *instrumental inequalities*. This novel approach has helped in resolving various conundrums in observational studies in epidemiology, computer science, and social sciences [20–22]. Despite a fairly recent development, mostly outside of physics, those methods have already influenced the research in quantum foundations (see, e.g., Refs. [23–31]).

In this paper, we employ the tools of causal inference to analyze measurement disturbance and its impact on postselection in the three-box paradox. This approach allows differentiating between the various mechanisms in which measurement disturbance can propagate. We also bring to light some implicit assumptions about realism that are typically made, which explains where the clash with our intuition might come from.

II. THREE-BOX PARADOX IN A NUTSHELL

Consider an experiment with a system prepared at time t_0 in some state initial state ρ_0 on which at a later time t_2 a projective measurement $\mathcal{M}_2$ is made checking for state Π^{post}. Let us agree to retain only the positive results $M_2 = 1$, which are deemed a success, and discard all the rest $M_2 = 0$. This is the so-called PPS scenario. To make it more interesting we

[*] pawel.blasiak@ifj.edu.pl
[†] ewa.borsuk@ifj.edu.pl

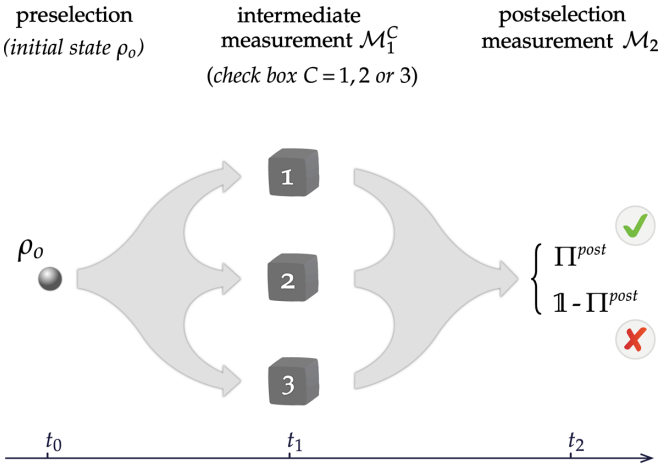

FIG. 1. Three-box experiment. Consider a particle that can be in one of three boxes labeled 1, 2, and 3. The system is preselected in state ρ_0 and postselected in state Π^{post}. In each experimental trial we choose a box $C = 1, 2, 3$ checking whether the particle is there or not. Quantum mechanics makes a puzzling prediction that whichever box $C = 1$ or 2 we choose to look at, the particle will be always found there [see Eq. (4)].

allow ourselves to make a measurement $\mathcal{M}_1^C$ in some intermediate time t_1 ($t_0 < t_1 < t_2$) described by a projection-valued measure (PVM) $\{\Pi_i^C\}$, where C stands for the *choice* of measurement in a given experimental run. Then, the conditional probability of obtaining outcome $M_1 = i$ in the PPS scenario is given by the ABL rule [8]

$$P(M_1 = i|M_2 = 1, C) = \frac{\text{Tr}[\Pi^{\text{post}}\,\Pi_i^C\,\rho_0\,\Pi_i^C]}{\sum_l \text{Tr}[\Pi^{\text{post}}\,\Pi_l^C\,\rho_0\,\Pi_l^C]}. \tag{1}$$

It is a straightforward application of Bayes' theorem to the joint probability

$$P(M_1 = i, M_2 = j|C) = \text{Tr}\big[\Pi_j^{\text{post}}\,\Pi_i^C\,\rho_0\,\Pi_i^C\big], \tag{2}$$

which is obtained by the usual von Neumann–Lüders rule. Here the final measurement $\mathcal{M}_2$ is described by the PVM $\{\Pi_j^{\text{post}}\}_{j=0,1} \equiv \{\mathbb{1} - \Pi^{\text{post}}, \Pi^{\text{post}}\}$.

A three-box paradox [7] is a specific realization of the PPS scenario in which the intermediate measurements $\mathcal{M}_1^C$ are assigned deterministic conditional outcomes. It concerns a single particle that can be localized in one of three boxes labeled 1, 2, and 3 described by the respective quantum states $|1\rangle$, $|2\rangle$, and $|3\rangle$. Suppose in the intermediate measurements we check whether the particle is in a given box $C = 1, 2, 3$, or not, which is implemented by the PVM $\{\Pi_i^C\}_{i=0,1} \equiv \{\mathbb{1} - |C\rangle\langle C|, |C\rangle\langle C|\}$. See Fig. 1. Now, if we choose for the pre- and postselected states, $\rho_0 = |\phi\rangle\langle\phi|$ and $\Pi^{\text{post}} = |\psi\rangle\langle\psi|$, the following nonorthogonal pair,

$$|\phi\rangle = \tfrac{|1\rangle+|2\rangle+|3\rangle}{\sqrt{3}} \quad \text{and} \quad |\psi\rangle = \tfrac{|1\rangle+|2\rangle-|3\rangle}{\sqrt{3}}, \tag{3}$$

then, from the ABL rule Eq. (1), we get

$$P(M_1 = 1|M_2 = 1, C) = 1 \qquad \text{for } C = 1, 2. \tag{4}$$

Hence, we have a paradoxical conclusion that whatever box we check, $C = 1$ or 2, the particle is always there.

