From locality to factorizability: a novel escape from Bell's theorem

G. S. Ciepielewski and E. Okon
Universidad Nacional Autónoma de México, Mexico City, Mexico.

While initial versions of Bell's theorem captured the notion of locality with the assumption of *factorizability*, in later presentations, Bell argued that factorizability could be derived from the more fundamental *principle of local causality*. Here we show that, contrary to what is commonly assumed, in order to derive factorizability from the principle of local causality, a non-trivial assumption, similar but strictly independent of settings independence, is required. Loosely speaking, such an extra assumption demands independence between the states of the measurement apparatuses. We conclude that it is possible to construct a model, satisfying both the principle of local causality and settings independence, but that, in virtue of violating this additional assumption—and thus factorizability—is able to break Bell's inequality.

1 Introduction

Through his famous theorem, John Bell challenged the capacity of local models to reproduce the predictions of quantum mechanics. In early versions of the theorem, Bell imposed the notion of locality with the assumption of *factorizability*—in broad terms, the idea that correlations between distant systems must have local explanations. However, in later presentations, he argued that factorizability could be derived from his more general *principle of local causality*.

In this work, we show that, contrary to what is commonly accepted, to derive factorizability from the principle of local causality, an additional, non-trivial assumption is required. Loosely speaking, such an extra assumption demands independence between the states of the two measurement apparatuses. As such, the new assumption is similar, but strictly independent of *settings independence*—an auxiliary assumption in Bell's theorem, demanding statistical independence between measurement settings and systems to be measured.

Given that, to derive factorizability from the principle of local causality, an extra assumption is required, we end up concluding that it is possible to construct a model, satisfying both the principle of local causality and settings independence, but which, in virtue of violating this additional assumption—and thus factorizability—can break Bell's inequality. Therefore, in light of experimental violations of the inequality, in the same way that one can retain locality by breaking settings independence (see Ciepielewski et al. (2021)), one could retain both locality and settings independence by violating the new assumption; of course, as with violations of settings independence, the question of how *reasonable* it is to violate the new assumption is independent of the purely logical point that it is possible to do so.

In order to make our case, our manuscript is organized as follows. In section 2, we review the standard story regarding Bell's theorem and the connection between the principle of local causality and factorizability. Then, in section 3, we examine and criticize well-known discussions regarding such a connection. Next, in section 4, containing the main result of this work, we show that in order to derive factorizability from the principle of local causality, an additional assumption is required. We also explore the nature of the new assumption and the possibility of violating it. Finally, in section 5, we offer our conclusions.

2 The standard story

Bell's construction of the theorem starts with an ensemble of pairs of particles in the singlet state. The particles of each pair are then sent to two spatially separated regions, 1 and 2, where spin measurements are performed. We denote by a, b the spin directions measured in 1 and 2, respectively, and by A, B the corresponding results (with spin-up corresponding to $+1$ and spin-down to -1). Next, Bell denotes by λ the complete state of each pair. If quantum mechanics is complete, then λ would be given by the quantum state; if quantum mechanics is not complete, then λ could be the quantum state supplemented with extra variables or it could be something entirely different. Either way, Bell does not impose determinism on the models considered.

Since Bell is interested in exploring the viability of local models, he assumes that the models under consideration have such a feature. To characterize them, he demands the probabilities $P\left(A, B|a, b, \lambda\right)$, predicted by such models for the experiment in question, to satisfy

$$P\left(A, B|a, b, \lambda\right) = P\left(A|a, \lambda\right) P\left(B|b, \lambda\right). \tag{1}$$

This condition, usually referred to as *factorizability*, is intended to capture the idea that, for local theories, all correlations between distant systems must have local explanations. That is, once one conditionalizes on the complete state λ of a pair, correlations between distant measurements completely disappear.

Now, given that Bell considers measurements over an ensemble of pairs, such an ensemble can be characterized by a distribution over the fundamental states, $\rho(\lambda)$. The point is that, even though all pairs are prepared in the singlet state, the complete description given by λ may change from pair to pair. Bell then assumes that the values of λ and the settings a and

b are statistically independent, that is

$$\rho(\lambda|a,b) = \rho(\lambda). \tag{2}$$

We call this condition *settings independence*. Intuitively, it entails that the settings a, b and λ are not correlated. Settings independence seems reasonable because one can set up things in such a way that the settings can be chosen before, during, or after the generation of the pair, and they can be chosen using a variety of methods. Moreover, an analogous assumption is (implicitly) accepted in all experimental scenarios across all sciences.

To compose the theorem, Bell shows that, for any model satisfying factorizability (and settings independence), the expectation value of the product AB over the whole ensemble,

$$E(a,b) = \int \sum_{A,B} A\,B\,P(A,B|a,b,\lambda)\,\rho(\lambda|a,b)\,d\lambda, \tag{3}$$

necessarily obeys

$$|E(a,b) + E(a,b') + E(a',b) - E(a',b')| \leq 2. \tag{4}$$

That is, the predictions of all models satisfying factorizability (and settings independence), necessarily satisfy this inequality.

Next, Bell considers the quantum prediction for the expectation value in Eq. (3), which is given by

$$E^{QM}(a,b) = -\cos(\theta) \tag{5}$$

with θ the angle between a and b. Moreover, he notes that, if one takes a, a', b, b' on the same plane, with a 90^{o} angle between a and a' and b and b', and a 45^{o} angle between a and b, then

$$|E^{QM}(a,b) + E^{QM}(a,b') + E^{QM}(a',b) - E^{QM}(a',b')| = 2\sqrt{2}. \tag{6}$$

It is clear, then, that quantum mechanics and models satisfying factorizability (and settings independence), make different predictions for the experiment in question.

