

Measurements, Disturbances and the Quantum Three Box Paradox

Abstract

A quantum pre- and post-selection paradox involves making measurements at two separate times on a quantum system, and making inferences about the state of the system at an intermediate time, conditional upon the observed outcomes. The inferences lead to predictions about the results of measurements performed at the intermediate time, which have been well confirmed experimentally, but which nevertheless seem paradoxical when inferences about different intermediate measurements are combined. The three box paradox is the paradigm example of such an effect, where a ball is placed in one of three boxes and is shuffled between the boxes in between two measurements of its location. By conditionalising on the outcomes of those measurements, it is inferred that between the two measurements the ball would have been found with certainty in Box 1 and with certainty in Box 2, if either box been opened on their own. Despite experimental confirmation of the predictions, and much discussion, it has remained unclear what exactly is supposed to be paradoxical or what specifically is supposed to be quantum, about these effects. In this paper I identify precisely the conditions under which the quantum three box paradox occurs, and show that these conditions are the same as arise in the derivation of the Leggett-Garg Inequality, which is supposed to demonstrate the incompatibility of quantum theory with macroscopic realism. I will argue that, as in Leggett-Garg Inequality violations, the source of the effect actually lies in the disturbance introduced by the intermediate measurement, and that the quantum nature of the effect is that no classical model of measurement disturbance can reproduce the paradox.

Keywords: pre- and post-selection, three box paradox, quantum theory, Leggett-Garg Inequality, macrorealism

1. Introduction

Pre- and post-selection statistics involve making measurements on a system, at two separate times, and using the observed outcomes to make inferences about the state of a system at an intermediate time. Classical probability theory suggests nothing problematical about such inferences. However, when the system and the measurements are quantum mechanical, apparently paradoxical conclusions can be reached.

The quantum three box paradox[1] is the paradigm example of such pre- and post-selection (PPS) paradoxes. A ball is placed in one of three boxes, and is shuffled between the boxes between the two measurements. The Aharonov-Bergman-Lebowitz rule gives the quantum mechanical pre- and post-selection rule for the probability of finding the ball in one of the boxes, if that box was opened at times between the two measurements. Paradoxically it predicts that with certainty the ball would be found in Box 1 and with certainty the ball would be found in Box 2, had either box been opened on their own.

The effect has been experimentally confirmed in a number of different contexts[2, 3] with tests[4] violating a measure of classicality by over seven standard deviations. Yet, despite much discussion in the literature, there is no clear consensus as to what is actually quantum about such effects, and which specific classical properties might be ruled out by such experiments. Questions raised include the validity of the counterfactual use of the Aharonov-Bergman-Lebowitz rule[5, 6, 7], connections to quantum contextuality proofs and measurement disturbance[8, 10] and whether classical systems can simulate the essential properties of a PPS paradox[9, 10, 11, 12].

This paper will analyse precisely what is non-classical about the three box paradox. The problem will be introduced in terms of an adversarial game between Alice and Bob. A PPS paradox occurs when Bob has good reason to believe that the game is fair, but Alice still wins disproportionately often. By carefully identifying the circumstances under which Bob's beliefs would seem justified, I will argue that the existence of a true PPS paradox requires a novel kind

of measurement disturbance which is necessarily *invasive*, but *operationally non-disturbing*. By relating this effect to the construction of a Leggett-Garg Inequality[13], I will show why no classical model of measurement disturbance can reproduce this effect.

Section 2 will introduce pre- and post-selections, describe the adversarial game and show how quantum theory seems to allow paradoxical effects. Section 3 will examine the role of measurement disturbance in the paradox. I will use the ontic model framework (introduced in [14]) to show how measurement disturbance plays a key role in such effects: for a true quantum PPS paradox, the measurement must be operationally non-disturbing, while in fact disturbing the actual state of the system¹. Previous models suggested[9, 10] are of a different kind: they involve operationally disturbing measurements that are not present in the three box paradox, so they cannot reproduce the same statistics.