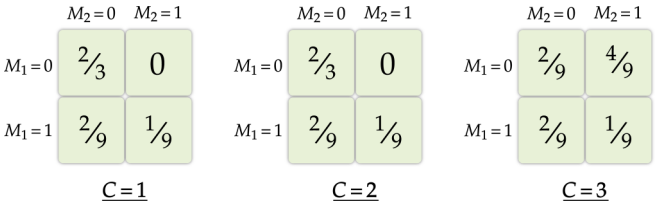

FIG. 2. Full statistics in three-box experiment. The joint probability $P(M_1, M_2|C)$ of obtaining measurement outcomes $M_1 = 0, 1$ (i.e., particle not found or found) and $M_2 = 0, 1$ (i.e., postselection is a failure or success) for the choice of experiment $C = 1, 2, 3$ (i.e., which box to check).

The full statistics observed in the experiment is given in Fig. 2, which readily follows from Eq. (2). This shows that for $C = 1, 2$ postselection succeeds with the probability equal to $P(M_2 = 1|C) = 1/9$.[1] Let us note in advance that although the original formulation of the paradox in Eq. (4) concerns just two (out of three possible) experimental choices $C = 1, 2$, it becomes more revealing and weird when we look at the full statistics $C = 1, 2, 3$.

III. CAUSAL PICTURE OF THE THREE-BOX EXPERIMENT

Let us consider the possible causal structures hiding behind the experiment which will be further assessed against their capacity for generating the three-box statistics in Fig. 2. We need to decide about the variables deemed relevant for the description of the experiment and then ponder their causal relationships.

A. Pure causal setting

In the description of the three-box experiment there are three *observed* variables:

C : choice of measurement setting ($C = 1, 2, 3$),

M_1 : outcome of measurement $\mathcal{M}_1^C$ ($M_1 = 0, 1$),

M_2 : outcome of measurement $\mathcal{M}_2$ ($M_2 = 0, 1$).

Furthermore, let us postulate the existence of some *unobserved* variable:

Λ : hidden (or latent) variable.

This variable is aimed to describe any other factors relevant for the experiment (e.g., the details of the preparation procedure). It is left unspecified in order not to restrict the range of possible explanations behind the observed correlations. We call it a *pure causal* setting as it involves the least number of assumptions (this should be compared with the more restricted framework below).

B. Realist causal setting

It is very instructive to consider in parallel the additional *realism* assumption. This is a position which derives from the worldview in which physical objects and their properties exist,

[1]We remark that this probability is the same as when no intermediate measurement $\mathcal{M}_1^C$ is made, in which case we also have $P(M_2 = 1|$ without $\mathcal{M}_1^C) = \text{Tr}[\Pi^{\text{post}}\,\rho_0] = |\langle\psi\,|\,\phi\rangle|^2 = 1/9$.

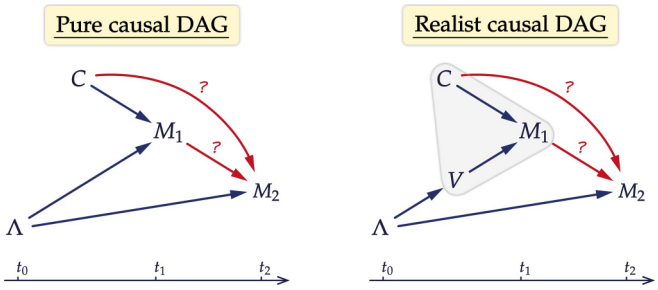

FIG. 3. Causal diagrams for the three-box experiment. In both diagrams, the variables C, M_1, and M_2 are observed in the experiment, whereas Λ and V are unobserved (latent or hidden) variables. These are directed acyclic graphs (DAGs) which are allowed by the temporal structure of the experiment (no retro-causation) and assuming C to be a free variable. The diagram on the right includes an additional structure (shaded in gray) which reflects the realism assumption. In both diagrams the red arrows (with question marks) are responsible for different types of measurement disturbance propagating in the system, i.e., whether it is just the outcome M_1 or the full measurement context C that affects M_2.

and the measurements reveal those preexisting values. In our case this view boils down to the existence of an additional variable

$$V : \text{position of the particle } (V = 1, 2, 3),$$

having a definite value before the measurement is made (possibly a derivative of the hidden variable Λ). Since the measurement $\mathcal{M}_1^c$ answers the question "Is the particle in a given box C?", we have the following consistency condition:

$$M_1(C, V) := \delta_{C,V}. \tag{5}$$

It is a direct expression of the requirement that the measurement reveals the property of the particle being in a given box. We remark that those additional structural components make the *realist causal* setting more restrictive compared to the *pure causal* setting, as we shall see shortly.

C. Causal diagrams for the experiment

Let us draw the diagrams compatible with the above two descriptions. In Fig. 3 the arrows represent cause-and-effect relationships between the variables. Observe that the temporal structure of the experiment allows us to eliminate certain arrows in the diagrams. Namely, we assume only *forward-in-time causation*. Also, since the choice of measurement is considered to be a *free variable*, we assume there is no arrow incoming to C. [Note that, although in principle in the *realist causal directed acyclic graph* (DAG) (right) there could be an arrow $V \to M_2$, it has not been drawn since we can always incorporate it in the arrow $\Lambda \to M_2$ (by appropriately modifying Λ).]

We note that the diagrams in Fig. 3 include *all* arrows compatible with the experiment. In this paper we pose the question about the *necessity* of arrows $M_1 \to M_2$ and $C \to M_2$. Both are responsible for the causal effects of the intermediate measurement $\mathcal{M}_1^c$ in the experiment. The lack of both arrows means no measurement disturbance. Conversely, their presence is a sign of different types of disturbance propagating in the system, i.e., whether it is just the measurement outcome M_1 or the full context specified by the choice of measurement (parameter) C that affects the final measurement M_2. Using the terminology borrowed from the analysis of Bell nonlocality [32], we call those arrows as

$$M_1 \to M_2 \ : \text{outcome dependence},$$
$$C \to M_2 \ : \text{parameter dependence}$$

Having specified causal structures of interest we may assess their potential for explaining the statistics of the three-box paradox in Fig. 2.

