PROCEEDINGS A

royalsocietypublishing.org/journal/rspa

Research

Cite this article: Adlam E, Hance JR, Hossenfelder S, Palmer TN. 2024 Taxonomy for physics beyond quantum mechanics. *Proc. R. Soc. A* **480**: 20230779.
https://doi.org/10.1098/rspa.2023.0779

Received: 24 October 2023
Accepted: 10 June 2024

Subject Category:
Physics

Subject Areas:
quantum physics

Keywords:
quantum, taxonomy, interpretations, models, theories

Author for correspondence:
Jonte R. Hance
e-mail: jonte.hance@newcastle.ac.uk

THE ROYAL SOCIETY
PUBLISHING

Taxonomy for physics beyond quantum mechanics

Emily Adlam[1], Jonte R. Hance[2,3], Sabine Hossenfelder[4] and Tim N. Palmer[5]

[1]The Rotman Institute of Philosophy, Western University, 1151 Richmond Street, London, Ontario N6A 5B7, Canada
[2]School of Computing, Newcastle University, 1 Science Square, Newcastle upon Tyne NE4 5TG, UK
[3]Quantum Engineering Technology Labs, Department of Electrical and Electronic Engineering, University of Bristol, Woodland Road, Bristol BS8 1US, UK
[4]Munich Center for Mathematical Philosophy, Ludwig-Maximilians-Universität München, Geschwister-Scholl-Platz 1, Munich 80539, Germany
[5]Department of Physics, University of Oxford, Oxford OX1 3PU, UK

JRH, 0000-0001-8587-7618

We propose terminology to classify interpretations of quantum mechanics and models that modify or complete quantum mechanics. Our focus is on models which have previously been referred to as superdeterministic (strong or weak), retrocausal (with or without signalling, dynamical or non-dynamical), future-input-dependent, atemporal and all-at-once, not always with the same meaning or context. Sometimes, these models are assumed to be deterministic, sometimes not, the word deterministic has been given different meanings, and different notions of causality have been used when classifying them. This has created much confusion in the literature, and we hope that the terms proposed here will help to clarify the nomenclature. The general model framework that we will propose may also be useful to classify other interpretations and modifications of quantum mechanics. This document grew out of the discussions at the 2022 Bonn Workshop on Superdeterminism and Retrocausality.

1. Introduction

Quantum mechanics, despite its experimental success, has remained unsatisfactory for a variety of reasons, notably due to its tension with locality, and due to the measurement problem [1,2]. Different authors have

formulated their unease with quantum mechanics in different ways. Analysing the origin of this unease is not the purpose of this present article. The purpose is instead to sort out the confusion in the terminology used to describe this unease.

We will below introduce a framework to distinguish between interpretations of quantum mechanics and modifications thereof. Our hope is that it may offer researchers a guide to classify their own approach. Our aim here is not to judge the promise or correctness of any of these approaches, but to make it easier to communicate with each other about what we are working on in the first place. An accompanying paper [3] will discuss some applications of this terminology to showcase its use.

This paper is organized as follows. In the next section, we will outline the general model framework and most importantly introduce different types of models. We believe that this classification of what we even mean by a model in and by itself will serve to alleviate the confusion of nomenclature. We will then, in §3 and §4 introduce some properties that such models typically have. In §5, we will briefly explain how to use this classification scheme, and then we conclude in §6. A list of acronyms can be found in the appendix.

2. General model framework

We will start with explaining the general framework that we want to use in the following, so that we can distinguish between different types of models and their properties.

(a) Calculational models

At its most rudimentary level, the task of physics is to provide an useful method for calculating predictions (or postdictions) for observables in certain scenarios. (Other potential tasks associated with physics (e.g. explaining phenomena, telling us what exist, examining what is built up from what, how it changes over time and why, or producing knowledge of physical reality) are both more debatable than this, and require such an useful method themselves as a basis, so we stick with this rudimentary description as a minimal example of what the task of physics is). A scenario is most often an experiment, and this is the situation we are typically concerned with in quantum foundations. But in some areas of physics—such as cosmology and astrophysics—one does not actually conduct experiments, but instead one observes a natural variety of instances that occurred in the past. For this reason, we will use the more neutral term 'scenario'. A scenario is not itself a mathematical structure; it is the real-world system that we want to describe using mathematical structures.

The basic logical flow of such a calculation is outlined in figure 1. We will call the centre piece of this calculation a *calculational model*, or *c-model* for short. We do not call it a 'computational model' because 'computational' suggests a numerical simulation, which is an impression we want to avoid. A calculational model could be numerical, but it could also be purely analytic. For more on the relation between calculational models and computer models, see §2f.

We will, in the following, use a notation in which curly brackets $\{\cdot\}$ denote sets of mathematical assumptions. The set $\overline{\{\cdot\}}$ is the set of all assumptions that can be derived from $\{\cdot\}$ (including the elements of $\{\cdot\}$ themselves).

The different components of the modelling framework have the following properties:

Inputs (of a calculational model): The inputs $\mathcal{I}$ of a calculational model are a (at most countably infinite) set of mathematical assumptions—each of which is an input, I_i—that describes the scenario. To be part of the inputs, an assumption must differ between at least two scenarios. We will denote this set as $\mathcal{I} := \{I_i \mid i \in K \subseteq \mathbb{N}^+\}$, were K is the (at most countably infinite) index range of the assumptions.

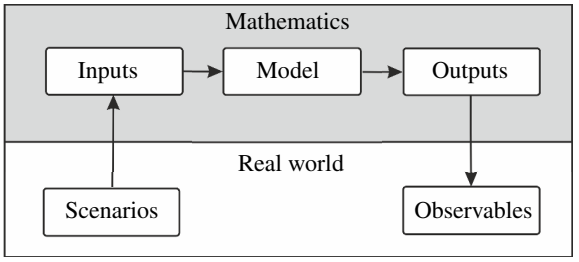

Figure 1. The logical flow of the model framework.

To aid readability, and because the index set will not matter in the following, we will from here on assume that all sets are at most countably infinite and suppress the index range. That is, we will just write $\{I_i\}$ rather than $\{I_i \mid i \in K \subseteq \mathbb{N}^+\}$.

Loosely speaking, you can think of the inputs as the mathematical representation of a scenario. A typical example might be the temperature in the laboratory, or the frequency of a laser. However, the inputs of a model do not necessarily have to correspond to definite observable properties of the scenario. They could also be expressing ignorance about certain properties of the scenario and thus be random variables with probability distributions, or they might indeed be entirely unobservable. We will come back to this point in §3a.

The inputs of a c-model are often assignments of values to variables. For example, a c-model may work with an unspecified variable x that is an element of some space. The input may then assign the value $x = 3$ to this variable for a specific scenario. For such an input, we will refer to the value it assigns as the *input value*. We want to stress that the input value is not the same as the input. The input is '$x = 3$', the input value is '3'.

A typical example of a c-model that we are all familiar with would be the Harmonic oscillator. In this case, the model would be the differential equation $m\ddot{x} = -kx$ and a complete set of inputs would be value assignments for k and m, plus two initial values for, say, $x(t = 0)$ and $\dot{x}(t = 0)$.

However, inputs of a calculational model are not necessarily value assignments. They might also be constraint-equations, or boundary conditions, or something else entirely. For example, when studying stellar equilibrium, it is quite common to enter an equation of state as the input to the Tolmann–Oppenheimer–Volkov (TOV) equations. In this case, the TOV equations are the same for all stellar objects that one may want to consider; they are hence part of the model. The equation of state, on the other hand, changes from one type of star to another; it is hence part of the input.

While our main interest is in models that describe the real world, it is also possible to study a model's properties with scenarios that do not exist in reality. We will refer to those as *hypothetical scenarios*. They include, but are not limited to, counterfactual realities, as well as universes with different constants of nature. (Note Frigg & Nguyen [4], among others, have also discussed representation in scientific modelling.)

Calculational model: A set $\mathcal{C}$ of mathematical assumptions A_x that are independent of the scenario. We will denote this set as $\mathcal{C} := \{A_k\}$.
Set-up: The union of a calculational model and its inputs. $\mathcal{F} := \mathcal{I} \cup \mathcal{C}$.
Outputs (of a calculational model): All mathematical statements that can be deduced from set-up, but from neither the model nor its inputs in isolation $\mathcal{O} = \overline{\mathcal{F}} \setminus (\overline{\mathcal{I}} \cup \overline{\mathcal{C}})$.

Predictions from observables are obtained from the outputs of the model. However, not all outputs of a model need to be observable. The prime example is quantum mechanics, where the outputs contain the time-evolution of the wavefunction, but the wavefunction itself is not observable. But there are many other examples, such as the production of gravitational waves

by a black hole merger. Given suitable inputs, the model (General Relativity) will output a mathematical description for the creation and propagation of gravitational waves, but we only measure their arrival on Earth, and only measure that through the waves' influence on matter, which is a small part of all the model outputs. And like the inputs, the outputs of a model do not necessarily have to result in definite values for observables; they could instead merely give rise to probability distributions for values of observables.

For a calculational model to be useful, we further need a prescription to encode a scenario with specific inputs, and a way to identify the observables from the outputs. This identification, since it is not restricted to the mathematical world, is not one that we can strictly speaking define. Science, after all, is not just mathematics. What property this identification with the real world must have is an interesting question, but it is one of many that we will not address here, because it is not relevant for what follows.

When using hypothetical scenarios, these have to be distinguished from the set-up as a matter of definition. If the hypothetical scenarios are not identified as such by definition, it becomes impossible to tell whether one changes the hypothetical scenario or the model.[1]

Another property, relevant for our purposes, that we want the set-up of a calculational model to have is: that they are *irreducible*, in the sense that we cannot split the combined set $\mathcal{C}:=\{I_j, A_k\}$ into two sets $\mathcal{C}_1:=\left\{I_j^1, A_k^1\right\}$ and $\mathcal{C}_2:=\left\{I_j^2, A_k^2\right\}$, where $\{I_j\}=\left\{I_j^1\right\}\cup\left\{I_j^2\right\}$ and $\{A_k\}=\left\{A_k^1\right\}\cup\left\{A_k^2\right\}$, so that both $\mathcal{C}_1$ and $\mathcal{C}_2$ are each set-ups of calculational models and the combination of their outputs is the same as the outputs of $\mathcal{C}$. A simpler way to put this is that, if we split the set-up of an irreducible model into two, some of the output can no longer be calculated. This approach is reminiscent of identifying particles in quantum field theory from the irreducible representations of the Poincaré group.