Section 4 will show how this kind of disturbance can be related to violations of the Leggett-Garg Inequality[13], and that the three box paradox holds if, and only if, a related Leggett-Garg Inequality is violated. This is critical to understanding what is specifically quantum about the three box paradox: the type of measurement disturbance required cannot be reproduced in classical toy models, and violation of the Leggett-Garg Inequality gives a measure of the non-classicality of the PPS effect. This analysis has already been applied in experimental tests of the three box paradox[4], showing large violations of this classicality measure. The three box paradox does introduce a novel twist to the Leggett-Garg Inequality: in 2-dimensional Hilbert spaces the Leggett-Garg Inequality can only be violated with operationally disturbing measurements. The experimental test of the three box paradox was the first violation of the Leggett-Garg Inequality using operationally non-disturbing measurements.

Finally Section 5 I will consider quantum PPS paradoxes in general, and argue that they are specifically testing the quantum nature of measurement disturbance, and not quasi-classical conditions such as macrorealism or non-contextuality.

2. Pre- and Post-Selection Statistics and the Three Box Paradox

Classical probability theory makes no distinction between pre-selection and post-selection, and presents no problems for combining the two. Consider some property, Q , of a given system at three successive times: $t_1 < t_2 < t_3$, and call the value possessed by Q at time t_i , Q_i . The joint probability distribution for Q at the three successive times, $P(Q_1, Q_2, Q_3)$ can be conditionalised upon a particular value occurring at t_1 , to make predictive inferences about later times:

$$P(Q_2, Q_3|Q_1) = \frac{P(Q_1, Q_2, Q_3)}{\sum_{Q_2, Q_3} P(Q_1, Q_2, Q_3)} \quad (1)$$

This is a pre-selection of Q_1 at time t_1 . The expression $P(Q_2, Q_3|Q_1)$ is, under normal circumstances, expected to be the same as would be obtained if the property were initially prepared to take the value Q_1 at time t_1 . Strictly speaking, however, a pre-selection is different from determinately preparing a system to take a particular value for the property. A pre-selection involves the system potentially having many possible values for the property at t_1 , but then only considering those cases in which the property takes a particular value.

A post-selection at t_3 is equally straightforward:

$$P(Q_1, Q_2|Q_3) = \frac{P(Q_1, Q_2, Q_3)}{\sum_{Q_1, Q_2} P(Q_1, Q_2, Q_3)} \quad (2)$$

Again, a post-selection involves the system potentially having many possible values for the property at t_3 , but only considering the cases in which the property takes a particular one of those values. Post-selections differ from pre-selections in that typically there are no feasible equivalent processes for determinately ensuring a system takes a particular value for the property at the time t_3 . Nevertheless, performing post-selections, and making retrodictive inferences based upon them, is neither uncommon nor problematical.

¹Although this is similar to a result obtained by Leifer and Spekkens[10], they proved only that a non-contextual model must introduce measurement disturbances. They do not demonstrate that a contextual model must involve disturbance, nor do they give an example of a non-contextual model that reproduces the three box paradox statistics.

Combining the two creates a pre- and post-selection:

$$P(Q_2|Q_1, Q_3) = \frac{P(Q_1, Q_2, Q_3)}{\sum_{Q_2} P(Q_1, Q_2, Q_3)} \quad (3)$$

There is still nothing problematical about this, provided the pre- and post-selected values are not incompatible ($\sum_{Q_2} P(Q_1, Q_2, Q_3) \neq 0$). The pre- and post-selection only considers those cases for which the specific values occur at t_1 and t_3 , and inferences are made about the properties of the system at an intermediate time $t_1 < t_2 < t_3$.

Turning now to quantum theory, applying the pre- and post-selection rule to a sequence of three projective measurements gives the Aharonov-Bergman-Lebowitz rule[15]:

$$P(Q_2|Q_1, Q_3) = \frac{|\langle Q_1 | Q_2 \rangle|^2 |\langle Q_2 | Q_3 \rangle|^2}{\sum_{Q_2} |\langle Q_1 | Q_2 \rangle|^2 |\langle Q_2 | Q_3 \rangle|^2} \quad (4)$$

This rule follows from normal quantum measurement statistics, is experimentally well verified, and no-one is suggesting that it is incorrect, so how can it seemingly lead to contradictory inferences about Q_2 at time t_2 ? To explore this, I will now give an overview of the three box paradox.