IV. MAIN RESULTS

A. Inspection of causal structures

In the causal inference field we are interested in verifying whether a given causal structure can explain the observed experimental behavior, i.e., whether the statistics can be reproduced by some *structural causal model* (SCM) consistent with a given *causal DAG* (see Refs. [16,18–29,31]). Since adding arrows to the diagram extends its expressive power, we are looking for structures with the fewest number of arrows which can still explain the observed experimental behavior.

As noted, the realist causal setting is more restrictive than the pure causal setting (see Fig. 3), i.e., not all behaviors compatible with the pure causal DAG (left) are admitted by the realist causal DAG (right). In the following we consider both cases separately. Furthermore, we also distinguish between the statistics in the three-box experiment in Fig. 2 for the *full* choice of three measurements $C = 1, 2, 3$, and the case *limited* to the two choices $C = 1, 2$ (as in the original exposition of the paradox). This will make an interesting case regarding our perception of the paradox and its further ramifications as explained below.

Our results are summarized in Fig. 4. For the proofs of necessity of the respective arrows see Appendix A (which employs the tools of causal inference [16]: the d-separation criterion and instrumental inequalities). The proofs of sufficiency are given in Appendix B (this requires explicit construction of SCMs in each case).

B. Conclusions

Having analyzed possible causal explanations of the three-box experiment it is natural to ask why people find it surprising. Notably, the original formulation of the paradox [7] concerns just two (out of three possible) experimental choices $C = 1, 2$; here the postselected behavior is deterministic, which makes it better suited for human judgment. In this case the paradox seems to arise from the tension between the assumption of realism and the deceptive impression, from how the paradox is phrased, as to the lack of measurement disturbance in the experiment. We showed that both assertions are intrinsically contradictory. The requirement of realism necessitates measurement disturbance of some sort [see Fig. 4(b)], i.e.,

$$\text{Realism} \underset{C=1,2}{\Rightarrow} \text{measurement disturbance}.$$

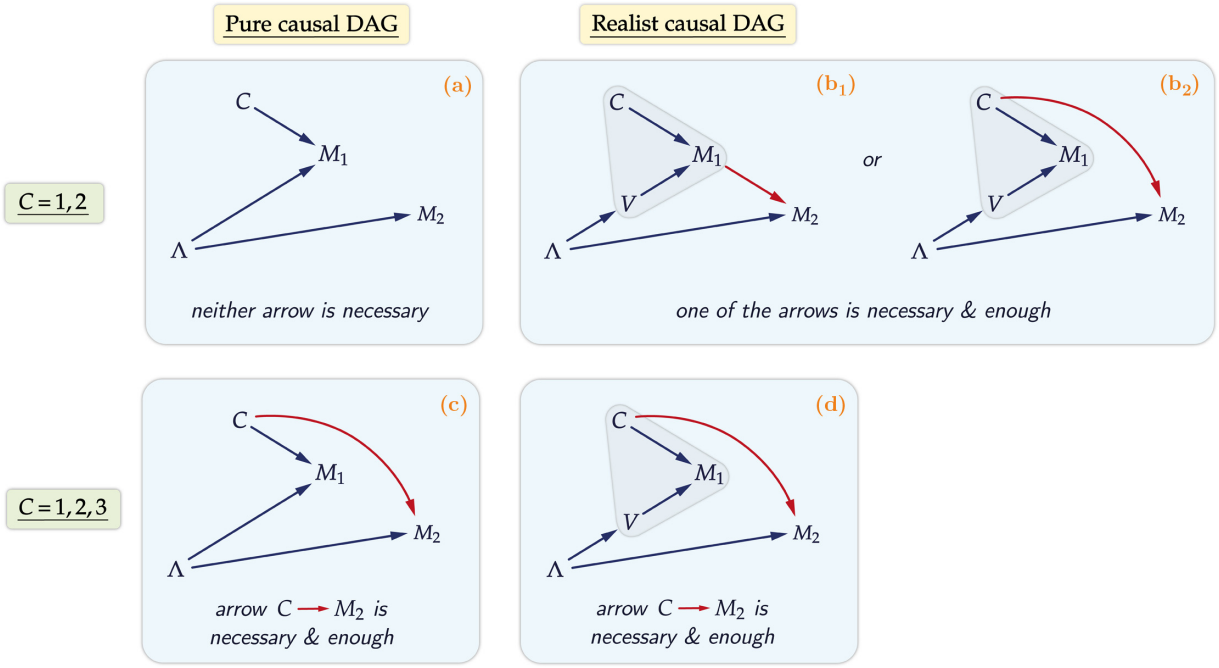

FIG. 4. Summary of the results. The table answers the question of which red arrows in the two causal diagrams in Fig. 3 can be removed while still retaining their capacity for generating the statistics in the three-box experiment in Fig. 2. There is a difference whether the full statistics is considered $C = 1, 2, 3$ or just its part related to the "paradoxical" choice of measurements $C = 1, 2$. This shows what kind of measurement disturbance (*outcome* vs *parameter dependence*) is required depending on the preferred worldview (pure vs realist) and the selection of measurements under consideration.

Accordingly, if the disturbance is taken on board, then the paradoxical correlations in the postselected regime can be explained as an instance of the *selection bias* [16] [see the proof of Fig. 4(b) in Appendix A].

Interestingly, in the pure causal setting (no realism) the paradox can be explained without measurement disturbance of any sort. In this case we showed, by constructing the explicit SCM, that confounding is just enough to explain the effect [see Fig. 4(a)], i.e.,

$$\text{No realism} \underset{C=1,2}{\not\Rightarrow} \text{measurement disturbance}.$$

Thus without claiming realism the question of measurement disturbance turns out immaterial, and hence the paradox becomes less of a surprise.