The final step is to consider experimental realizations of Bell's scenario. Those experiments have been done and have established clear violations of the inequality (Aspect et al., 1981, 1982; Weihs et al., 1998; Giustina et al., 2015; Shalm et al., 2015; Hensen et al., 2015). It seems, then, that models satisfying factorizability (and settings independence) are unable to correctly describe our world. These, in a nutshell, are Bell's theorem and its highly non-trivial implications.

Above we saw that the key condition employed by Bell to derive the inequality is factorizability. Initially, such a condition was assumed by Bell as a defining characteristic of local theories. However, in later presentations of the theorem, Bell argued that factorizability could be *derived* from a more fundamental postulate, which he calls the *principle of local causality* (see, e.g., Bell (1976, 1990)).

According to Bell's principle of local causality, a model is local if the probability it assigns to q_χ, the value of some physical quantity at spacetime event χ, is such that

$$P(q_\chi|\lambda_\sigma) = P(q_\chi|\lambda_\sigma, q_\xi), \tag{7}$$

with λ_σ a complete specification of the physical state on σ, a spatial slice of the past light cone of χ, and q_ξ the value of any physical quantity on ξ, an event spacelike separated from χ and outside of the causal future of σ (see Figure 1). In words, the predictions that a local model makes for an event, given complete information on a slice of its past light cone, do not change with additional information about regions outside of the future of that slice.

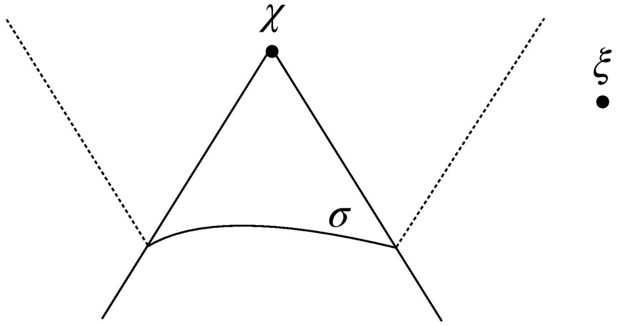

Figure 1: According to Bell's principle of local causality, a theory is local if $P(q_\chi|\lambda_\sigma) = P(q_\chi|\lambda_\sigma, q_\xi)$.

Bell makes two important comments regarding his principle. First, it is crucial for ξ to lie outside of the causal future of σ. If not, for a local, indeterministic theory, ξ could contain information about a stochastic process to the future of σ, but in the common past of χ and ξ, that could improve the prediction of the model for χ. Bell also stresses that it is essential for the specification of the physical state on σ to be complete. Otherwise, even for local models, ξ could contain part of the missing information, which could help improve the prediction.

As we said above, Bell claimed, and it is generally accepted (see, e.g., Bell (1976, 1990); Goldstein et al. (2011); Maudlin (2011); Myrvold et al. (2021); Ciepielewski et al. (2021)), that factorizability follows from the principle of local causality. The connection is argued for along these lines. One considers the joint probability $P(A,B|a,b,\lambda_\Sigma)$, with λ_Σ the complete state over Σ (see Figure 2), and uses the product rule to write it as

$$P(A,B|a,b,\lambda_\Sigma) = P(A|a,b,B,\lambda_\Sigma)P(B|a,b,\lambda_\Sigma). \tag{8}$$

Next, one employs the principle of local causality to note that

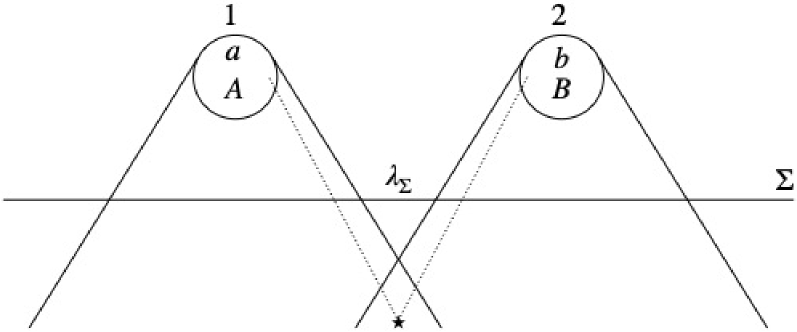

Figure 2: Space-time diagram of Bell's experimental scenario.

$$P(A|a,b,B,\lambda_\Sigma) = P(A|a,\lambda_\Sigma) \quad \text{and} \quad P(B|a,b,\lambda_\Sigma) = P(B|b,\lambda_\Sigma), \tag{9}$$

from which it follows that

$$P(A,B|a,b,\lambda_\Sigma) = P(A|a,\lambda_\Sigma)P(B|b,\lambda_\Sigma). \tag{10}$$

This looks like factorizability, but it isn't because, while factorizability is written in terms of λ, the complete state of the pair, the equation above is written in terms of λ_Σ, the complete state over Σ. To obtain factorizability, one then argues that λ_Σ can be decomposed into two (not necessarily independent) parts: λ and λ_E. As a result, Eq. (8) is written as

$$P(A,B|a,b,\lambda,\lambda_E) = P(A|a,\lambda,\lambda_E)P(B|b,\lambda,\lambda_E). \tag{11}$$

By explicitly having in the conditional a, b, λ and λ_E, it is finally argued that one can remove λ_E from the above equation. The idea is that, if one knows the complete state of the pair, λ, and one knows what is going to be measured, a and b, then all additional information becomes irrelevant. With this, one finally arrives at

$$P(A, B|a, b, \lambda) = P(A|a, \lambda)P(B|b, \lambda), \tag{12}$$

that is, factorizability.

3 A revision of the literature

Before delving, in the next section, into a detailed analysis of the connection between factorizability and the principle of local causality, in this section we examine a number of well-known discussions regarding the issue.