We need this requirement because otherwise, we could just join different set-ups to form a new one, which would make it impossible to classify any. Note that this does not mean the composition of two set-ups is not a set-up. On the contrary, a composition of two set-ups *is* a set-up, it is just that the combined set-up is no longer irreducible. The issue is, if we allowed reducible combinations of set-ups, then we could not meaningfully assign properties to them. It would be like asking which fruit is fruit salad. Once we have, however, succeeded in identifying properties of irreducible set-ups, we can join those with the same properties together, and meaningfully assign the same property to the reducible set-up. Using the above fruit example, if we have identified several fruits as apples, we can join them and be confident we have apple salad.

A particularly relevant case of a reducible set-up is one in which some inputs or assumptions can be removed without changing anything about the outputs. This may be because the assumptions are not independent (in the sense that some can be derived from the others), or because an assumption is simply not used to calculate the outputs.

One might be tempted to add to the requirements of a model that its assumptions are consistent (given problems such as the Principle of Explosion with inconsistent models [5]). However, as is well known, for recursively enumerable sets, Gödel's theorem [6] tells us that we cannot in general prove the consistency of the assumptions. We might then try to settle on the somewhat weaker requirement that at least the assumptions should not be known to be inconsistent. However, it sometimes happens in physics that a model works well in certain parameter ranges despite being inconsistent in general. An example may be the Standard

[1]An example for this is the case of the Standard Model with its fundamental constants not fixed but taken as variable inputs. This cannot be a correct model for scenarios in our universe because in our universe the constants are constant, hence they cannot be inputs. One can, however, consider hypothetical scenarios with different values of the constants, often interpreted as a type of multiverse. But, of course, one could alternatively interpret these hypothetical scenarios as different versions of the Standard Model that, alas, happen to not agree with our observations. The point is that whether the Standard Model with fundamental constants that do not agree with our observations describes a hypothetical alternative universe, or is just a wrong model for our universe, is a matter of definition.

Model without the Higgs field and its boson [7]. For this reason, we will here not add any requirement about consistency. One may justify this by taking a purely instrumental approach. We only care whether a set of assumptions is any good at describing observations.

The set-up of a calculational model whose inputs are all value assignments that, alas, have not been assigned is what is usually called a model in the causal models literature [8,9].

(b) Mathematical models

We defined a calculational model and its inputs as sets of mathematical assumptions. Such sets can be expressed in many different ways that are mathematically equivalent. We will call an equivalence class of calculational models a *mathematical model, or m-model* for short, that is the term m-class means the same as m-model. We will denote this equivalence class with $[\,\cdot\,]_{\mathrm{m}}$ and refer to it as an *m-class*; it is a set of sets. For example, if $\mathcal{C} = \{A_k\}$ is a calculational model, then $\mathcal{M} := [\mathcal{C}]_{\mathrm{m}}$ is the m-model that encompasses all mathematically equivalent formulations

$$[\mathcal{C}]_{\mathrm{m}} := \{\{B_j\} : \{B_j\} \Leftrightarrow \{A_i\}\}\,. \tag{2.1}$$

Calculational models in the same m-class will be called *m-equivalent*. By a mathematical equivalence of the sets ($\Leftrightarrow$), we here mean that either one can be derived from the other. We thereby adopt the notion of model equivalence advocated by Glymour [10,11].[2]

It was argued by Weatherall [12,13] that mathematical equivalence might overdistinguish models in certain circumstances, and that it might be better to use a weaker form of structural equivalence based on category theory. We do not use this proposal of categorial equivalence here, not because we object to it, but because it is neither widely accepted nor understood. Most importantly, it is at present not practical because few physicists would know how to apply it. Mathematical equivalence, in contrast, is widely understood and comparably straightforward (though not necessarily easy) to prove.

Among the models we will discuss here, mathematical models are closest to what physicists typically mean by a 'model'. They do not distinguish between the particular mathematical formulations that one might use to make a calculation. For example, we could express Maxwell's Equations using differential forms with $\mathrm{d}\star$, $\mathrm{d}\wedge$ and $\star\mathrm{d}$ operations, or we could do it with the more old-fashioned $\vec{\nabla}$, $\vec{\nabla}\cdot$ and $\vec{\nabla}\times$ operators. These two definitions would be two different calculational models, but the same mathematical model. A similar equivalence covers switching from the Schrödinger picture to the Heisenberg picture in quantum mechanics. We believe that most physicists would not call these different models.

The inputs of a mathematical model are likewise an equivalence class: it is the class of all sets of assumptions which change with the scenario that, together with the mathematical model, produce equivalent outputs.

The inputs of a mathematical model are usually not just the set of assumptions that are mathematically equivalent to the inputs of one of the calculational models. This is because the assumptions of the model itself may render certain inputs equivalent that are not equivalent without the model. An example that will be relevant for what follows is that a model with a time-reversible evolution operator will produce identical outputs for input states at times t_1 and $t_2 > t_1$, if the input state at t_2 is the forward evolution of the state from t_1. In this case, the input is m-equivalent, though the inputs at different times are not equivalent to each other without the model.

We will refer to different calculational models as *representations* of their m-class.

[2]Glymour uses the term 'theory equivalence', not 'model equivalence'. This is not a relevant distinction for the purposes of this present paper: please see §2e for discussion.

Representation (of an m-model): A calculational model within the m-class of that m-model.

In Argaman [14], a similar concept, that of two c-models within the same m-class, was referred to as 'reformulations' of each other.

(c) Physical models

The previous definitions do the heaviest lifting for the model framework, and the remaining ones are now straightforward. It can happen that mathematical models are different, but they nevertheless give rise to the same observables for all scenarios. We will combine all such mathematical models into yet another, even bigger class and say they constitute the same *physical model, or p-model* for short. We will denote this equivalence class with $[\,\cdot\,]_{\mathrm{p}}$ and refer to it as the *p-class*. Note that we could take either the *p*-equivalence class of a computational model or that of a mathematical model. Models in the same *p-class* will be called *p-equivalent*.

By calling these models physical, we do not mean to imply that we only consider observable properties as physically real. The reader should not take our nomenclature to imply any statement about realism or empiricism. We call them physical just because they describe what physicists in practice can distinguish with physical measurements.

This now gives us a way to define what we mean by an interpretation:

Interpretation (of a p-model): Any one of the mathematical models in the p-class.

Note that according to this nomenclature,[3] each interpretation has its own representations. That is, a c-model is a representation of an m-class, but more specifically it is a representation of a particular interpretation.

For example, if we take the Copenhagen Interpretation (CI) in one of its common axiomatic formulations (e.g. as given in Zurek [16]), that set of axioms will constitute a particular representation, hence, a c-model. The CI *per se* is the class of all mathematically equivalent models. We can then further construct the physical equivalence class of the CI, which we will in the following refer to as Standard Quantum Mechanics (SQM:=$[\mathrm{CI}]_{\mathrm{P}}$). Any other mathematical model that is physically but not mathematically equivalent to the CI is also an interpretation of SQM.

For the purposes of this article, we will not need to specify exactly what we mean by CI. However, we will take it to be only a first-quantized theory. That is, it is not a quantum field theory. If one were to specify further details, one would in particular have to decide whether one considers the relativistic version, or only the non-relativistic case, as some alternative interpretations work only non-relativistically [17].

That we have interpretations and not just representations is a possibility which appears in quantum mechanics because it has outputs that are unobservable in principle. A different computational model which affects only unobservable outputs might not be mathematically equivalent and yet give rise to the same physics.

(d) Empirical models

Finally, we define an *empirical model, or e-model*, $\mathcal{E}$, as the class of all physical models that cannot be distinguished by observations so far. We will denote this class as $[\,\cdot\,]_{\mathrm{e}}$, refer to it as an *empirical class or e-class*. Models in the same e-class are *empirically equivalent or e-equivalent*. We will call anything in the same empirical class as SQM, but not in the same physical class, a modification of quantum mechanics:

[3]Note others have given different definitions of interpretations for quantum mechanics, e.g. [15].

Modification or completion or extension (of a p-model): Another p-model that is not physically equivalent but so far empirically equivalent.

Figure 2 summarizes the relations between these various types of models.

We here use the terms modification, completion and extension loosely, and will not distinguish them below because it is not necessary for what follows. In the literature, these terms are used with a somewhat different meaning. A modification or extension of a model typically employs similar mathematical structures, whereas a completion introduces new mathematical structures. However, the distinction between the two is not always clear. Again, just how to distinguish them is an interesting question in its own right, but not one we will address here.

With a slight abuse of terminology that is, however, unlikely to cause misunderstandings, we can also use the above definition to refer to an m-model or a c-model as a modification. That is, an m-model (c-model) is a modification of another m-model (c-model) if the two are not physically equivalent, but so far empirically equivalent.

One could further refine the class of empirically equivalent models by specifying various sets of experiments. For example, we could ask what models can be distinguished by experiments in the near future, and what exactly we mean by near future. Or, we could talk about models that can be distinguished by experiment in principle, and then discuss whether, say, waiting 100 billion years is possible in principle. Or, whether it is possible in principle to make measurements inside a black hole, and so on. All of these are very interesting questions; however, they will not concern us in the following, so we will not elaborate on them.

(e) Theories

We will in the following not distinguish between models and theories. Loosely speaking, a theory is a class of models that can be used for a large number of scenarios, congruent with the semantic approach of Suppes [18] and van Fraassen [19]. However, physicists do not tend to use the terms theory and model in a well-defined way.