As noted, both conclusions above concern the original formulation of the paradox with just two boxes being considered for inspection ($C = 1, 2$).

Our discussion relies on a proper treatment of measurement disturbance as a genuine causal notion. This approach allows us to make a further distinction between various types thereof, that is, outcome vs parameter dependence represented by the respective arrows $M_1 \to M_2$ and $C \to M_2$ in Fig. 3. For this purpose the full statistics, which includes checking the third box ($C = 1, 2, 3$), appears to be more interesting. It allows us to show that parameter dependence is actually necessary in both pure and realist frameworks [see Figs. 4(c) and 4(d)], i.e.,

$$\text{Full statistics} \underset{C=1,2,3}{\Rightarrow} \text{parameter dependence}.$$

We also showed, by construction of the explicit SCM, that parameter dependence is sufficient to explain the observed statistics (this can be also deduced from the general property regarding saturation of the model by the single arrow $C \to M_2$, as proved in Ref. [33]).

Let us emphasize that a statement that an arrow is unnecessary does not mean that in reality it is not present (this only means that one can explain the statistics without its assistance). However, the necessity of an arrow implies that it cannot be replaced by any other arrow and still correctly reproduce the statistics. In this sense the causal DAGs in Fig. 4 are the minimal structures compatible with the observed statistics under the given realist or pure assumption.

V. DISCUSSION

In this paper we focused on the assessment of the role of the various causal mechanisms capable of generating the statistics observed in the quantum three-box paradox. Such an approach seems to be more revealing regarding the structure of measurement disturbance than the mere acknowledgment of its presence of some sort. The analysis based on the instrumental inequalities shows the necessity of the parameter dependence in the system (but interestingly, only when the full statistics is taken into account). Furthermore, the use of the d-separation criterion allows us to see the paradox as a case of the selection bias [16] (see Refs. [34,35] for a related issue of postselection in the definition of weak values [36]). Note that in our discussion we take the conservative point of view with single measurement outcomes (in contrast to the many-worlds interpretation) as well as exclude backward-in-time causation.

We remark that for the three-box paradox there is no physical principle that would prohibit the propagation of measurement disturbance in the experiment (like the locality principle for Bell experiments [1–3]). This is because the PPS paradigm consists of a sequence of two measurements $\mathcal{M}_1^c$ and $\mathcal{M}_2$, with the second one being made on the whole system (i.e., even if the boxes can be kept separate while checking a given box C in the measurement $\mathcal{M}_1^c$, all of them have to merge together at the end to implement the measurement $\mathcal{M}_2$). See Fig. 1. It is also clearly seen in the experimental realizations of the three-box paradox [37,38]. This makes all the information about the outcome $\mathcal{M}_1^c$ and the parameter C in principle available upon postselection, and hence neither of the disturbance arrows can be excluded by locality arguments. It is in fact possible to construct a generic local hidden variable model of a single particle in arbitrary linear optical circuits (where measurement disturbance propagates locally too) [39].

It is instructive to observe that the quantum description of the paradox, if interpreted literally in the causal terms, works according to the pure causal DAG in Fig. 3 with *both* arrows $M_1 \to M_2$ and $C \to M_2$ present. This is readily seen from the corresponding SCM which boils down to the identification $\Lambda \equiv \rho_0$ and the structural equations implementing the standard quantum recipe. See Appendix C for details. Note that such a quantumlike model is not optimal, as it features the unnecessary arrow $M_1 \to M_2$ (outcome dependence) [see Fig. 4(c) as well as Ref. [33]]. However, it is an interesting open question regarding the scope of scenarios in which outcome dependence can be taken out of the picture in favor of parameter dependence alone (note that in the standard Bell scenario either one of those arrows is just enough).

The aforementioned quantumlike causal model reflects the standard view of quantum theory which turns down the discussion of unobserved properties. It falls within the pure causal setting [see Fig. 3 (left)]. This should be compared with the concept of measurement as an act of observation which reveals some preexisting property of the system. The latter entails including additional structure in the causal modeling of the system as indicated in the realist causal setting [see Fig. 3 (right)]. It is not without consequences for the expressive power of the considered causal structures, which in turn may entail the necessity of certain mechanisms of measurement disturbance in the system (see the comparison in Fig. 4).

Let us note that the PPS paradoxes can be turned into the proofs of contextuality [40–43]. Since in principle all those effects can be attributed to measurement disturbance, it is natural to ask about the possible causal mechanisms allowing for this to be fully attained. This makes the question about the causal resources (here taken as different sorts of arrows) that are enough to explain a given class of contextual effects an interesting research problem. This paper provides a strong hint that parameter dependence is in general indispensable. However, we have also seen that it may be superfluous in some restricted settings (see case $C = 1, 2$ vs $C = 1, 2, 3$ in Fig. 4). We note in passing an interesting question regarding the role of signaling in our argument. Observe that the marginals of M_2 in the behavior in Fig. 2 change only when $C = 3$ is also considered, and it is where the parameter dependence in the three-box statistics can be proved (this should be compared with the nonsignaling Popescu-Rohrlich boxes for which the outcome dependence is just enough). We leave it as a curious remark regarding contextuality in the presence of signaling [44,45].

In conclusion, we mention the existence of several related PPS paradoxes discussed in the literature, e.g., Refs. [46–51]. Their resemblance to the three-box paradox suggests that the causal approach might shed more light there too.

ACKNOWLEDGMENTS

We acknowledge helpful discussions with J. Duda, M. Markiewicz, and R. Staszewski. We especially thank E. Wolfe for bringing Ref. [33] to our attention and comments on the overall framework of this paper.