3.1 La Nouvelle Cuisine

We start this revision by exploring the use that Bell himself, in *La Nouvelle Cuisine* (Bell, 1990), gives to the principle of local causality for the derivation of the theorem. After stating the principle, and setting up the experimental scenario, Bell considers the joint probability of the results, A and B, conditional on four different terms: a, b, λ and c; i.e., $P(A, B|a, b, c, \lambda)$. The first three terms on the conditional are, as expected, the settings and the state of the pair. Regarding c, Bell takes it to represent *values of other variables describing the experiment*. However, he also assumes that c and λ give a specification which is *complete*, at least for the intersection of Σ with the union of the past light cones of regions 1 and 2 (see Figure 2). It seems, then, that counter to what Bell alleges, c must contain much more than a description of the experiment and is much closer to what we called λ_E. That is, c contains whatever is needed, besides λ, to make the description complete.

Next, Bell writes the joint probability he is considering as

$$P(A, B|a, b, c, \lambda) = P(A|B, a, b, c, \lambda)P(B|a, b, c, \lambda), \tag{13}$$

and, using the fact that the conjunction of λ and c is assumed to be complete, he employs the principle of local causality to arrive at

$$P(A, B|a, b, c, \lambda) = P(A|a, c, \lambda)P(B|b, c, \lambda). \tag{14}$$

At that point, and in contrast to what was done at the end of section 2, Bell does not attempt to remove c from the conditional. Instead, he keeps it in all conditionals throughout the derivation of the theorem. For instance, he writes the settings independence assumption as

$$\rho(\lambda|a,b,c)=\rho(\lambda|c). \tag{15}$$

Moreover, he assumes that c remains constant during the whole sequence of experiments. In fact, when he is done, he writes the inequality as

$$|E(a,b,c)-E(a,b',c)|+|E(a',b,c)+E(a',b',c)|\leq 2. \tag{16}$$

There is, however, a problem with retaining c for the whole derivation and, in particular, with assuming it constant throughout the complete sequence of measurements. As we argued above, since the information provided by c and λ is assumed complete (which is what allowed for the principle of local causality to be used to arrive at Eq. (14)), c is required to contain much more information than the experimental arrangement. For instance, it must contain information regarding the temperature of the lab, the time of day or the material of the shoes of the experimentalists. It is fully unwarranted, then, to expect for c not to change from run to run. We conclude, then, that Bell's derivation of the theorem, as presented in *La Nouvelle Cuisine*, is simply invalid, as it depends on the invalid assumption that c remains constant.

What about letting c vary from run to run? That also leads to trouble. The issue is that, in that case, one would have to average over the distribution of c, call it $\mu(c)$, over the whole ensemble. However, given the definition of c, it is clear that such a distribution will *not* be statistically independent of the settings a and b, i..e, $\mu(c|a,b)\neq\mu(c)$. And, in the same way that, in the absence of settings independence, a possible dependence of λ on the settings blocks the derivation of the inequality, a dependence of c on a and b prevents the construction of the theorem (in fact, a similar averaging procedure was proposed in Bell (1976), but criticized along the ideas of this paragraph in Shimony et al. (1976)).

3.2 Scholarpedia

Next, we explore the discussion of the relation between factorizability and the principle of local causality in the popular Scholarpedia entry on Bell's theorem (Goldstein et al., 2011). As we did in section 2, (and translating everything into our notation), they start with the identity[1]

$$P(A,B|a,b,\lambda_\Sigma)=P(A|a,b,B,\lambda_\Sigma)P(B|a,b,\lambda_\Sigma) \tag{17}$$

[1] Instead of using the surface Σ to screen off, Goldstein et al. (2011) employs a wider region (see their figure 2). Such a difference is not relevant for the following discussion.

and use the principle of local causality to arrive at

$$P(A, B|a, b, \lambda_\Sigma) = P(A|a, \lambda_\Sigma)P(B|b, \lambda_\Sigma). \tag{18}$$

They readily acknowledge that such equality looks like factorizability, but it isn't. They recognize that λ_Σ includes much more than λ and that, while it is reasonable to assume that λ is independent of the settings, it is not reasonable to assume that λ_Σ is.

At this point, they claim that the principle of local causality alone is not sufficient to prove that there exists a λ, a subset of λ_Σ, which is independent of a and b and satisfies factorizability. The claim, in particular, is that the existence of this λ also depends on the settings independence condition, which they call "no conspiracy". Moreover, they argue that by depending on the distinction between settings, controllable by experimenters, and other variables, the "no conspiracy" assumption involves anthropocentric elements. For this reason, they think, it is not possible to formalize the "no conspiracy" condition, nor to give a mathematical proof that, for a "non-conspiratorial" local theory, there exists a λ which is independent of the settings and satisfies factorizability. Still, without any pretension of mathematical precision, they offer the following argument to the effect that, for a "non-conspiratorial" local theory, a λ with such properties does exist.

The starting point of the argument is to distinguish between different subsets of λ_Σ. First, there are those values that are irrelevant for the experiment and can be ignored. Of those that are relevant, some, call them α, will determine or influence the setting a and some, call them β, the setting b. Finally, call whatever is left and relevant λ_S. Next, they notice that, for a "non-conspiratorial" theory, one must be able to define α, β and λ_S such that λ_S is independent of a and b. Finally, to arrive at the claim that, if the theory is local, then factorizability must hold for this λ_S, they argue as follows. Since only λ_S and α are relevant for A and since α is relevant to A only through a, then the motivation behind the principle of local causality leads to the conclusion that $P(A|a, b, B, \lambda_S) = P(A|a, \lambda_S)$ (and analogously for region 2). If this is so, factorizability follows.