For example, we use the terminology of Quantum Field Theory in general, and the Standard Model in particular. We refer to General Relativity as a theory, and to ΛCDM as a model. This agrees with what we said above. But we also refer to Fermi's theory of beta decay, and the BCS theory, though those would better be called models. To make matters worse, we also sometimes call supersymmetry a theory, when it is really a property of a class of m-models, and so on. For the purposes of this article, we will not need to distinguish theories from models, so we will not bother with a precise definition of the term 'theory'.

(f) Computer models and simulations

Another type of model which is commonly used by scientists of all disciplines are computations performed on computers. There are two ways to think of these models.

One is that they are really simulations that represent one real-world system with another real-world system. That is to say, they are not models in the sense that we have considered here—the models we considered here are mathematical. A suitably programmed computer is instead a physical stand-in for the system one wants to make a prediction for. This is also how quantum simulations work [20,21].

However, there is another way to look at computational models, which is to use their program as a definition for a calculational model. This then falls into the classification scheme we discussed above. But, in this case, the computational model is typically not mathematically equivalent to the analytical, calculational model that one used for a scenario. This is because computers are in almost all cases digital, and only approximate the continuum. The exception

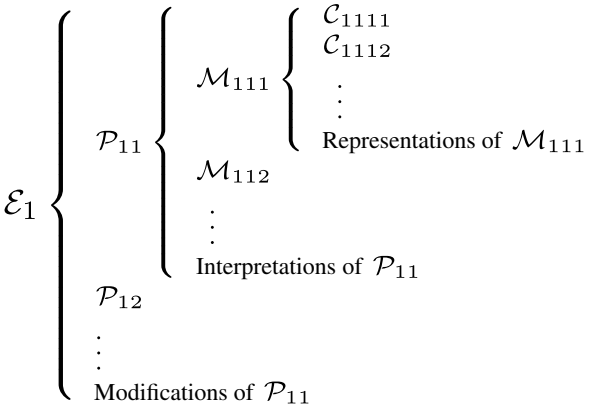

Figure 2. A mathematical model $\mathcal{M}_{ijk}$ is an equivalence class of computational models $\mathcal{C}_{ijkl}$. A physical model $\mathcal{P}_{ij}$ is an equivalence class of mathematical models $\mathcal{M}_{ijk}$. An empirical model $\mathcal{E}_i$ is an equivalence class of physical models $\mathcal{P}_{ij}$.

are certain types of analogue computers, which however can better be understood as simulations again.

That is to say, when one wants to classify a computer model according to our scheme, one should take its algorithmic definition as a set of assumptions, and use that to define a calculational model and its inputs. This calculational model, defined by its executable algorithm, will be different from the calculational model that uses an analytic expression.

(One can then further ask to what accuracy will the algorithm, when executed on a physical computer, approximate the output of the analytical model. This is a relevant point which was previously brought up in [22]. While an analytically defined calculational model might be time-reversible, an algorithmic approximation of it run on a computer will in general no longer be time-reversible. This is because errors will build up differently depending on which direction of time one runs the algorithm. This is particularly obvious for time-evolutions which are chaotic. In such cases, the forward-evolution and the backward-evolution of the algorithm as executed on the computer will actually be two different calculational models, and both are different from the analytical calculational model that they approximate).

3. General model properties

In this section, we will discuss some properties that we can assign to models and their inputs. The point of this section is to specify which properties are m-class properties (do not change with mathematical redefinitions), and which are c-class properties (can change with mathematical redefinition).

We want to stress that we will not prove that these assignments are correct. To do that, we would have to add many more assumptions (e.g. about what we mean by space, time, measurements and so on). What we will do instead is specify what we believe is the way that an expression has been previously dominantly used in the literature, and this specification will then in turn imply properties for the concepts we did not specify.

To make this procedure clear, if we state for example that the property of 'time-reversibility' does not change with a mathematical redefinition, then this implicitly means: if it did change, we would not refer to the property as time-reversibility. That is, there are certain properties of models which we want to not depend on just exactly how we write down the mathematics.

Quite possibly, some readers will disagree with some of our assessments. This is fine. Our aim here is not to end the discussion, but to propose a language in which we can talk about these properties in the first place.

The term 'model' without further specification (c/m/p/e) will refer to any type of model.

(a) Atemporal properties

Following the terminology of Cavalcanti & Wiseman [23], we will call an m-model 'deterministic' if its observables can be calculated with certainty from the outputs obtained from the set-up:

Deterministic: An m-model is deterministic if[4] the probabilities for predictions of observables are all either 0 or 1.

This property was termed 'holistic determinism' in Adlam [24] and differs from a more common definition of determinism, that connects one moment in time with a later one. We will elaborate on this in the next subsection. For now, we further follow [23] and distinguish determinism from predictability:

Predictable: An m-model is predictable if it is deterministic, and the inputs are all derived from observable properties of the scenario.

Both determinism and predictability are m-model properties, because we expect a redefinition of the mathematics that one uses to process inputs to not change predictions for observable output. If that was the case, we would assume something is wrong with our idea of what is observable.

The distinction between deterministic and predictable is that a model may have inputs that are unknowable in principle. Typically, these are value assignments for variables that are usually referred to as 'hidden variables'. A model with such hidden variables, according to the above terminology, may be deterministic and yet unpredictable.

Hidden variables: Hidden variables, that we will denote κ, are input values to a c-model that cannot be inferred from any observation on the scenario.

We want to stress that these hidden variables are in general not localized and sometimes not even localizable in any meaningful way. We will say more about localizable variables in the next subsection, but a simple example is that we could use Fourier-transforms of space–time variables, or just extensive quantities that are properties of volumes. Note that hidden variables are defined for c-models, not for m-models. This is because hidden variables can be redefined into (not necessarily local) random variables with an equivalent mathematical outcome. That is to say, mathematically, it makes no difference whether a variable was unknowable or indeed random.

While it may sound confusing at first to distinguish determinism from predictability, it will be useful in what follows. Indeed, the reader may have recognized that Bohmian Mechanics is deterministic yet not predictable. In Bohmian Mechanics, observables can be calculated with certainty if the inputs are specified, yet the inputs are also assumed to be partly unobservable in principle, so predictions can still not be made with certainty. Since determinism is a property of an m-model that cannot be removed with a redefinition, it follows that Bohmian Mechanics is not a representation of SQM. We elaborate on this further in the companion paper [3].

(b) Local and temporal properties

We will now add some properties of models that we commonly use in physics, starting with those that refer to locations in space and time. Clearly, we can only speak of locations in space and time if a model has some notion of space and time, and a distance measure on them, to begin with. In the simplest case, that would be the usual space–time manifold and its metric

[4]We use the common mathematical abbreviation 'iff' for 'if and only if.

distance. But it could alternatively be some other structure that performs a similar function, such as a lattice, or a discrete network with a suitably defined metric.

To make sense of space and time, we will in the following consider only a restricted m-class associated with a calculational model, that is one in which space and time are consistently defined throughout the entire class. To see why this is necessary, imagine if we were to redefine one direction of 'space' as 'time' and vice versa. This is mathematically possible, but it makes no physical sense. It would screw up any notion of locality and causality just by nomenclature. We will hence assume that space and time and a distance measure on them are consistently defined throughout the m-class.

However, sometimes a model just does not have a description of space–time or a notion of locality. This might sound odd, but is not all that uncommon in the foundations of quantum mechanics, where many examples deal with qubit states that do not propagate and are not assigned space–time locations [25]. We will refer to such models as alocal:

Alocal: An m-model that does not have a well-defined notion of space, space–time, locality or propagation.

One cannot fix a lack of definition by switching to a different but mathematically equivalent definition, so if a c-model is alocal, then so will be any other c-model its m-class.

For alocal models, we cannot make any statements about whether they are local or not. Similarly, a model may just not have a notion of time or a time-evolution. This again is not all that uncommon in the foundations of quantum mechanics. Many elaborations on correlations and classicality bounds, for example, do not specify a time-evolution for states; and the process matrix formalism is specifically designed to study possible quantum processes without a predefined time order [26,27].

Atemporal: An m-model that does not have a well-defined notion of time.

For models which have a proper notion of space–time, and a distance measure on it, we then want to identify how local they are. For this, we first have to identify the input values that can be assigned to space–times which Bell coined 'local beables' [28]. According to Bell, local beables 'can be assigned to some bounded space–time region'. We will take 'assigned to' to mean that the variable is the value of a function whose domain is space–time, or whatever the stand-in for space–time is in the model at hand. Note 'region' might be a point set.

For example, if space–time is parameterized by coordinates $\left(\vec{x}, t\right)$, and we have a function $f:\left(\vec{x}, t\right) \to \mathbb{C}$, then $f\left(\vec{x}, t\right)$ is a local beable. The domain of the SQM wavefunction is generically configuration space and it is therefore not a local beable, though under certain circumstances local beables can be obtained from it (such as the single-particle wavefunction in position space). Local beables are not necessarily observable.

Bell's definition of a local beable makes the assignment to a space–time region optional (can be assigned). However, if the assignment was optional, then it could be omitted, and a model with an assumption that can be omitted is reducible. Since we only deal with irreducible models, we therefore already know that if a variable is assigned to a space–time region, then this assignment is actually necessary. For this reason, we define

Local beable: An input value of a c-model that is assigned to a compact region of space–time.

In the following, $\mathcal{I}(A)$ will refer to the local beables in space–time region A.

It becomes clear here why we required our models to be irreducible: so that it is actually necessary, and not just possible, to assign the local beables to space–time regions. It has for example been rediscovered a few times independently [29,30] that any theory can be made local (in pretty much any reasonable definition of the term) by copying the local beables from

any space–time location into any other space–time location and postulating them to 'be there' without any other consequences.

Concretely, suppose we have localized variables $f(x,t)$ over a space–time manifold M, then we can just define the state $f(x,t) \otimes f(x',t') \in M \otimes M$ and call that a local beable at (x,t), just that now we have the information about any (x',t') also available at the same point. Depending on perspective, one might consider such a model as either ultra-local or ultra-non-local. It is ultra-local in that we do not need to know the state at any other location than (x,t) to calculate the time-evolution. It is ultra-non-local because any point in space–time contains a copy of the entire universe.