APPENDIX A: PROOFS OF NECESSITY

Here we answer the question of which of the two red arrows in the causal DAGs in Fig. 3 is necessary in order to reproduce the statistics in Fig. 2. For the proofs we use modern tools of causal inference [16], i.e., the d-separation criterion [Fig. 4(b)] and instrumental inequalities [Figs. 4(c) and 4(d)].

Proof of Fig. 4(b). Let us consider the realist causal framework with just two measurement choices $C = 1, 2$. In that case, from Eq. (4), we have $P(M_1 = 1|M_2 = 1, C) = 1$. As a consequence of the assumption Eq. (5) it means that $V = C$ whenever $M_2 = 1$. This necessitates the conditional dependence $V \not\perp\!\!\!\perp C \,|\, M_2$. However, in the realist causal DAG with both arrows $M_1 \to M_2$ and $C \to M_2$ absent [see Fig. 3 (right)], the variables C and V are d-separated conditioned on M_2 (since the only path $C \to M_1 \leftarrow V$ is blocked by the collider M_1), which entails their statistical independence $V \perp\!\!\!\perp C \,|\, M_2$. This contradiction means that at least one of those arrows must be present in the realist causal DAG. Indeed, either arrow $M_1 \to M_2$ or $C \to M_2$ lifts the d-separation by opening the respective path $V \leftarrow \Lambda \to M_2 \leftarrow M_1 \leftarrow C$ or $V \leftarrow \Lambda \to M_2 \leftarrow C$ (here M_2 is not a collider because of conditioning). This opens a way for both variables to become correlated (whether it is enough with a single arrow is proved by an explicit SCM in Appendix B). In the field of causal inference the phenomenon of correlation due to conditioning is known as a *selection bias* or *Berkson's paradox* [16].

Proof of Figs. 4(c) and 4(d). Now we are concerned with the full choice of measurements $C = 1, 2, 3$. For the proof it is enough to consider the pure causal DAG, i.e., Fig. 4(c) [since this framework is more permissive, the necessity for Fig. 4(c) automatically carries over to the realist case Fig. 4(d)].

Let us reformulate our problem in causal inference terms. Suppose we admit the arrow $M_1 \to M_2$ in the pure causal DAG [see Fig. 3 (left)] and ask about the necessity of the arrow $C \to M_2$. (Note that proving the necessity of arrow $C \to M_2$ in this case will entail its necessity when the arrow $M_1 \to M_2$ is missing, since adding the latter only increases the expressive power of the DAG.) This can be recast as the instrumental scenario [16,52], where the instrument C is used for determining causal influence of M_1 on M_2, both of which are affected by some unobserved (or latent) variable Λ. The crucial assumption in this scenario is the absence of any arrow incoming or outgoing from C except for $C \to M_1$.

It appears that this assumption can be tested by the so-called instrumental inequalities [16,52] which adopted to our case take the form

$$\max_i \sum_j \left[\max_k P(M_1=i, M_2=j|C=k)\right] \leqslant 1, \quad \text{(A1)}$$

where $i, j = 0, 1$ and $k = 1, 2, 3$. Violation of those inequalities by the observed statistics testifies to the presence of some other arrow incoming or outgoing from C (whatever the character of the arrow $M_1 \to M_2$). In our case this could be only the arrow $C \to M_2$, and hence a way to check its necessity.

To unfold the condition Eq. (A1) we observe that it is equivalent to the following set of equations:

$$\begin{aligned}
&P(M_1=0, M_2=0|C=k) + P(M_1=0, M_2=1|C=l) \leqslant 1,\\
&P(M_1=1, M_2=0|C=k) + P(M_1=1, M_2=1|C=l) \leqslant 1,\\
&P(M_1=0, M_2=1|C=k) + P(M_1=0, M_2=0|C=l) \leqslant 1,\\
&P(M_1=1, M_2=1|C=k) + P(M_1=1, M_2=0|C=l) \leqslant 1,
\end{aligned} \quad \text{(A2)}$$

where $kl = 12, 13, 23$. Now it is straightforward to check that the statistics in Fig. 2 violates those inequalities for $kl = 13$ and 23. For example, for $kl = 23$ in the first line in Eq. (A2) we get

$$2/3 + 4/9 = 10/9 > 1. \quad \text{(A3)}$$

Therefore, if the full choice of measurements $C = 1, 2, 3$ is considered, then the arrow $C \to M_2$ must be present in the causal DAG in Fig. 3 (either pure or realist), if it is to reproduce the statistics in Fig. 2 (this single arrow is also enough as proved by an explicit SCM in Appendix B). We note that for the limited choice $C = 1, 2$ (i.e., $kl = 12$) the instrumental inequalities Eq. (A2) remain inviolate, which means that in that case the arrow is unnecessary [in full agreement with Figs. 4(a) and 4(b_1)].

APPENDIX B: PROOFS OF SUFFICIENCY

Here we give a repository of explicit SCMs proving the sufficiency of the considered causal DAGs for each case in Fig. 4. In the following we just specify the structural equations for the respective SCMs and observe that the joint probability distributions $P(M_1=i, M_2=j|C=k)$ are obtained by a straightforward application of the product decomposition (Markov property) for the associated causal DAGs.