What are we to make of the discussion? We start by pointing out that, according to Goldstein et al. (2011), what needs to be proven is not factorizability alone, but "the existence of a subset λ of the data codified by $[\lambda_\Sigma]$ that is independent of $[a, b]$ and for which [factorizability] holds". That is, from the unset, they formulate the issue in a way that factorizability and settings independence, or "no-conspiracy" as they call it, are linked. We believe that this approach only muddles the waters as factorizability and settings independence are two fully independent conditions. That is, it is perfectly possible to have either one without the other. For instance, the model in Ciepielewski et al. (2021) satisfies factorizability but forbids the validity of settings independence, and pilot-wave theory violates factorizability but allows for settings independence to be the case.

Next, it is argued that settings independence involves anthropocentric elements, so it cannot be cleanly defined mathematically. We disagree. While it is clear that “setting” is an anthropocentric concept, that does not mean that the settings independence assumption cannot be defined precisely. As we mentioned when we introduced it, $\rho(\lambda)$ is the actual distribution of physical states λ over the ensemble of measured pairs—and not a probability distribution. Moreover, settings independence consists of the assumption that such a distribution and the measurement settings, a and b, are statistically independent: $\rho(\lambda|a,b) = \rho(\lambda)$. That is, one assumes that, if out of the whole ensemble of measured pairs, one focuses on a subensemble of runs with any particular pair of settings, then the distribution of λ over that subensemble is the same as the distribution of the whole. It seems clear that, whether this is the case or not is a completely objective question, independent of the question of which variables are controllable by certain means (or even from whether there are controllable variables at all). That is, given a concrete experiment, and independently of how the settings were chosen, there will be an actual distribution of states and settings, and whether or not settings independence holds in that experiment is a completely objective question, with no anthropocentric elements rendering it subjective or vague.

It is important to add that the question of whether settings independence is satisfied or not is a question that can be asked of experiments, not models. Of course, there can be a models, such as the one in Ciepielewski et al. (2021), for which it is impossible to design a Bell-type experiment in which settings independence is the case. But that clearly does not mean that there are models for which settings independence is true for every experiment. For instance, if one is conducting a survey asking whether Real Madrid or Barcelona is the best football club in Spain, but conducts the survey in Barcelona, then the analog of settings independence will not be satisfied for that experiment—a fact that has little to do with any underlying model being able or not to satisfy settings independence.

Coming back to the discussion in Goldstein et al. (2011), we saw that, since factorizability and settings independence are blended, and since it is concluded that settings independence cannot be formalized, they settle for an informal argument. The starting point of that argument is the separation of λ_Σ into four subsets: the irrelevant sector, α influencing a, β influencing b, and the rest, which we called λ_S. After this, they argue that, for a “non-conspiratorial” model, λ_S thus defined is independent of a and b.

There are, however, a couple of complications with this reasoning. To begin with, as we just saw, λ_S is not defined as the complete state of the pair to be measured, as with Bell, but as whatever is left after removing from λ_Σ the irrelevant sector, α, and β. We find this definition problematic for various reasons. First, by defining λ_S this way, one also removes all the force behind the reasonableness of the settings independence assumption. That is, if λ is the state of the pair to be measured, given a well-designed experiment, it is quite reasonable to assume that λ and the settings are independent. This, however, is no longer the case for λ_S. It is perfectly possible for aspects of the measuring apparatuses, other than the settings, to be relevant for the result of an experiment (we will have much more to say about this in section 4). If that is the case, the λ_S defined above would include such aspects. But then it is clear that one would not be able to freely assume such features of the apparatuses, now part of λ_S, to be independent of the settings—i.e., one would not be able to justify the usage of settings independence. For instance, suppose that the result of the experiment depends on the y-position of a particular particle (so it is part of λ_S) and that the setting of one of the measuring devices depends on the x-position of the same particle (so it is part of, say, α). If the particle is then constrained to be, say, in a circle, then λ_S and the settings would clearly be correlated.

Of course, as pointed out in Goldstein et al. (2011), there are going to be models for which λ_S will, in fact, be independent of the settings. We contend, however, that the important question to ask is not whether models like this exist, which they do, but whether, given a well-designed experiment, it is *reasonable* to assume that λ_S and the settings are independent. Our point is that this definition for λ_S terribly weakens the force behind the reasonableness of this assumption.

Another potential problem with the definition of λ_S is that it is experiment-dependent. That is, for the same system to be measured, the associated λ_S would change when subjected to different measurements. This seems odd, as one would expect for λ to capture aspects particular to the system to be measured, independently of aspects of the measurement apparatuses. Moreover, we find it problematic that λ_S, as defined in Goldstein et al. (2011), might not coincide with what a given model actually takes to be the state of the system to be measured.

There is, however, a bigger issue with the derivation of factorizability in Goldstein et al. (2011); an issue that prevails even after overlooking the problems with λ_S mentioned above. The final step in their argument is the claim that, since only α and λ are relevant for A, and since α is relevant to A only through a, then $P(A|a, b, B, \lambda) = P(A|a, \lambda)$. But why is it that α is relevant to A only through a? It seems to us that there could be aspects of α, beyond the information that a will be the case, that could be useful to make a better prediction for A. Suppose, for instance, that the position of some particle influences, but not determines, the value of a and determines the value for A. In that case, it is clear that the position of such a particle would be a part of α, and it is so in a way that is relevant for A beyond the fact that the setting is a. The point is that, in such cases, it is no longer true that one can substitute α for a and β for b, which would completely block the derivation of factorizability. That is, one would arrive at

$$P(A,B|\alpha,\beta,\lambda_S) = P(A|\alpha,\lambda_S)P(B|\beta,\lambda_S), \tag{19}$$

but would not be able to derive factorizability from such an equation.