A similar problem would occur with parameters, such as $\hbar$, that could be transformed into fields $\hbar : (x,t) \to \delta(x-x')\delta(t-t')$ and then be used as inputs with random locations, leading to seemingly 'non-local' interactions in the Hamiltonian.

Such procedures to localize beables create new c-models that are in different m-classes, because the copying procedure is an additional assumption that could not have been derived from the original version of the c-model. Since such a localizing assumption is unnecessary to calculate outputs, the set-up of such a model is reducible and, as we have previously remarked, properties of reducible set-ups are ambiguous.

A word of caution is in order here. As mentioned above in the elaboration on alocality, many discussions of violations of Bell's inequalities do not explicitly state locality assumptions. These assumptions might hence appear unnecessary. This is indeed correct if one merely looks at the inequality. However, if one wants to draw conclusions about locality from a violation of Bell's inequality, one needs to at least make a statement about which parts of the experiment are causally connected or space-like separated. That is, we suspect that many alocal models do implicitly require locality statements to arrive at the desired conclusions. One should not be deceived by only looking at the explicitly stated assumptions.

Having introduced local beables, we can now define an alternative:

All-at-once (AAO) input: Input of a c-model that constrains or defines properties of at least one local beable for at least two temporal regions, and that cannot be decomposed into inputs in separate temporal regions.

Typical examples of such inputs are event relations, consistency requirements for histories, evolution laws, temporal boundary conditions or superselection rules. Something as mundane as an integral over time that depends on the scenario would also be an all-at-once input.

All-at-once (AAO) model: A c-model that uses all-at-once input.

The use of all-at-once input is *a priori* a property of c-models. That is to say, it might be possible to reformulate a model with AAO input into a mathematically equivalent model that does not have this property.

The principle of least action in classical mechanics, for example, uses All-At-Once input (the action), but given that the Lagrangian fulfils suitable conditions, the principle can equivalently be expressed by the Euler–Lagrange equations which do not require AAO input. Another example may be the cosmological model introduced by Carroll & Remmen [31], which uses a constraint on the space–time average of the Lagrangian density. The authors refer to this as non-local, and in some sense it is, but it is more importantly also an all-at-once input.

Now that we have localized variables, we can define a notion of locality. We will first leave aside causality and introduce a notion of Continuity of Action (CoA), as done in Wharton & Argaman [22]. The idea is that if one wants to calculate the outputs $\mathcal{O}(A)$ for a region A, then it is sufficient to have all the information on a closed hypersurface, S_1, surrounding the region, and adding local beables from another region outside S_1 provides no further information.

CoA (locality): An m-model has continuous action or fulfils CoA or is local iff it fulfils the condition $P(\mathcal{O}(A) \mid \mathcal{I}(S_1), \mathcal{I}(B)) = P(\mathcal{O}(A) \mid \mathcal{I}(S_1))$ for any region S_1 that encompasses A but not B (see figure 3*a*).

Non-locality: An m-model is non-local if it violates CoA.

Note that fulfilling CoA does not mean that the outputs in A are determined by the input values on S_1 to begin with. It is just that adding information from B does not make a difference.

A natural question to ask here is just how small the regions should be for the model to have continuous action. This is a most excellent question but luckily one that we do not need to answer because it resides on the level of empirical adequacy. A model in which variables cannot be localized sufficiently much will simply not reproduce certain observations. (For example, a model with a spatial resolution so rough that it cannot distinguish two detectors from each other can only give a total probability for both, not for each separately.) The typical size of the regions is what is often referred to as the coarse-graining scale of a model.

A subtlety of the definition for CoA was pointed out in Palmer [32], which is that some combinations of inputs may be mathematically possible, but are not in the physical state space of the model and hence, do not correspond to any scenario which can occur in reality. This might happen, for example, because certain combinations of variables are just forbidden by a mathematical constraint (as happens with the Fermi exclusion principle). In this case, one might have a situation where, e.g. $P(\mathcal{I}(S_1), \mathcal{I}(B)) = 1$ and $P(\mathcal{I}(S_1), \mathcal{I}(B)') = 0$ just because the latter combination is incompatible with the assumptions of the model [33]. It would then seem that $P(A \mid \mathcal{I}(S_1), \mathcal{I}(B)) \neq P(A \mid \mathcal{I}(S_1), \mathcal{I}(B)')$ and CoA is violated. However, as argued in Palmer [32], it is rather meaningless to talk about violations of locality for configurations which do not physically exist. Hence, if a model has input constraints, one should only apply the locality requirement to inputs that lead to physically possible scenarios.

We define CoA as an m-model property because, if it could be removed with a mathematical redefinition, we believe most physicists would not accept the definition as meaningful.

The term 'non-locality' has been used to refer to many other definitions in the foundations of physics in general, and quantum mechanics in particular. For example, in field theories, non-locality often refers to dynamical terms of higher order, a definition that is also used in General Relativity [34]. In quantum field theories, non-locality usually refers to the failure of operators to commute outside the light cone. Even in quantum foundations, non-locality may refer to different properties. For example, entanglement itself is sometimes considered non-local despite being locally created [35]. The latter in particular has created a lot of confusion, because while it has been experimentally shown that entanglement is a non-local correlation and does in fact exist, this does not imply that nature is non-local in the sense that the term has been used in Bell's theorem, which has of course not been shown [36].

Surveying all those different notions of non-locality is beyond the scope of this present work. However, we want to stress that as definitions, none of these notions of non-locality are wrong. They are just that: definitions. We chose our definition to resemble closely the 'spooky action at a distance' that Einstein worried about.

According to our definition, a calculational model fulfils CoA even if it cannot be directly evaluated whether it fulfils the requirement on the conditional probabilities, so long as a mathematically equivalent reformulation of the model fulfils it. The most relevant example for our purposes is that $[\mathrm{CI}]_\mathrm{m}$ does not fulfil CoA. This is because making a measurement in B provides information about the measurement outcome in A that was not available in S_1. This is Einstein's 'spooky action at a distance'.

That is, with the terminology we have introduced so far, the CI and all mathematically equivalent formulations are equally non-local. The question is then merely whether this is

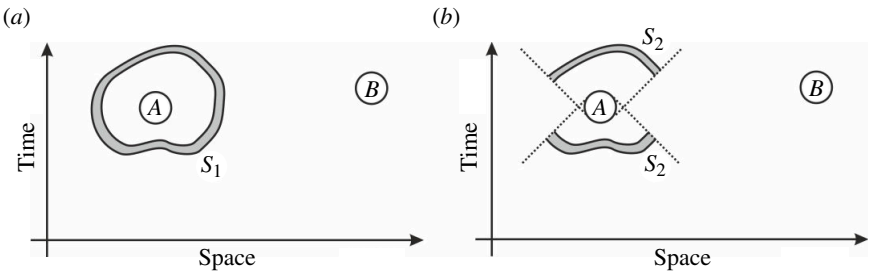

Figure 3. (*a*) CoA. (*b*) Strong CoA.

something that we should worry about. If one can understand the wavefunction as an epistemic state (a state of knowledge), then its non-local update is not *a priori* worrisome.

CoA loosely speaking means that localizable variables can only influence their nearest neighbours. However, it does not require that this influence lies within the light cones. To arrive at a stronger condition, we will therefore now as is customary assign the light-cones and their insides to a space–time region, A. We will denote with $L(A)$ the union of all space–time points that are light-like or time-like related to any point in A.

We can then cut out regions from the closed surface S_1 using the light-cones: $S_2 := S_1 \cap L(A)$ and arrive at a stronger version of CoA:

Strong CoA (locally subluminal): An m-model has Strong CoA if local beables outside the forward and backward light-cones of a region play no role for calculating outputs in that region $P(\mathcal{O}(A) \mid \mathcal{I}(S_2), \mathcal{I}(B)) = P(\mathcal{O}(A) \mid \mathcal{I}(S_2))\ \forall\ S_2$ that do not overlap with the light cones of B, i.e. $S_2 \cap L(B) = \varnothing$ (see figure 3*b*).

And correspondingly,

Weak CoA (locally superluminal): An m-model has Weak CoA if it fulfils CoA but not Strong CoA.

Weak CoA, roughly speaking, means that influences happen locally, but sometimes superluminally. What we call Strong CoA was called Einstein locality in Maudlin [37]. Local and non-local models can further be distinguished into those which are super- and subluminal. It is rather uncommon to consider subluminal non-locality, but it will be helpful in what follows to clearly distinguish non-locality from superluminality. We can define subluminal non-locality by requiring that the only local beables necessary to find out what happens at A are those within or on the light-cones of A, and adding the light-cones of B and their inside brings no further information.

Non-locally subluminal: An m-model is non-locally subluminal iff it is non-local, but local beables outside light-cones of a region play no role for calculating outputs in that region $P(\mathcal{O}(A) \mid \mathcal{I}(L(A)), \mathcal{I}(L(B))) = P(\mathcal{O}(A) \mid \mathcal{I}(L(A))$.

And, consequently,

Non-locally superluminal: An m-model is non-locally superluminal iff it is non-local, but not non-locally subluminal.

For a summary of those four terms, see figure 4.