Proof. Note that the joint probability distribution generated by the causal DAG in Fig. 4(a) has the following decomposition:

$$\begin{aligned}
&P(M_1=i, M_2=j|C=k)\\
&= \sum_\lambda P(M_1=i|C=k, \Lambda=\lambda)P(M_2=j|\lambda)P(\Lambda=\lambda).
\end{aligned} \quad \text{(B1)}$$

Now, in order to recover the statistics in Fig. 2 for $C = 1, 2$ it is enough to consider a Bernoulli distributed hidden variable $\Lambda \sim \text{Ber}(1/3)$, i.e., we have $\Lambda = 0, 1$ and

$$P(\Lambda = 0) = 2/3, \quad P(\Lambda = 1) = 1/3. \quad \text{(B2)}$$

Then, we set the following structural equations compatible with the diagram in Fig. 4(a):

$$M_1(C, \Lambda) := \Lambda, \quad \text{(B3)}$$

$$M_2(\Lambda, N) := \Lambda N, \quad \text{(B4)}$$

where $N \sim \text{Ber}(1/3)$ is an independent noise variable having a Bernoulli distribution. This defines the SCM, which via Eq. (B1) gives the correct statistics in Fig. 2 for $C = 1, 2$ (only).

Proof of Fig. 4(b_1). Let us start by justifying the sufficiency of the causal DAG in Fig. 4(b_1). In this case, the joint probability distribution has the following decomposition:

$$\begin{aligned}
&P(M_1=i, M_2=j|C=k)\\
&= \sum_{\lambda,v} P(M_1=i|C=k, V=v)P(M_2=j|M_1=i, \Lambda=\lambda)\\
&\qquad P(V=v|\Lambda=\lambda)P(\Lambda=\lambda).
\end{aligned} \quad \text{(B5)}$$

In this case we posit a uniformly distributed hidden variable $\Lambda \sim \text{Uni}(1, 3)$, i.e., we have $\Lambda = 1, 2, 3$ with

$$P(\Lambda = i) = 1/3 \qquad \text{for } i = 1, 2, 3. \quad \text{(B6)}$$

The following structural equations define an SCM compatible with the diagram in Fig. 4(b_1):

$$M_1(C, V) := \delta_{C,V}, \qquad \text{[see Eq. (5)]} \quad \text{(B7)}$$

$$V(\Lambda) := \Lambda, \quad \text{(B8)}$$

$$M_2(M_1, \Lambda) := M_1 N, \quad \text{(B9)}$$

where $N \sim \text{Ber}(1/3)$ is a noise variable with a Bernoulli distribution. This is enough to recover, via Eq. (B5), the statistics in Fig. 2 for $C = 1, 2$ (only).

The sufficiency of the DAG in Fig. 4(b_2) follows immediately from the model in Fig. 4(c) which works for all $C = 1, 2, 3$.

Proof of Fig. 4(c). Here the joint probability distribution generated by the causal DAG in Fig. 4(c) has the following decomposition:

$$\begin{aligned}
&P(M_1=i, M_2=j|C=k)\\
&= \sum_\lambda P(M_1=i|C=k, \Lambda=\lambda)\\
&\qquad P(M_2=j|C=k, \Lambda=\lambda)P(\Lambda=\lambda).
\end{aligned} \quad \text{(B10)}$$

We need an SCM which reconstructs the statistics in Fig. 2 for $C = 1, 2, 3$ (all of them). Consider a uniformly distributed hidden variable $\Lambda \sim \text{Uni}(1, 3)$ which takes three values $\Lambda = 1, 2, 3$. Then we define the following set of structural equations in accord with the diagram in Fig. 4(c):

$$M_1(C, \Lambda) := \delta_{C,\Lambda}, \qquad \text{[see Eq. (5)]} \quad \text{(B11)}$$

$$M_2(C, \Lambda) := \begin{cases} \delta_{C,\Lambda} N & \text{for } C = 1, 2,\\ (1-\delta_{C,\Lambda})(1-N) + \delta_{C,\Lambda} N & \text{for } C = 3, \end{cases} \quad \text{(B12)}$$

where $N \sim \text{Ber}(1/3)$ is a noise variable with a Bernoulli distribution. Such a definition of the SCM recovers, via Eq. (B10), the full statistics in Fig. 2 for all $C = 1, 2, 3$.

Proof of Fig. 4(d). This case is a straightforward extension of the SCM in Fig. 4(c). Here the causal DAG in Fig. 4(d) entails decomposition of the joint probability distribution in the form

$$P(M_1=i, M_2=j|C=k) = \sum_{\lambda,v} P(M_1=i|C=k, V=v)P(M_2=j|C=k, \Lambda=\lambda) P(V=v|\Lambda=\lambda)P(\Lambda=\lambda). \quad \text{(B13)}$$

Following the constriction in Fig. 4(c) we take the hidden variable $\Lambda = 1, 2, 3$ with a uniform distribution $\Lambda \sim \text{Uni}(1, 3)$. Then we introduce an additional variable $V = 1, 2, 3$ and equate it with Λ. This trivial extension provides the required SCM which takes the following explicit form [see Eqs. (B11) and (B12)]:

$$M_1(C, V) := \delta_{C,V}, \quad \text{[see Eq. (5)]} \quad \text{(B14)}$$

$$V(\Lambda) := \Lambda, \quad \text{(B15)}$$

$$M_2(C, \Lambda) := \begin{cases} \delta_{C,\Lambda} N & \text{for } C = 1, 2, \\ (1-\delta_{C,\Lambda})(1-N) + \delta_{C,\Lambda} N & \text{for } C = 3, \end{cases} \quad \text{(B16)}$$

where $N \sim \text{Ber}(1/3)$ is again a Bernoulli noise variable. Those structural equations comply with the diagram in Fig. 4(d) reconstructing, via Eq. (B13), the full statistics in Fig. 2 for all $C = 1, 2, 3$.