It could be argued that all aspects of α, beyond the information that a will be the case, and which could be useful to make a better prediction for A, would already be encoded in λ_S. But this is not so because, by definition, λ_S is what remains *after* removing α and β. Therefore, by definition, it does not contain anything belonging to α. Moreover, even if one allows for those aspects of α to be contained in λ_S, that would only strengthen our case that this definition of λ_S would make it unreasonable for settings independence to be the case.

In sum, the discussion in Goldstein et al. (2011) unnecessarily mixes the derivation of factorizability with the independent assumption of settings independence and introduces the non-standard λ_S, for which the reasons for assuming settings independence are removed. Moreover, the derivation of factorizability depends on the assumption that α is relevant to A only through a (and similarly for β, b and B), but such an assumption is unwarranted—rendering the derivation invalid.

3.3 Quantum Non-Locality and Relativity

In order to justify factorizability, following Bell (1976), chapter 4 of *Quantum Non-Locality and Relativity* (Maudlin, 2011) calls N the overlap of the past light cones of regions 1 and 2, Γ the remainder of the past light cone of 1 and M the remainder of the past light cone of 2 (see Figure 2). Moreover, it calls γ, μ and ν the complete specifications of the entire physical states of Γ, M and N, respectively.[2]

Next, two principles are invoked, Reichenbach's common cause principle and the principle of local causality. The first one is stated as entailing that if one takes into account all events that could play a causal role in bringing about a pair of events, then the correlations between them disappear. As for the second, it is taken to imply that all of the causes of an event must lie in its past light cone. Applying these principles to the particular case of Bell-type experiments leads to

$$P(A,B|\gamma,\mu,\nu) = P(A|\gamma,\nu)P(B|\mu,\nu). \tag{20}$$

This, of course, is not factorizability. To arrive at such a condition, Maudlin (2011)

[2] We use Γ and γ, instead of Maudlin's Λ and λ, to reserve λ for the state of the pair, as defined in section 2.

argues that, when analyzing the Bell correlations, one usually stipulates that all of the causally relevant facts in region N are contained in the state of the pair, which it's called k. Moreover, one assumes that the only event in Γ which is relevant to the outcome is the setting a, and similarly for M. If so, $P(A|\gamma,\nu) = P(A|a,k)$ and $P(B|\mu,\nu) = P(B|b,k)$, so factorizability is obtained.

Are these assumptions legitimate? In footnote 10 of chapter 4 in Maudlin (2011), it is argued that k can, in fact, be stipulated to include all causally relevant physical facts in N. That might be so, the problem is that, as with λ_S above, with this definition of k, all reasons to expect k to be independent of the settings, i.e., all reasons to expect settings independence to be valid, disappear. Regarding the assumptions for the settings, things are even worse. In the same footnote, it is straightforwardly accepted that the settings simply cannot be stipulated to be the only causally relevant events in the remainders of the cones. As a possible remedy, the reader is then referred to Bell's treatment of the issue in Bell (1976), in which he deals with other factors by averaging over them. The problem is that, as mentioned above, already in the year such a paper was published, Shimony et al. (1976) pointed out that the derivation in Bell (1976) was, in fact, invalid.

3.4 The Stanford Encyclopedia

To conclude this assessment, we explore the article on Bell's Theorem in the The Stanford Encyclopedia of Philosophy (Myrvold et al., 2021). The discussion there starts by asserting that, in Bell (1976, 1990), factorizability is derived from the principle of local causality. Then, it goes on to distinguish between two versions of such a principle offered in Bell (1990): the informal, referred to as PLC-1, which reads

> The direct causes (and effects) of events are near by, and even the indirect causes (and effects) are no further away than permitted by the velocity of light,

and the more formal, referred to as PLC-2, which is equivalent to the one we presented in section 2. According to Myrvold et al. (2021), PLC-2 follows from the conjunction of PLC-1 and Reichenbach's common cause principle.

Next, it is claimed that factorizability is the application of Bell's principle of local causality to the particular setting of Bell-type experiments. Moreover, after recalling the well-known fact (Jarrett, 1984) that factorizability can be derived from the conjunction of *parameter independence* (PI)

$$P(A|a,b,\lambda) = P(A|a,\lambda) \quad \text{and} \quad P(B|a,b,\lambda) = P(A|b,\lambda) \tag{21}$$

and *outcome independence* (OI)

$$P(A, B|a, b, \lambda) = P(A|a, b, \lambda)P(B|a, b, \lambda), \tag{22}$$

the following blunt remark is offered

> PI is a consequence of the causal locality condition PLC-1 alone, whereas OI requires in addition the assumption of the common cause principle.

In sum, according to Myrvold et al. (2021), the principle of local causality follows from PLC-1 and the common cause principle. Moreover, PI follows from PLC-1 and OI from PLC-1 and the common cause principle—and so from the principle of local causality. Putting everything together, factorizability, which follows from PI and OI, follows from the principle of local causality alone.

Are these claims correct? It is hard to say because no details about any of the derivations are given. Here, we will focus on the claims regarding PI and OI and show them to be problematic. We are also doubtful regarding the claim that the principle of local causality follows from PLC-1 and the common cause principle, but we will not explore that in detail.

Let's start with the claim that PLC-1 is sufficient for PI. To begin with, we point out that, while PLC-1 does not refer to probabilities, PI does. Therefore, one needs to add to PLC-1 a way of talking about probability. Granting this as unproblematic, the remaining task is with identifying as causes the settings a, b, and as effects the results A, B. If one could do this, PLC-1 would imply that the setting b cannot affect the result A and *vice versa*. How, then, would one argue that the settings can be thought of as causes and the results as effects?

One possibility is by invoking an *agency theory of causation*, according to which "the ordinary notions of cause and effect have a direct and essential connection with our ability to intervene in the world as agents" (Menzies and Price, 1993, p. 187). Since the settings a and b can, presumably, be decided by the experimentalists, one can identify them as possible causes. If so, PLC-1 would imply that the setting a (b) cannot affect the event B (A). The Agency Theory of Causation, however, is not without its problems, (Woodward, 2003).