We should also mention here that strong CoA is a weaker criterion than confining CoA inside the light cones, because of the requirement that one only considers regions S_2 which do not overlap with the light cones of region B. The reason for this requirement—as noted by Bell already—is that otherwise the outputs in B might well provide additional information that correlates with the inputs from S_2 (and hence, the outputs in A) without influences ever leaving the light-cones (see appendix for explanation). But, of course, if one extends the light cones of regions A and B far enough, they will always overlap. This means that one can always try to

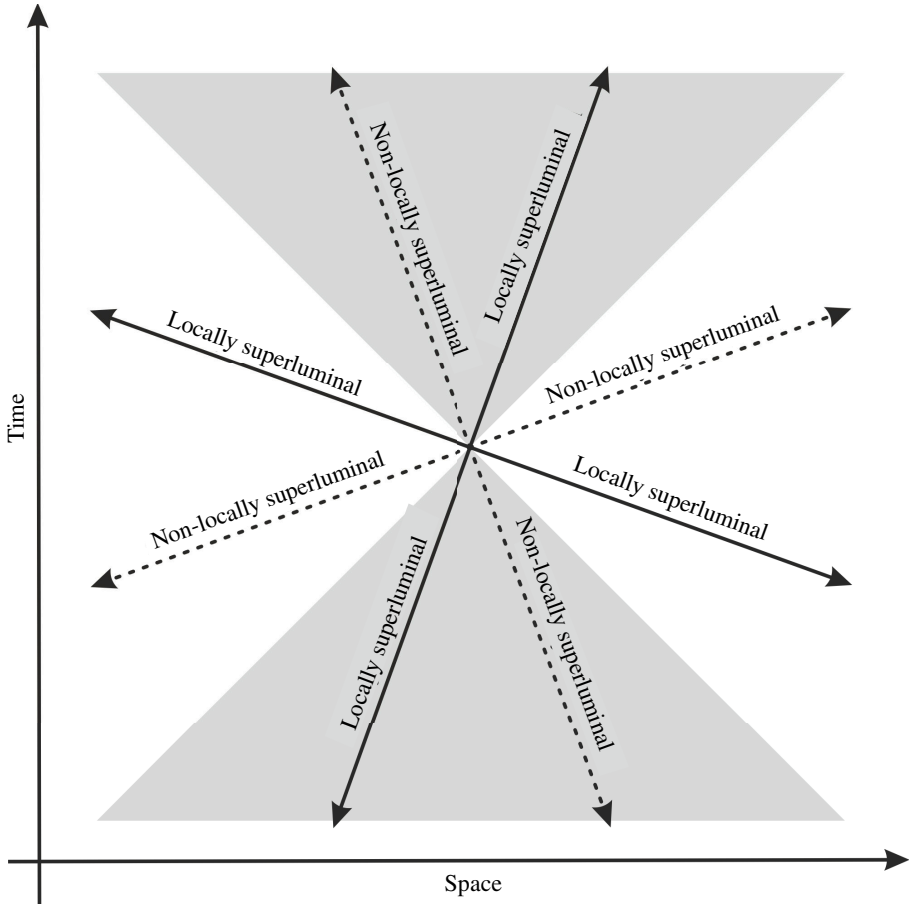

Figure 4. Difference between local, non-local and superluminal. Note that local and non-local are here distinguished by having solid/dotted lines respectively, and that each arrow is only one representative for the entire quadrant (e.g. one can imagine a reflected version of the left-to-right 'locally subluminal' arrow, going instead of right-to-left, which is also valid, so long as it is in both the past and future light cones).

explain violations of Strong CoA by locating an origin of correlations between A and B in the overlap of the light cones.[5] We will come back to this later.

(c) Properties of temporal order

Like with the local and temporal properties, to talk about temporal order, a model needs to have been provided with additional information. We need not only a notion of time, but also an arrow of time that allows us to distinguish past and future. This arrow of time is often not explicitly defined but implicitly assumed. Typically, it comes in because we assume that an experimenter can choose a setting in the future, but not one in the past. That is, the arrow of time sneaks in with what we consider to be a possible scenario.

Since an arrow of time is *a priori* a matter of definition, we have to specify that this definition has to be consistent for all mathematically equivalent expressions of a set-up.

There are many notions of arrows of time that have been discussed in the literature [38,39] and our aim here is not to unravel this discussion. We will merely note that a model needs to have such a notion for the following properties to be well-defined.

[5]This is often called a common 'cause'. However, all that is required here is a correlation, not a causation.

Like before, it is possible to have a model that just does not have a temporal order, or that does not distinguish past and future. Indeed, this is the case for many of the simplest models that we deal with, such as an undamped harmonic oscillator, or the two-body problem in Newtonian gravity.

Acausal: An m-model is acausal iff it does not have a well-defined arrow of time, and hence no notion of past or future.

An atemporal model is also acausal—one cannot have an arrow of time without having time to begin with. This feature is exhibited in the process matrix formalism, which can even be used to model processes for which it is impossible to specify a well-defined arrow of time [26,27]. One might plausibly argue that a model which is not time-reversible necessarily has an arrow of time, and hence cannot be acausal, but then we did not say who or what defines that arrow of time, so we cannot draw this conclusion.

As stressed earlier, some properties of models only make sense with an arrow of time that orders times. We now come to the first of those:

Temporally deterministic: A c-model is temporally deterministic if it is both deterministic, and has an arrow of time, according to which calculating localizable output values at time t' does not require inputs that are local beables at $t > t'$. An m-model is temporally deterministic if at least one of its c-models is.

This is a complicated way of saying that, in a temporarily deterministic model, a future state can be calculated with certainty from a past state, but not necessarily the other way round. This notion of temporally deterministic is what is often meant by the term 'deterministic'. Note that a temporally deterministic model might have other inputs besides value-assignments for local beables. In particular, a model with all-at-once inputs may still be temporally deterministic.

We have defined temporal determinism of an m-model from the requirement that at least one of its c-models has that property, because temporal determinism is easy to remove by redefining all variables so that they mix different times, or using (partially) time-like boundary conditions.

In the CI, so long as no measurement occurs, the state of the wavefunction at time t_a is temporally determined by the state of the wavefunction at time $t_b \neq t_a$ and the Hamiltonian operator. The state of the wavefunction after measurement, on the other hand, is generically not determined by the state of the wavefunction before measurement. Hence, the CI (c-model) is not temporally deterministic.

It does not follow from this that SQM, which we defined as $[\mathrm{CI}]_{\mathrm{m}}$, is also not temporarily deterministic. However, this is so because—as we saw earlier—SQM is not deterministic to begin with, so it cannot be temporarily deterministic either.

For temporal models, we can further define:

Time-reversible: An m-model is time-reversible iff it is both temporally deterministic, and also deterministic under the replacement $t \to -t$, where t measures time.

As mentioned earlier, this definition implicitly makes a statement about what properties we expect of time, and hence cannot stand on its own. That is not its purpose. Its purpose is rather to capture what properties physical properties like time and time-reversibility should have.

Time-reversibility should not be confused with invariance under time-reversal, which is a stronger requirement, but one that we will not consider here. Just because a model is time-reversible, does not mean that its time-reversed version is the same as the original.

Next, we recall a term previously introduced in Wharton and Argaman [22]:

Future input dependence (FID): A c-model has a FID iff, to produce output for time t, it uses local beables from at least one time $t' > t$ for at least one scenario.

FID is a property of the set-up of the c-model, and may simply be a matter of choice. For example, in any time-reversible c-model, one can replace a future input with a past input and get the same outputs. We define it here because it was previously used in Wharton & Argaman [22] and we want to make a connection to this definition below. However, we also want to define a somewhat-stronger property:

Future input requirement (FIR): A c-model has an FIR iff there is at least one scenario for which producing outputs for time t requires inputs from $t' > t$. An m-model has an FIR if all c-models in its equivalence class have an FIR.

Another way to phrase this is that a c-model with an FIR has at least one scenario for which the input cannot be entirely chosen in the past. A c-model with an FIR cannot be temporally deterministic. However, a model that is not temporally deterministic does not necessarily have an FIR.

A well-known example for an FIR is a space–time that is not globally hyperbolic, in which case the initial value problem is ill-posed. The time-evolution can then not be calculated unless further input is provided about what will happen.

Another example might be a model which evaluates possible policy paths to limit global temperature increase to certain goals within a certain period of time (say, 2, 3 or 4°C by 2050). Such a model requires specifying the desired boundary condition in the future. Of course, in this case, one is dealing with hypothetical scenarios, but the example illustrates that FIRs are used in practical applications.

(d) Properties of causal order

For models with a temporal order, we can distinguish the past/backwards and the future/forward light-cones of a region A, that we will denote $L_p(A)$ and $L_f(A)$, respectively. It is $L(A) = L_p(A) \cup L_f(A)$, and we define S_3: $= S_1 \cap L_p(A)$, the intersection of the shielding region S_1 with the past light-cone. With this, we can further refine CoA to a notion of local causality:

Local causality: An m-model is locally causal iff it fulfils strong CoA with local beables localized in the past light cone: $P(\mathcal{O}(A) \mid \mathcal{I}(S_3), \mathcal{I}(B)) = P(\mathcal{O}(A) \mid \mathcal{I}(S_3))\ \forall\ S_3$ that do not overlap with the light cones of B, i.e. $S_1 \cap L(B) = \varnothing$ (see figure 5*a*).

We want to stress that this notion of local causality is based on the mathematical structure of space–time, and so mixing it with other notions of causality can result in confusion. In particular, space–time causality is not *a priori* the same notion of causality that is used in large parts of the philosophical literature, which concerns itself with the question of cause-and-effect (one currently popular realization of which is the one based on causal models, that we will refer to as interventionist causality [8,9]).

According to space–time causality, of two causally related events, the one in the past is the cause by definition. One thus has to be careful when interpreting local causality using other notions of causality. This is especially important to keep in mind when we classify retrocausality.

If one interprets 'causal' as referring to space–time causality, then the term 'retrocausal' suggests influences that go inside the light cones, but backwards in time. The term retrocausal, however, does not necessarily imply locality. For example, the Transactional Interpretation [40,41], often referred to as retrocausal, is also non-local.