APPENDIX C: QUANTUMLIKE CAUSAL MODEL

The formalism of quantum theory in the considered sequential scenario indicates the following. The outcomes of the first measurement $\mathcal{M}_1^c$ are distributed as

$$P(M_1 = i|C) = \text{Tr}\left[\Pi_i^C \rho_0\right]. \quad \text{(C1)}$$

Then the initial quantum state ρ_0 gets updated according to the von Neumann–Lüders rule and the conditional distribution of the outcomes of the second measurement $\mathcal{M}_2$ is given by

$$P(M_2 = j|M_1 = i, C) = \frac{\text{Tr}\left[\Pi_j^{\text{post}} \Pi_i^C \rho_0 \Pi_i^C\right]}{\text{Tr}[\Pi_i^C \rho_0]}. \quad \text{(C2)}$$

Note that this is how Eq. (2) is derived.

Let us construct an SCM as suggested by Eqs. (C1) and (C2). It will have the structure of the pure causal DAG in Fig. 3 (left) with *both* arrows $M_1 \to M_2$ and $C \to M_2$ present, i.e., admit the following decomposition of the joint probability distribution:

$$P(M_1=i, M_2=j|C=k) = \sum_{\lambda} P(M_1=i|C=k, \Lambda=\lambda) P(M_2=j|C=k, M_1=i, \Lambda=\lambda)P(\Lambda=\lambda). \quad \text{(C3)}$$

The obvious choice follows from the trivial identification of the hidden variable Λ with the quantum state $\Lambda \equiv \rho$. In our case, the system is prepared in a given initial state ρ_0, so we have $P(\Lambda = \rho_0) = 1$. Now, the model needs to recover the probabilistic behavior in Eqs. (C1) and (C2). This can be attained with the help of two noise variables N_1 and N_2 with a uniform distribution over interval [0,1]. They are assumed to be independent and correspond, respectively, to M_1 and M_2. Thus we define the following structural equations:

$$M_1(C, \Lambda) := \begin{cases} 0 & \text{if } N_1 < \text{Tr}\left[\Pi_0^C \rho_0\right], \\ 1 & \text{otherwise,} \end{cases} \quad \text{(C4)}$$

and

$$M_2(C, M_1, \Lambda) := \begin{cases} 0 & \text{if } N_2 < \frac{\text{Tr}\left[\Pi_0^{\text{post}} \Pi_{M_1}^C \rho_0 \Pi_{M_1}^C\right]}{\text{Tr}\left[\Pi_{M_1}^C \rho_0\right]}, \\ 1 & \text{otherwise.} \end{cases} \quad \text{(C5)}$$

Clearly, such a defined SCM agrees with the structure of the pure causal DAG in Fig. 3 (left) with *both* arrows $M_1 \to M_2$ and $C \to M_2$ present. Furthermore, the statistics generated by the model, as obtained from Eq. (C3), recovers the desired distribution in Eq. (2) or equivalently Eqs. (C1) and (C2).

[1] J. S. Bell, *Speakable and Unspeakable in Quantum Mechanics* (Cambridge University, Cambridge, England, 1987).

[2] N. Brunner, D. Cavalcanti, S. Pironio, V. Scarani, and S. Wehner, Bell nonlocality, Rev. Mod. Phys. **86**, 419 (2014).

[3] V. Scarani, *Bell Nonlocality* (Oxford University, New York, 2019).

[4] S. Kochen and E. Specker, The problem of hidden variables in quantum mechanics, J. Math. Mech. **17**, 59 (1967).

[5] J. Thompson, P. Kurzyński, S.-Y. Lee, A. Soeda, and D. Kaszlikowski, Recent Advances in Contextuality Tests, Open Syst. Inf. Dyn. **23**, 1650009 (2016).

[6] C. Budroni, A. Cabello, O. Gühne, M. Kleinmann, and J. Å. Larsson, Quantum contextuality, arXiv:2102.13036.

[7] Y. Aharonov and L. Vaidman, Complete description of a quantum system at a given time, J. Phys. A **24**, 2315 (1991).

[8] Y. Aharonov, P. G. Bergmann, and J. L. Lebowitz, Time symmetry in the quantum process of measurement, Phys. Rev. **134**, B1410 (1964).

[9] R. E. Kastner, The three-box "paradox" and other reasons to reject the counterfactual usage of the ABL rule, Found. Phys. **29**, 851 (1999).

[10] L. Vaidman, The meaning of elements of reality and quantum counterfactuals: Reply to Kastner, Found. Phys. **29**, 865 (1999).

[11] J. Finkelstein, What is paradoxical about the "Three-box paradox"? arXiv:quant-ph/0606218.

[12] K. A. Kirkpatrick, Classical three-box "paradox," J. Phys. A: Math. Gen. **36**, 4891 (2003).

[13] T. Ravon and L. Vaidman, The three-box paradox revisited, J. Phys. A **40**, 2873 (2007).

[14] K. A. Kirkpatrick, Reply to 'The three-box paradox revisited' by T. Ravon and L. Vaidman, J. Phys. A **40**, 2883 (2007).