More serious complications besiege the derivation of OI from PLC-1 and the common cause principle—complications we have encountered before. To derive OI from the common cause principle, one would start by pointing out that, in a Bell-type experiment, there are correlations between the results A and B, i.e., $P(A, B|a, b) > P(A|a, b)P(B|a, b)$. Then, by invoking the principle, there must be a common cause, C, such that, conditional on C, events A and B are independent. In other words,

$$P(A,B|a,b,C) = P(A|a,b,C)P(B|a,b,C). \tag{23}$$

This surely looks like OI, but it isn't because C may be different from λ: one is the Reichenbachian common cause and the other the complete state of the pair of particles—and it is perfectly possible for those two things not to be the same. One could then simply define λ as C, but that is also no good because, in that case, as with previous attempts, the reasons we have for assuming λ to be independent of the settings disappear.

More generally, there are two issues with attempts (see also Cavalcanti and Wiseman (2021)) at deriving factorizability from the principle of local causality and the common cause principle. The first, as we just saw, is that he have no control over the scope and nature of the common cause C, so it cannot be assumed to be λ, nor to be independent of the settings. The second issue is that, since, for instance, a and C cannot be assumed to give a complete physical description over a slice of the past light cone of region 1, one cannot apply the principle of local causality to remove b from the first term in the right of the last equation, nor a from the second. We conclude that, to properly derive factorizability, some other assumption must be included. We turn to that issue in the next section.

4 From locality to factorizability revisited

In the previous section, we reviewed well-known attempts at deriving factorizability from the principle of local causality—and found them all wanting. In this section, we revisit the issue anew, looking for missing elements in past efforts. To do so, we start with the equality

$$P(A,B|a,b,\lambda) = P(A|B,a,b,\lambda)P(B|a,b,\lambda), \tag{24}$$

with the standard definition for λ, i.e., that of the complete state of the pair to be measured. As discussed above, for such a λ there are good reasons to assume, independently from factorizability and the principle of local causality, for it to be independent of the settings a and b—i.e., there are good reasons to assume settings independence.

Now, to obtain factorizability from (24) what we need is to remove b and B from the first term on the right, and a from the second. Can we use the principle of local causality by itself to do so? Clearly not. Since a and λ do not constitute a complete description of a slice of the past light cone of the measurement in region 1, even though b and B are outside of such a cone, they cannot be simply removed, as they might contain information, not contained in a and λ, useful for the prediction. Now, the fact that b and B *might* contain such information does not imply that they do so. Therefore, the key question at this point is, under what circumstances would they? In other words, assuming the principle of local causality, under what circumstances would it be *illegitimate* to remove b or B from the first term on the right-hand side of (24)?

For that to be the case, two conditions are required. First, there must be something, besides a and λ, which is relevant for the prediction. Second, that something must be (at least partially) encoded in b or B. Let's explore these two conditions with some care.

Regarding there being something besides a and λ relevant for the prediction, a straightforward possibility (already considered above), is for there being aspects of the measurement apparatus or its environment, other than the setting, determining or influencing the result. That is, there could be aspects of the full microscopic state of the apparatus and its environment in region 1, call it λ_a,[3] which clearly go beyond the information contained in λ and a, that could be relevant for the result A. One concrete example in which this happens, adapted from a discussion of this possibility in Spekkens (2005, p. 5), is of a classical system and a classical measurement device, which generates an outcome by rolling one of several differently weighted dice with the choice of the die being determined both by the state of the system (call it λ) and the setting of the apparatus (call it a). It is quite clear, in this case, that the full microscopic state of the apparatus, (call it λ_a), would greatly improve upon the predictions one could make, conditional on a and λ alone: while the latter would limit to probabilistic statements, based on the die chosen, the former, being the underlying model deterministic, would allow for non-probabilistic, determined predictions.

We take a short detour to point out an important consequence of the previous discussion. It is often argued that any deterministic theory necessarily violates PI (Myrvold, 2016; Butterfield, 2018). This is important because PI seems to be incompatible with Lorentz invariance. Therefore, if determinism does imply violations of PI, the only chance for building a Lorentz invariant theory would be through indeterminism. The connection between determinism and violations of PI is argued for as follows. Any empirically adequate theory must violate Bell's inequality, so it must violate factorizability. Moreover, since factorizability can be derived from OI and PI, all such theories must violate one of these conditions. Next, it is argued that all deterministic theories trivially satisfy OI because, once the state and the settings are given, the probability of obtaining any result is either 1 or 0. Therefore, since OI must be the case in all deterministic theories, and because of the need to violate factorizability, it is concluded that any deterministic theory must violate PI. The problem with this argument, as we just saw, is that it is not the case that determinism implies satisfaction of

[3] One could think of λ_a (λ_b) as the complete physical state on a spacial slice of the past light cone of the apparatus on region 1 (2), and its immediate environment, moments before the measurement takes place.

OI, so it does not imply violations of PI. That is, even for a deterministic theory, it is not necessarily the case that the probability for A, once λ and a are given, must be either 1 or 0: the result may also depend on other aspects of λ_a. If all this is correct, it opens up the possibility for a deterministic, Lorentz invariant theory.

Coming back to the main discussion, what about the second condition, namely, for λ_a to be, at least partially, encoded in b or B? Well, as with settings independence, one must detach the purely logical question of its possibility with the more physical question of its reasonableness. Regarding the former, it is clear that the measurement apparatuses share a (not that distant) common past. Therefore, in the same way that it is possible for λ to be correlated with the settings, it is perfectly possible for the microstate of the apparatus (and its environment) on one side to be correlated with the setting or result of the apparatus on the other. Moreover, it is clear that this condition is fully independent of settings independence: even if λ and the settings are statistically independent, the microstate of the apparatus on one side could be correlated with the setting or result of the apparatus on the other.