We will hence proceed by endowing the classification of the four local properties, summarized in figure 4 with an additional temporal distinction that is either forward in time or backward in time, according to the arrow of time that we have assumed exists. We would like to stress that since we have an arrow of time, we can meaningfully distinguish forward and backward in time also outside the light-cones. The reader may want to think of the arrow of

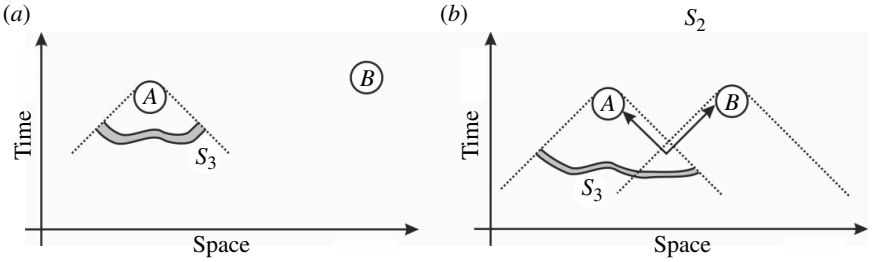

Figure 5. (*a*) Local causality. (*b*) A subtlety of local causality.

time as a preferred slicing in space–time. This will give us a total of eight distinctions. It is then straightforward to define:

Local retrocausality: An m-model is locally retrocausal iff it fulfils Strong CoA and has a future input requirement.
Non-local causality: An m-model is non-locally causal iff it is non-local, subluminal and has no future input requirement.

Non-local Retrocausality: An m-model is non-locally retrocausal iff it is non-local, subluminal and has a future input requirement.

As we noted already, the term 'causality' might either refer to the light-cone structure of space–time or to interventionist causality. However, since the term 'local causality' is extremely widely used and refers to space–time causality, we will use the term 'retrotemporal' to refer to models with a future input requirement that do not respect the light-cone structure. This gives us the remaining four definitions:

Locally superluminal retrotemporal: An m-model is locally superluminal retrotemporal if it fulfils Weak CoA and has a future input requirement.
Locally superluminal temporal: An m-model is locally superluminal temporal iff it fulfils Weak CoA but has no future input requirement.
Non-locally superluminal retrotemporal: An m-model is non-locally superluminal retrotemporal iff is non-locally superluminal and has a future input requirement.
Non-locally superluminal temporal: An m-model is non-locally superluminal temporal if it is non-local and superluminal, but has no future input requirement.

Since one can combine forward and backward causes to a zigzag [42], a locally retrocausal model might appear to be superluminal. However, whether such combinations are possible depends on the model.

The above types of retrocausality and retrotemporality are properties of the mathematical structure: a future input requirement cannot be removed by a mathematical redefinition, it is therefore a property of a representation of a model. An example for a locally retrocausal model might be General Relativity in space–times that have time-like closed curves.

However, there is a weaker notion of retrocausality that concerns the use of a c-model rather than its mathematical structure, the associated m-model. We can again distinguish four cases, which are the same as above, except that instead of a future input requirement, we merely have a future input dependence:

Local pseudo-retrocausality: A c-model is locally pseudo-retrocausal if it fulfils Strong CoA, and has a future input dependence.

One can similarly define the terms non-local pseudo-retrocausality, locally superluminal pseudo-retrotemporarity and non-locally superluminal pseudo-retrotemporality, by taking the definition without the 'pseudo' and replacing the future input requirement with a future input dependence.

The reason we use the prefix 'pseudo' is because according to our earlier definition a future input dependence is a matter of choice. It can be replaced with a past input, at least in principle. This does not mean that a future input dependence is unimportant, however. This is because it could be that removing the future input much increases the complexity of using the model. That is, future input dependence is linked to the question of whether a model is fine-tuned which will be further explored in the companion paper [3].

Causal and retrocausal non-locality can only be distinguished in the presence of an arrow of time, which for all practical purposes defines a space–time slicing. The same is the case for superluminal causal and superluminal retrocausal models—absent an arrow of time, they cannot be told apart, since Lorentz-transformations can convert one into the other.

We want to stress that pseudo-retrocausality is a property of a c-model, but not a property of an m-model. A retrocausal c-model cannot be temporally deterministic. A pseudo-retrocausal c-model, in contrast, can be temporally deterministic. It follows that temporal determinism is a simple way to tell pseudo-retrocausality from retrocausality. Note that pseudo-retrocausality was referred to just as retrocausality in Wharton and Argaman [22]. The practical use of pseudo-retrocausality will be discussed in the accompanying paper [3].

The reader who at this point despairs over the many different types of retrocausality will maybe understand now why the literature on the topic is so confusing, and hopefully also why a common terminology is necessary.

Some properties about causal structure just come from the definition of the arrow of time. In particular, we have to distinguish oriented and non-oriented arrows of time. A non-oriented arrow of time is one that allows forming loops in time (figure 6):

Dynamical retrocausality: An m-model is dynamically retrocausal iff its retrocausality is due to a non-oriented arrow of time.

Dynamical retrocausality may or may not be due to a space–time structure with closed time-like curves. It is important to emphasize that dynamical retrocausality is a property of the model, and not a property of the way the model uses inputs. Depending on how the arrow of time is defined, it may not be particularly meaningful. What makes an arrow of time meaningful, however, is not a question we want to unravel here.

Just for completeness, we also define:

Counter-causal: An m-model is counter-causal iff its time-reversed version is locally causal.

A model with such a property would makes one strongly suspect that the arrow of time was just defined the wrong way round to begin with. However, possibly there were other reasons to define an arrow of time that way.

A general comment is in order here, which is that the term 'retrocausal' is somewhat linguistically confusing. It does not so much refer to causes generally going backwards, but rather to them sometimes going against an arrow of time that was defined from something else. That is, it is really a mix of different directions of time that mark a retrocausal model, the already mentioned property that has previously been referred to as the possibility of zigzags in space–time [42]. Note that the zigzag property itself is defined against the presumed-to-exist GR arrow of time.

(e) Agent-based properties

Bell [28] further introduces local beables that are *controllable*, and those which are *not controllable*, a distinction that we will also use below. This notion is somewhat vague, but for our purposes, we do not need a precise definition. We will take a controllable local beable to be one whose value can in practice be set by the action of an experimentalist—typically this is a detector setting. For a more mathematically minded definition, see [43]. Note that for a local beable to be

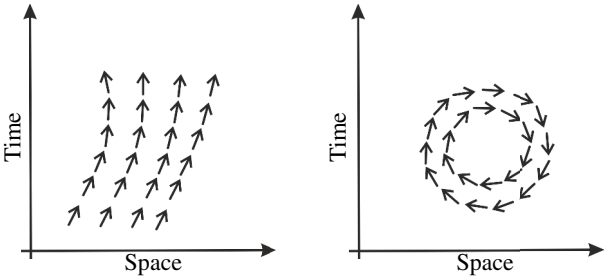

Figure 6. Orientable and non-orientable arrows of time. The arrows in the figure indicate the hypothetical flow of internal time of an observer, which might differ from the coordinate time. That is, the arrows are not in all places time-like, according to the coordinate time. This is supposed to illustrate the often-used example in which a spaceship that can travel faster than light in one frame actually seems to go back in time in another frame. If that was possible, one could use it to construct loops that seem to be 'timelike' from the point of view inside the spaceship (i.e. permissible motion for massive objects), but not according to the coordinate time.

controllable, an experimenter need not have free will, whatever that might mean, and they also do not have to control the local beable themselves; it could be done by some kind of apparatus.

If controllable input is correlated with observable output, we will speak of signalling. Signalling is particularly interesting if it is outside the forward light cones.

Superluminal signalling: A c-model allows superluminal signalling iff it is superluminal and has controllable inputs which are local beables in a region A that are correlated with observable outputs that are local beables in a region outside $L(A)$. An m-model allows superluminal signalling if at least one c-model in its class does.

SQM is non-local but does not allow superluminal signalling [44].

Since the time-order of space-like separated events can change under Lorentz-transformations, superluminal signalling is usually assumed to imply the possibility of signalling back in time. However, non-local signalling does not necessarily imply the possibility of signalling back in time when one has a time-order given by an arrow of time. In general relativity, for example, this may be a time-like vector field. One can then constrain non-local signals to only be forward in time relative to the vector field. For this reason, the two phenomena—signalling non-locally and signalling back in time—are distinct.

Let us then move on to signals that actually travel back in time:

Retrocausal signalling: A c-model allows retrocausal signalling if it is retrocausal and there is at least one scenario for which observables at time t are correlated with required controllable inputs from $t' > t$. An m-model allows retrocausal signalling if at least one c-model in its equivalence class does.

This retrocausality signalling could either be local or non-local. Like for the non-signalling case discussed in §3d, causal and retrocausal non-local signalling can only be distinguished in the presence of an arrow of time, and the same is the case for temporal and retrotemporal superluminal signalling, since Lorentz-transformations can mix both cases.

The problem with retrocausal signalling is that the observables at time t are, well, observable. If they can be affected by input at a later time $t' > t$, then the result may disagree with what one had already observed. This is what opens the door to causal paradoxes.

We will not introduce a notion of pseudo-retrocausal signalling, as that would be a technically possible definition, but rather oxymoronic. If a future input was removable and therefore not necessary for an earlier observable, then no signal was sent (though in such a case an agent might still have an illusion of signalling).

4. Specific model properties

In this section, we will now introduce properties that are specific to models typically used in the foundations of quantum mechanics.

Using the terminology of the previous sections, we can summarize the key conundrum of SQM by saying that models mathematically equivalent to the CI, $[\mathrm{CI}]_\mathrm{m}$, do not fulfil Strong CoA. However, SQM also has the odd property that observables generically do not have definite values before a measurement. This opens the possibility that one may still find a physical or empirical equivalent to the CI that fulfils Strong CoA.

Of course, one may be interested in interpretations or modifications of quantum mechanics for other reasons. For example, one may want to find a realist interpretation, or to return to determinism. However, the models we are here mainly concerned with are those which reinstate Strong CoA. Such models usually introduce new input variables. Using the terminology of Maudlin [45], we will call them additional variables:

Additional variables: Input values for an interpretation or modification of $[\mathrm{CI}]_\mathrm{m}$ that have no equivalent expression in $[\mathrm{CI}]_\mathrm{m}$.

Additional variables are not necessarily localized, or even localizable, and they are not necessarily hidden, though all of that might be the case. For example, in Bohmian Mechanics, particle positions are localizable, localized and hidden. Bohmian Mechanics, however, as we noted previously, does not fulfil CoA.