[15] O. J. E. Maroney, Measurements, disturbances and the quantum three box paradox, Stud. Hist. Phil. Mod. Phys. **58**, 41 (2017).
[16] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. (Cambridge University, Cambridge, England, 2009).
[17] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect* (Basic, New York, 2018).
[18] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, 2nd ed. (MIT, Cambridge, MA, 2000).
[19] J. Pearl, M. Glymour, and N. P. Jewell, *Causal Inference in Statistics: A Primer* (Wiley, New York, 2016).
[20] M. A. Hernan and J. M. Robins, *Causal Inference: What If* (Chapman and Hall, London, 2020).
[21] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference Foundations and Learning Algorithms* (MIT, Cambridge, MA, 2017).
[22] D. B. Rubin and G. W. Imbens, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge University, Cambridge, England, 2015).
[23] C. J. Wood and R. W. Spekkens, The lesson of causal discovery algorithms for quantum correlations: Causal explanations of Bell-inequality violations require fine-tuning, New J. Phys. **17**, 033002 (2015).
[24] R. Chaves, R. Kueng, J. B. Brask, and D. Gross, Unifying Framework for Relaxations of the Causal Assumptions in Bell's Theorem, Phys. Rev. Lett. **114**, 140403 (2015).
[25] K. Ried, M. Agnew, L. Vermeyden, D. Janzing, R. W. Spekkens, and K. J. Resch, A quantum advantage for inferring causal structure, Nat. Phys. **11**, 414 (2015).
[26] M. Ringbauer, C. Giarmatzi, R. Chaves, F. Costa, A. G. White, and A. Fedrizzi, Experimental test of nonlocal causality, Sci. Adv. **2**, e1600162 (2016).
[27] J.-M. A. Allen, J. Barrett, D. C. Horsman, C. M. Lee, and R. W. Spekkens, Quantum Common Causes and Quantum Causal Models, Phys. Rev. X **7**, 031021 (2017).
[28] R. Chaves, G. Carvacho, I. Agresti, V. Di Giulio, L. Aolita, S. Giacomini, and F. Sciarrino, Quantum violation of an instrumental test, Nat. Phys. **14**, 291 (2018).
[29] R. Chaves, G. B. Lemos, and J. Pienaar, Causal Modeling the Delayed-Choice Experiment, Phys. Rev. Lett. **120**, 190401 (2018).
[30] P. Blasiak, E. Borsuk, and M. Markiewicz, On safe post-selection for Bell tests with ideal detectors: Causal diagram approach, Quantum **5**, 575 (2021).
[31] P. Blasiak, E. M. Pothos, J. M. Yearsley, C. Gallus, and E. Borsuk, Violations of locality and free choice are equivalent resources in Bell experiments, Proc. Natl. Acad. Sci. USA **118**, e2020569118 (2021).
[32] J. P. Jarrett, On the physical significance of the locality conditions in the Bell arguments, Nous **18**, 569 (1984).
[33] R. J. Evans, Graphs for margins of Bayesian networks, Scand. J. Statist. **43**, 625 (2016).
[34] C. Ferrie and J. Combes, How the Result of a Single Coin Toss Can Turn Out to be 100 Heads, Phys. Rev. Lett. **113**, 120404 (2014).
[35] L. Vaidman, Comment on "How the result of a single coin toss can turn out to be 100 heads", arXiv:1409.5386.
[36] Y. Aharonov, D. Z. Albert, and L. Vaidman, How the Result of a Measurement of a Component of the Spin of a Spin-1/2 Particle Can Turn Out to be 100, Phys. Rev. Lett. **60**, 1351 (1988).
[37] K. J. Resch, J. S. Lundeen, and A. M. Steinberg, Experimental realization of the quantum box problem, Phys. Lett. A **324**, 125 (2004).
[38] R. E. George, L. M. Robledo, O. J. E. Maroney, M. S. Blok, H. Bernien, M. L. Markham, D. J. Twitchen, J. J. L. Morton, G. A. D. D. Briggs, and R. Hanson, Opening up three quantum boxes causes classically undetectable wavefunction collapse, Proc. Natl. Acad. Sci. USA **110**, 3777 (2013).
[39] P. Blasiak, Local model of a qudit: Single particle in optical circuits, Phys. Rev. A **98**, 012118 (2018).
[40] M. S. Leifer and R. W. Spekkens, Logical pre- and post-selection paradoxes, measurement-disturbance and contextuality, Int. J. Theor. Phys. **44**, 1977 (2005).
[41] M. S. Leifer and R. W. Spekkens, Pre- and Post-Selection Paradoxes and Contextuality in Quantum Mechanics, Phys. Rev. Lett. **95**, 200405 (2005).
[42] M. F. Pusey, Anomalous Weak Values Are Proofs of Contextuality, Phys. Rev. Lett. **113**, 200401 (2014).
[43] M. F. Pusey and M. S. Leifer, Logical pre- and post-selection paradoxes are proofs of contextuality, EPTCS **195**, 295 (2015).
[44] E. N. Dzhafarov, J. V. Kujala, and J.-A. Larsson, Contextuality in three types of quantum-mechanical systems, Found. Phys. **45**, 762 (2015).
[45] J. V. Kujala, E. N. Dzhafarov, and J.-A. Larsson, Necessary and Sufficient Conditions for an Extended Noncontextuality in a Broad Class of Quantum Mechanical Systems, Phys. Rev. Lett. **115**, 150401 (2015).
[46] Y. Aharonov and L. Vaidman, How one shutter can close N slits, Phys. Rev. A **67**, 042107 (2003).
[47] N. Aharon and L. Vaidman, Quantum advantages in classically defined tasks, Phys. Rev. A **77**, 052310 (2008).
[48] L. Vaidman, Past of a quantum particle, Phys. Rev. A **87**, 052104 (2013).
[49] Y. Aharonov, S. Popescu, D. Rohrlich, and P. Skrzypczyk, Quantum Cheshire Cats, New J. Phys. **15**, 113015 (2013).
[50] Y. Aharonov, F. Colombo, S. Popescu, I. Sabadini, D. C. Struppa, and J. Tollaksen, Quantum violation of the pigeonhole principle and the nature of quantum correlations, Proc. Natl. Acad. Sci. USA **113**, 532 (2016).
[51] Y. Aharonov, E. Cohen, A. Landau, and A. C. Elitzur, The case of the disappearing (and re-appearing) particle, Sci. Rep. **7**, 531 (2017).
[52] J. Pearl, On the testability of causal models with latent and instrumental variables, in *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, 1995), pp. 435–443.