Summing up, removing from the right-hand side of Eq. (24) what is required to obtain factorizability is *illegitimate* if i) besides the state and the settings, there are aspects of the apparatuses that are relevant for the prediction and ii) the microstate of the apparatus (or its environment) on one side has information regarding the setting or result on the other. Therefore, in order to enable the transition from locality to factorizability, at least one of these conditions must be blocked. What about denying the first condition, i.e., stipulating that only the state and the settings can be relevant for the prediction? The problem with this route is that it is hard to justify. As mentioned before, there are models for which additional variables, besides a and λ, do influence A. Therefore, denying i) by fiat would amount to denying such models a priori. This seems antithetical to the whole project, manifested in Bell's work, of finding predictions that *any local theory* would make.

A better option is to deny the second condition. In particular, in order to enable the transition from locality to factorizability, it is sufficient to assume that the microstate of the apparatus and its and environment, on one side, are independent of the setting and result of the apparatus, on the other. That is, to demand

$$P(\lambda_a|b,B) = P(\lambda_a) \quad \text{and} \quad P(\lambda_b|a,A) = P(\lambda_b), \tag{25}$$

which we call (for lack of a better term) the *microstate independence* assumption.[4]

One may think that it would be simpler to enunciate the assumption as demanding

[4] In fact, condition (25) is a little too strong, as all one actually needs in not independence between b and B, and all of λ_a, but independence between b and B, and the sector of λ_a which is relevant for A, beyond a and λ (and analogously for the other side).

independence between λ_a and λ_b. That would be easier to state, and it would accomplish the same objective. The problem, though, is that independence between λ_a and λ_b is not a demand which is reasonable to make since, in general, the microstates of the measuring apparatuses will be correlated. For instance, if both are in the same lab, or even in the same city, we would expect correlations in their temperatures or in vibrations due to imperceptive earthquakes. Is the assumption, as stated above, more reasonable? We do think so, but we will explore the issue in due time.

With all this, we can enunciate the main result of this work: factorizability follows from the conjunction of the principle of local causality and the assumption of microstate independence. To see this, we start with Eq. (24) and note that the principle of local causality implies that anything outside of the past light cone of 1, which does not enhance the prediction for A given a and λ, can be removed from the conditional in the probability of A (and similarly for region 2). Then, we notice that, because of microstate independence, this is the case for b and B in the first term on the right and for a in the second, so such terms can be removed, leading to factorizability. Now, since Bell's theorem follows from the conjunction of settings independence and factorizability, we conclude that Bell's theorem follows from the principle of local causality, settings independence and microstate independence. That, of course, means that it is possible to construct a model, satisfying both the principle of local causality and settings independence, which, in virtue of violating microstate independence—and thus factorizability—is able to break Bell's inequality.

To see how this would work in a concrete example, we can employ the game considered in Maudlin (2011), in which two friends try to reproduce the quantum correlations displayed in Bell-type experiments. The rules of the game are as follows. The friends start in a room, where they can devise a strategy to coordinate their answers. Then, they leave by different doors and each is asked one of a list of possible questions, which they can answer with either "up" or "down". To enforce locality, it is assumed that, after leaving the room, the friends can no longer communicate. Of course, the point of the game is to show that, assuming settings independence—i.e., that there is no correlation between the strategy selected by the friends and the questions to be answered—the friends will not be able to reproduce the quantum correlations.

What we want to show now is that, if one violates microstate independence, then the friends would be able to succeed in the game. To model a violation of microstate independence in this scenario, suppose that there is a correlation between the question to be asked on one side and, say, the color of the shirt of the examiner on the other (suppose, for concreteness, that the correlation is perfect). That is, in direct opposition to microstate independence, the setting (question) on one side is correlated with a feature of the apparatus (the color of the shirt of the examiner) on the other. It is quite clear that, under such circumstances, the friends would have no problem devising a strategy to reproduce the quantum statistics. For instance, by noting that the correlation between the color of the shirt on one side and the question on the other effectively means that one of the friends knows the question being asked to the other, the strategy could be to employ pilot-wave theory to produce the answers. That is, the friends could assume a singlet as the initial state and randomly choose in the room an initial condition analogous to the initial position of the pilot-wave particles. Then, they would choose the answers based on the corresponding predictions of pilot-wave theory for the analogous situation. In this way, all quantum predictions for the Bell experiment would be reproduced by a fully local procedure.

So far we have discussed the mere possibility of violating microstate independence, what about the reasonableness of doing so? Well, it could be argued that reasons very similar to those offered in support of settings independence, which we do find completing (although see Ciepielewski et al. (2021)), would also be applicable here. It seems to us that this is indeed the case for the dependence between, say, λ_a and b. That is, it seems to us that the same type of precautions usually considered to avoid a dependence between the settings and the state would also prevent, to a reasonable degree, dependencies between λ_a and b. The independence between λ_a and B, on the other hand, seems more tricky. The issue is that we seem to able to design experiments in which the settings are chosen in such a way that it is quite reasonable to assume for them to be independent of both the state and the microstate of the apparatus on the other side. And that seems to be helpful in preventing correlations between λ_a and b. However, since we have no control over B, it is not that clear whether all these precautions would actually prevent B from being correlated with λ_a. That is, while the reasons for assuming independence between λ_a and b seem strong, it is not clear that the same could be said regarding independence between λ_a and B.