One might worry here that since the variables are 'additional', they are unnecessary to produce the same output as SQM, and therefore just make a model's set-up reducible. However, this does not have to be the case, because one normally introduces the additional variables to remove another assumption from $[\mathrm{CI}]_\mathrm{m}$. This is typically the measurement update postulate, since it is the one that leads to a violation of Strong CoA [1].

The purpose of additional variables reveals itself when one interprets the violation of Strong CoA in the CI as being due to the epistemic character of the wavefunction. One assumes that the real (ontic) state is not the wavefunction, but one that respects CoA, one just does not know which ontic state one is dealing with until a measurement was made. Quantum mechanics, in this picture, is just an incomplete description of nature.

We know from Bell's theorem that all such ensemble models with additional variables which determine the measurement outcome will violate an assumption to this theorem commonly known as statistical independence (SI) [46]:

SI: An m-model fulfils SI iff $P(\lambda \mid X, Y) = P(\lambda)$, where λ is (a set of) local beables on S_3 and X and Y quantify the settings of two detectors in regions A and B (compare with figure 5*a*).

From this, we can tell that all local interpretations or modifications of quantum mechanics can be classified by the ways in which they violate SI.[6] We will hence refer to them all as SI-violating models.

SI is also sometimes referred to as 'measurement independence,' or the 'free choice assumption' or the 'free will assumption' in Bell's theorem. In rare occasions, we have seen it being referred to as 'no conspiracy'. In recent years, theories which violate SI have also been dubbed 'contextual' [48], though the class of contextual models is larger than just those which violate SI (there is more 'context' to an experiment than its measurement setting).

Not all models that are being used in quantum foundations reproduce all predictions of quantum mechanics. Many of them only produce output for certain experimental situations, typically Bell-type tests, interferometers or Stern–Gerlach devices. We can then ask, in a

[6]It is sometimes argued that the Many Worlds Interpretation is a counterexample to this claim [47]. However, as laid out in the companion paper [3], the Many Worlds Interpretation is either not empirically equivalent to SQM, or violates CoA.

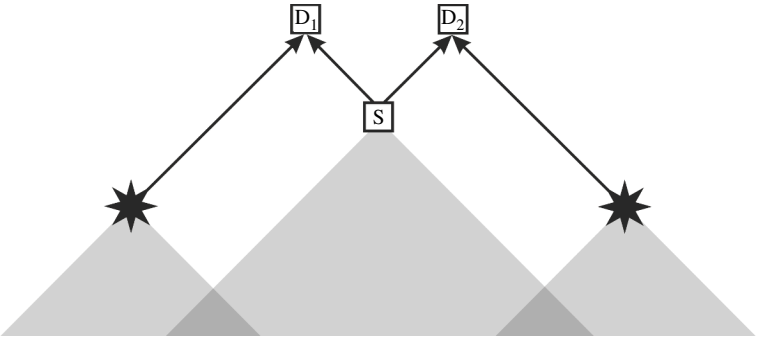

Figure 7. Sketch of Cosmic Bell test with source S and two detectors. The settings of the detectors D^1 and D^2 are chosen by using photons from distant astrophysical sources, usually quasars. Any past cause that could have given rise to the observed correlations must then have been very far back in time. Light cones indicated by grey shading.

hopefully obvious generalization of the above classification, whether these models are either representations or interpretations of $[CI]_m$ as it applies to the same experiments.

Traditionally, a distinction has been made between SI-violating models that are either retrocausal or superdeterministic. However, this distinction has remained ambiguous for three reasons.

First because—as we have seen already—retrocausality itself has been used for a bewildering variety of cases. If one then defines superdeterminism as those SI-violating models which are not retrocausal, one gets an equally bewildering variety. Second, since Bell (who coined the term 'superdeterminism') did not distinguish retrocausality from superdeterminism, one could reasonably argue that superdeterminism should be equated with SI-violation in general, and then consider retrocausality to be a variant of superdeterminism. Third, not everyone agrees that superdeterministic models have to be deterministic to begin with [49,50].

There is no way to define the term so that it agrees with all the ways it has previously been used. We will therefore just propose a definition that we believe agrees with the way it has most widely been used, based on the following reasoning:

(i) We are not aware of any superdeterministic model which is not also deterministic. Leaving aside that it is terrible nomenclature to speak of 'non-deterministic superdeterminism', there is not even an example for it. For this reason, we will assume that superdeterministic models are deterministic.
(ii) We want a model that fulfils Strong CoA, because this is the major reason why violations of SI are interesting, and it is also the context in which Bell coined the expression.
(iii) Most of the literature seems to consider superdeterminism and retrocausality as two disjoint cases, so we will do the same, even though Bell seems to not have used this distinction when he coined the term. This point together, with the previous one, implies that the model has to be locally causal.
(iv) We want a distinction that refers to the m-model, not to its particular realization as a c-model, to avoid new ambiguities.
(v) The model should reproduce the predictions of quantum mechanics, at least to a reasonable extent. This assumption is relevant because without it Newtonian mechanics would also be superdeterministic which makes no sense.
(vi) It is a one-world model that violates SI.

Some readers might argue that requirement 6 is strictly speaking unnecessary because it follows from Bell's theorem given the previous five requirements. However, since we did not explicitly list the assumptions to Bell's theorem, and it is somewhat controversial which of those have to be fulfilled in any case, we add it as an extra requirement.

Taking this together, we arrive at the following definition:

Superdeterminism: An m-model with additional variables which is deterministic, locally causal, violates SI, and is empirically equivalent to SQM (as in $[\mathrm{CI}]_e$).

With this definition, a superdeterministic model is distinct from a retrocausal model. However, a superdeterministic c-model can still be pseudo-retrocausal. This distinction has previously been made in Nikolaev and Vervoort [51]. The authors of this article used the term 'Soft Superdeterminism' for what in our nomenclature would be a pseudo-retrocausal superdeterministic c-model, and they use the term 'Hard Determinism' which in our nomenclature would be a not pseudo-retrocausal, superdeterministic c-model. In both cases though, the m-class belonging to the model is superdeterministic, and not retrocausal.

The notion of common-cause superdeterminism refers even more specifically to a certain type of superdeterministic models that are not pseudo-retrocausal. In these models, one assumes that the additional variables which determine the measurement outcomes in a Bell-type experiment are correlated with the settings, because they had a common cause in the past.

We want to emphasize that the additional requirement of common-cause superdeterminism is that there was a *common* cause, a more or less localized event that serves as what some would call an 'explanation' for the correlation that gives rise to violations of SI. This is opposed to the general superdeterministic models in which the correlation does not require explanation, but rather *is* the explanation (for our observations). The correlation that gives rise to violations of SI of course always has a space–time cause—in the most extreme case that would be the initial condition at the beginning of the universe. But, in general, the hidden variables and detector settings have no common cause in the interventionist sense, nor do they need have to have one.

That said, this particular type of common-cause superdeterminism can be tested by using methods to choose the detector settings that are so far apart from each other that any common cause would have had to be in the very early universe. This is the idea behind the Cosmic Bell test [52,53], sketched in figure 7.

However, while common-cause superdeterminism is a logical possibility, we are not aware of any superdeterministic model which is of this type, or of anyone who has advocated such a model.

Unfortunately, the results of Cosmic Bell tests are often overstated in the literature, quite frequently claiming to 'close the superdeterminism loophole', rather than just testing a particular type of superdeterministic model, which no one works on to begin with.

The relation between pseudo-retrocausality and finetuning (a.k.a. 'conspiracies') will be explored in a companion paper [3].

5. Classification

We want our classification scheme to be practically useful, so we will here provide a short guide to how it works.

The question we want to address is this: suppose you have a c-model for quantum mechanics—that is the thing you are doing your calculations with—what should you call it? We will assume that your model is presently empirically equivalent to SQM in the non-relativistic limit. (If it is not, you have bigger problems than finding a name for it.) You then proceed as follows:

(i) To classify the model, you first have to make sure that its set-up irreducible. If it is reducible, the set-up cannot be classified, so please remove all assumptions that are not necessary to calculate outputs. Assumptions that state the 'physical existence' (whatever that may be) of one thing or another are typically unnecessary for any calculation.

(ii) Figure out whether your model is physically equivalent to SQM. If it is not, it is a *modification*. If it is, it is an *interpretation*.

(iii) Figure out whether your model is mathematically equivalent to any already known interpretation. If it is, your model is a *representation* (of the interpretation it is mathematically equivalent to). If it is not, you have a representation of a new interpretation.

(iv) Go through the list of properties in§3 and §4 and note down whether your model has them, starting with the m-model properties, then the c-model properties.

6. Summary

We have proposed a classification scheme for models in the foundations of quantum mechanics. It is most central element is the distinction between different types of models: calculational, mathematical, physical and empirical. After distinguishing these different classes of models, we have defined some of their properties that are discussed most commonly in the foundations of quantum mechanics, with special attention to those concerning locality and causality. We hope that the here proposed terminology can aid to clarify which problems of quantum mechanics can and cannot be solved by interpretation.

Data accessibility. This article has no additional data.

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. E.A.: conceptualization, formal analysis, investigation, methodology, writing—original draft, writing—review and editing; J.H.: conceptualization, formal analysis, investigation, methodology, project administration, writing—original draft, writing—review and editing; S.H.: conceptualization, formal analysis, investigation, methodology, project administration, supervision, writing—original draft, writing—review and editing; T.P.: conceptualization, formal analysis, investigation, methodology, supervision, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein

Conflict of interest declaration. We declare we have no competing interests.

Funding. J.R.H. acknowledges support from Hiroshima University's Phoenix Postdoctoral Fellowship for Research, the University of York's EPSRC DTP grant EP/R513386/1 and the UK Quantum Communications Hub, funded by the EPSRC grants EP/M013472/1 and EP/T001011/1.

Acknowledgements. We want to thank Louis Vervoort and Ken Wharton for valuable feedback and all participants of the 2022 Bonn workshop on Superdeterminism and Retrocausality for helpful discussion.