Before moving on, it is useful to explore in some detail a possible dependence between the microstate on one side and the result on the other. First of all, it seems clear that a dependence between λ_a and B and between λ_b and A, even in the absence of a dependence between λ_a and b and between λ_b and a, would be sufficient for the inequality to be broken—as that would block the derivation of factorizability (note that a correlation between, say, λ_a and B, in the absence of correlations between λ_b and A, λ_a and b, and λ_b and a, would not block the derivation). Moreover, a model of that sort would display quite peculiar characteristics. Of course, it would share with standard quantum mechanics the ability to break OI. However, whereas in standard quantum mechanics $P(A|B,a,b,\lambda) \neq P(A|B,a,\lambda)$ and $P(A|B,a,\lambda) = P(A|a,\lambda)$, these models would satisfy the opposite, i.e., $P(A|B,a,b,\lambda) = P(A|B,a,\lambda)$ and $P(A|B,a,\lambda) \neq P(A|a,\lambda)$. That is, a model of this sort would manage to break IO through a very different path than standard quantum mechanics. We leave the analysis of models that employ these mechanisms to break the inequality for future work.

5 Conclusions

In early versions of Bell's theorem, the notion of locality was imposed via the factorizability assumption. However, in later presentations, Bell introduced the more general and fundamental principle of local causality and argued that factorizability could be derived from such a principle. In this work, we have analyzed prominent attempts at such a derivation and have shown that, contrary to what is usually affirmed, the derivation requires the introduction of an additional, non-trivial assumption—which we call microstate independence. This new assumption is similar to, but strictly independent of the well-known settings independence assumption, and demands independence between the microstate of the apparatus on one side and the setting and result of the apparatus on the other.

Given the fact that, to derive factorizability from the principle of local causality, this additional assumption is required, we conclude that it is possible to construct a model that satisfies settings independence and the principle of local causality, but which, in virtue of breaching this new assumption—and thus factorizability—is able to break Bell's inequality. In fact, we display a toy model in which this is the case. We close by pointing out that, in view of clear experimental violations of Bell's inequality, in the same way that one can cling to locality by giving up settings independence, one could hold on to, both, locality and settings independence, by dropping the microstate independence assumption.

Regarding the reasonableness of doing so, we maintain that, for the independence between microstates on one side and settings on the other, arguments very similar to those usually provided in support of settings independence, which we do find (for the most part) completing, would also apply in this case. Such arguments might be considered less effective in defending the independence between microstates on one side and results on the other, as we have no control over results. Still, it seems to us that correlations between microstates on one side and results on the other would strike most as quite conspiratorial; we leave further discussion of this point for the future.

Acknowledgments

We thank Shelly Goldstein for valuable comments. We acknowledge support from CONACYT grant 140630 and PAPIIT grant IN405320.

References

Aspect, A., Dalibard, J., and Roger, G. (1982). Experimental test of Bell's inequalities using time-varying analyzers. *Physical Review Letters*, 49:1804–1807.

Aspect, A., Grangier, P., and Roger, G. (1981). Experimental tests of realistic local theories via Bell's theorem. *Physical Review Letters*, 47:460–6443.

Bell, J. S. (1976). The theory of local beables. *Epistemological Letters*, 9:11–24.

Bell, J. S. (1990). La nouvelle cuisine. In Sarlemijn, A. and Kroes, P., editors, *Between Science and Technology*. Elsevier Science Publishers.

Butterfield, J. (2018). Peaceful Coexistence: Examining Kent's Relativistic Solution to the Quantum Measurement Problem. In Ozawa, M. et al., editors, *Reality and Measurement in Algebraic Quantum Theory*. Springer.

Cavalcanti, E. G. and Wiseman, H. M. (2021). Implications of Local Friendliness Violation for Quantum Causality. *Entropy*, 23(8):925.

Ciepielewski, G. S., Okon, E., and Sudarsky, D. (2021). On Superdeterministic Rejections of Settings Independence. *The British Journal for the Philosophy of Science*.

Giustina, M. et al. (2015). Significant-loophole-free test of Bell's theorem with entangled photons. *Physical Review Letters*, 115(25):1–7.

Goldstein, S., Norsen, T., Tausk, D., Victor, D., and Zanghi, N. (2011). Bell's theorem. *Scholarpedia*, 6(10):8378.

Hensen, B. et al. (2015). Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature*, 526(7575):682–686.

Jarrett, J. P. (1984). On the Physical Significance of the Locality Conditions in the Bell Arguments. *Noûs*, 18(4):569.

Maudlin, T. (2011). *Quantum Non-Locality and Relativity: Metaphysical Intimations of Modern Physics*. Wiley-Blackwell, Malden, MA, 3rd ed edition.

Menzies, P. and Price, H. (1993). Causation as a Secondary Quality. *The British Journal for the Philosophy of Science*, 44(2):187–203.

Myrvold, W., Shimony, A., and Genovese, M. (2021). Bell's Theorem. *The Stanford Encyclopedia of Philosophy*.

Myrvold, W. C. (2016). Lessons of Bell's Theorem: Nonlocality, Yes; Action at a Distance, Not Necessarily. In Bell, M. and Gao, S., editors, *Quantum Nonlocality and Reality*, pages 238–260. Cambridge University Press, Cambridge.

Shalm, L. K. et al. (2015). Strong loophole-free test of local realism. *Physical Review Letters*, 115:250402.

Shimony, A., Horne, M. A., and Clauser, J. F. (1976). Comment on "The theory of local beables". *Epistemological Letters Hidden Variables and Quantum Uncertainty*, 13.

Spekkens, R. W. (2005). Contextuality for preparations, transformations, and unsharp measurements. *Physical Review A*, 71(5):052108.

Weihs, G., Jennewein, T., Simon, C., Weinfurter, H., and Zeilinger, A. (1998). Violation of Bell's inequality under strict Einstein locality conditions. *Physical Review Letters*, 81:5039.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation.* Oxford Studies in Philosophy of Science. Oxford University Press, Oxford.