Appendix

If the region S_3 was allowed to intersect with the inside of the past-lightcone of B, then in a theory which is not temporally deterministic, correlations could be created later which were not contained on S_3. In this case then, local beables at B could provide extra information for what happens at A though the information got there locally and inside the light-cones. For illustration, see figure 5*b*.

References

1. Hance JR, Hossenfelder S. 2022 What does it take to solve the measurement problem? *J. Phys. Commun.* **6**, 102001. (doi:10.1088/2399-6528/ac96cf)
2. Adlam E. 2023 Do we have any viable solution to the measurement problem? *Found. Phys.* **53**, 44. (doi:10.1007/s10701-023-00686-x)
3. Hossenfelder S. 2023 Quantum confusions, cleared up (or so I hope). *arXiv.* (doi:10.48550/arXiv.2309.12299)
4. Frigg R, Nguyen J. 2020 *Modelling nature: an opinionated introduction to scientific representation*. Cham: Springer. (doi:10.1007/978-3-030-45153-0)

5. MacFarlane J. 2020 *Philosophical logic: a contemporary introduction*. New York, NY: Routledge. (doi:10.4324/9781315185248)
6. Gödel K. 1931 Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatsh. f. Math. Physik.* **38–38**, 173–198. (doi:10.1007/BF01700692)
7. Smith CL. 1973 High energy behaviour and gauge symmetry. *Phys. Lett. B* **46**, 233–236. (doi:10.1016/0370-2693(73)90692-8)
8. Pearl J. 2009 *Causality*. Cambridge, UK: Cambridge University Press. (doi:10.1017/CBO9780511803161)
9. Spirtes P, Glymour C, Scheines R. 1993 *Causation, prediction, and search*. New York, NY: Springer. (doi:10.1007/978-1-4612-2748-9)
10. Glymour C. 1970 Theoretical realism and theoretical equivalence. In *PSA: proceedings of the biennial meeting of the philosophy of science association*, pp. 275–288, vol. **1970**. Cambridge, UK: Cambridge University Press. (doi:10.1086/psaprocbienmeetp.1970.495769)
11. Glymour C. 1980 *Theory and evidence*. NJ: Princeton University Press.
12. Weatherall JO. 2019 Part 1: Theoretical equivalence in physics. *Philos. Compass.* **14**, e12592. (doi:10.1111/phc3.12592)
13. Weatherall JO. 2019 Part 2: Theoretical equivalence in physics. *Philos. Compass.* **14**, e12591. (doi:10.1111/phc3.12591)
14. Argaman N. 2018 A lenient causal arrow of time? *Entropy* **20**, 294. (doi:10.3390/e20040294)
15. Muller F. 2013 Circumveiloped by obscuritads. the nature of interpretation in quantum mechanics, hermeneutic circles and physical reality, with cameos of James Joyce and Jacques Derrida. https://philsci-archive.pitt.edu/10166/
16. Zurek WH. 2003 Decoherence, einselection, and the quantum origins of the classical. *Rev. Mod. Phys.* **75**, 715–775. (doi:10.1103/RevModPhys.75.715)
17. Wallace D. 2022 The sky is blue, and other reasons quantum mechanics is not underdetermined by evidence. *Eur. J. Philos. Sci.* **13**, 54. (doi:10.48550/arXiv.2205.00568)
18. SuppesP. 1961 *In a comparison of the meaning and uses of models in mathematics and the empirical sciences*, pp. 163–177. Dordrecht: Springer Netherlands. (doi:10.1007/978-94-010-3667-2)
19. van Fraassen B. 1989 *Laws and symmetry*. Oxford, UK: Clarendon Press. (doi:10.1093/0198248601.001.0001)
20. Georgescu IM, Ashhab S, Nori F. 2014 Quantum simulation. *Rev. Mod. Phys.* **86**, 153–185. (doi:10.1103/RevModPhys.86.153)
21. Hangleiter D, Carolan J, Thébault KP. 2022 *Analogue quantum simulation*. Cham: Springer. (doi:10.1007/978-3-030-87216-8)
22. Wharton KB, Argaman N. 2020 Colloquium: bell's theorem and locally mediated reformulations of quantum mechanics. *Rev. Mod. Phys.* **92**, 021002. (doi:10.1103/RevModPhys.92.021002)
23. Cavalcanti EG, Wiseman HM. 2012 Bell nonlocality, signal locality and unpredictability (or what Bohr could have told Einstein at Solvay had he known about Bell experiments). *Found. Phys.* **42**, 1329–1338. (doi:10.1007/s10701-012-9669-1)
24. Adlam E. 2022 Determinism beyond time evolution. *Eur. J. Philos. Sci.* **12**, 73. (doi:10.1007/s13194-022-00497-3)
25. Spekkens RW. 2007 Evidence for the epistemic view of quantum states: a toy theory. *Phys. Rev. A* **75**, 032110. (doi:10.1103/PhysRevA.75.032110)
26. Oreshkov O, Giarmatzi C. 2016 Causal and causally separable processes. *New J. Phys.* **18**, 093020. (doi:10.1088/1367-2630/18/9/093020)
27. Oreshkov O, Costa F, Brukner C. 2012 Quantum correlations with no causal order. *Nat. Commun.* **3**, 1092. (doi:10.1038/ncomms2076)
28. Bell JS. 1975 *The theory of local beables*. Technical report CERN-TH-2053 CERN. See https://cds.cern.ch/record/980036?ln=en.
29. Brassard G, Raymond-Robichaud P. 2019 Parallel lives: a local-realistic interpretation of "nonlocal" boxes. *Entropy* **21**, 87. (doi:10.3390/e21010087)
30. Ciepielewski GS, Okon E, Sudarsky D. 2020 On superdeterministic rejections of settings independence. *Br. J. Philos. Sci.* **74**, 435–467. (doi:10.1086/714819)
31. Carroll SM, Remmen GN. 2017 A nonlocal approach to the cosmological constant problem. *Phys. Rev. D* **95**, 123504. (doi:10.1103/PhysRevD.95.123504)

32. Palmer TN. 2019 Bell inequality violation with free choice and local causality on the invariant set. *arXiv* (doi:10.48550/arXiv.1903.10537)
33. Hance JR, Hossenfelder S, Palmer TN. 2022 Supermeasured: violating bell-statistical independence without violating physical statistical independence. *Found. Phys.* **52**, 81. (doi:10.1007/s10701-022-00602-9)
34. Hehl FW, Mashhoon B. 2009 Nonlocal gravity simulates dark matter. *Phys. Lett. B* **673**, 279–282. (doi:10.1016/j.physletb.2009.02.033)
35. Croke S. 2022 Multipartite subspaces containing no locally inaccessible information. *Phys. Rev. A* **106**, 032421. (doi:10.1103/PhysRevA.106.032421)
36. Hance JR, Hossenfelder S. 2022 Bell's theorem allows local theories of quantum mechanics. *Nat. Phys.* **18**, 1382. (doi:10.1038/s41567-022-01831-5)
37. Maudlin T. 2012 *Philosophy of physics: space and time*. Princeton, NJ: Princeton University Press. (doi:10.1515/9781400842339)
38. Price H. 1996 *Time's arrow & Archimedes' point: new directions for the physics of time*. USA: Oxford University Press.
39. Albert DZ. 2000 *Time and chance*. MA: Harvard University Press. (doi:10.4159/9780674020139)
40. Cramer JG. 1986 The transactional interpretation of quantum mechanics. *Rev. Mod. Phys.* **58**, 647–687. (doi:10.1103/RevModPhys.58.647)
41. Kastner RE. 2012 *The transactional interpretation of quantum mechanics: the reality of possibility*. Cambridge University Press. (doi:10.1017/CBO9780511675768)
42. Price H, Wharton K. 2015 Disentangling the quantum world. *Entropy* **17**, 7752–7767. (doi:10.3390/e17117752)
43. Walleczek J, Grössing G. 2016 Nonlocal quantum information transfer without superluminal signalling and communication. *Found. Phys.* **46**, 1208–1228. (doi:10.1007/s10701-016-9987-9)
44. Ghirardi GC, Rimini A, Weber T. 1980 A general argument against superluminal transmission through the quantum mechanical measurement process. *Lett. Nuovo Cim.* **27**, 293–298. (doi:10.1007/BF02817189)
45. Maudlin T. 1995 Three measurement problems. *Topoi* **14**, 7–15. (doi:10.1007/BF00763473)
46. Hance JR, Hossenfelder S. 2022 The wave function as a true ensemble. *Proc. R. Soc. A* **478**, 20210705. (doi:10.1098/rspa.2021.0705)
47. Waegell M, McQueen KJ. 2020 Reformulating Bell's theorem: the search for a truly local quantum theory. *Stud. Hist. Philos. Sci. B Stud. Hist. Philos. Mod. Phys.* **70**, 39–50. (doi:10.1016/j.shpsb.2020.02.006)
48. Kupczynski M. 2023 Contextuality or nonlocality: what would John Bell choose today? *Entropy* **25**, 280. (doi:10.3390/e25020280)
49. Sen I, Valentini A. 2020 Superdeterministic hidden-variables models I: non-equilibrium and signalling. *Proc. R. Soc. A* **476**, 20200212. (doi:10.1098/rspa.2020.0212)
50. Sen I, Valentini A. 2020 Superdeterministic hidden-variables models II: conspiracy. *Proc. R. Soc. A* **476**, 20200214. (doi:10.1098/rspa.2020.0214)
51. Nikolaev V, Vervoort L. 2023 Aspects of superdeterminism made intuitive. *Found. Phys.* **53**, 17. (doi:10.1007/s10701-022-00648-9)
52. Gallicchio J, Friedman AS, Kaiser DI. 2014 Testing Bell's inequality with cosmic photons: closing the setting-independence loophole. *Phys. Rev. Lett.* **112**, 110405. (doi:10.1103/PhysRevLett.112.110405)
53. Rauch D *et al.* 2018 Cosmic bell test using random measurement settings from high-redshift quasars. *Phys. Rev. Lett.* **121**, 080403. (doi:10.1103/PhysRevLett.121.080403)