2.1. Alice and Bob's adversarial game

Alice proposes a game to Bob([16]). She has three indistinguishable boxes, into the third one she places a ball. Then she places a cover over the boxes and shuffles the ball between the boxes. The cover is removed and Bob is allowed to look in his choice of either Box 1 or Box 2. Alice is blindfolded while Bob looks, so she is unaware which box Bob looks in or whether he sees the ball (a trusted impartial umpire is present to ensure Bob does not cheat). The boxes are covered again, and the umpire then shuffles the ball between the boxes, according to instructions already given by Alice. Finally Alice gets to check if the ball is now in Box 3, and is then allowed to choose to play or pass. On rounds when she plays, she wins if Bob saw the ball in the box he opened, and loses whenever Bob did not. If she passes, the game is considered a draw. She offers Bob odds that are significantly better than fifty-fifty that she will win when she plays.

It is straightforward to show that, provided Bob's test leaves no mark for Alice to read (either of which box Bob checked, or whether the ball was seen) and Bob is equally likely to choose either box, Alice has no better than a fifty-fifty chance of winning.

Bob is understandably suspicious. He thinks carefully, and demands a number of checks² before playing:

- Condition (I) Bob reasons that if the final shuffle somehow provides information to Alice about which box he looked in, she can improve her odds of success. He therefore checks that, no matter which box he opens before the final shuffle, the relative frequency with which the ball ends up in Box 3 is the same as when no box is opened.
- Condition (II) Bob now wonders if some mark is being left by the act of opening a box. He is allowed to choose a method to test if the ball is in a given box, which he believes cannot disturb the box or the ball. For example, the boxes could be left open under the cover, and he carefully reaches his hand under the cover and inside his chosen box without touching the sides. If the box is empty, Bob will discover this without anything being disturbed. Or there could be a spring connected to the top of each box and Bob could gently stretch the spring until it exerts a lifting force on the box slightly greater than the weight of the box, but well below that of the box and ball together. While an empty box is lifted, a full box should be left undisturbed.
- Condition (III) As a further check, he is allowed to verify before playing that his test is not leaving any mark he can find. He can place the ball in any given box, and perform his test on one of the boxes. He then compares the measurement statistics of any further checks he likes after his test, with the measurement statistics for simply placing the the ball in the box and making those further checks *without* performing his test. He finds that, regardless of the box in which he places the ball, his test seems to have no measureable effect on the later statistics.

²The significance of which may not be immediately apparent.

After thinking a little more, Bob requires one additional condition:

Condition (IV) If Bob chooses, he can instead look in each of the two boxes in turn. If Bob finds a ball in both boxes, then he wins immediately. Otherwise he loses immediately.

Finally Bob decides to play the game. He can't see any way Alice can know which box Bob checked or that there is any mark being left of whether he saw the ball.

Alice then chooses to play whenever she sees the ball in Box 3, and wins every time she plays.

2.2. The quantum description of the game

Of course, Alice is using quantum theory to win.

The quantum state for the ball being in Box Q will be given by $|Q\rangle$, and Q_i represents the the ball being in Box Q at time t_i . The initial preparation of the system is to be in state $|3\rangle$ at time t_1 . Alice's initial shuffle is any unitary which satisfies

$$U_I |3\rangle = \frac{1}{\sqrt{3}} (|1\rangle + |2\rangle + |3\rangle)$$

Three possibilities are allowed for Bob's choice of intervening measurement at time t_2 :

- $B1$: A projective measurement onto $|1\rangle\langle 1|, (|2\rangle\langle 2| + |3\rangle\langle 3|)$. Outcome 1_2 represents the ball being found in Box 1 at time t_2 , and $\neg 1_2$ represents Box 1 being empty at time t_2 ;
- $B2$: A projective measurement onto $|2\rangle\langle 2|, (|1\rangle\langle 1| + |3\rangle\langle 3|)$. Outcome 2_2 represents the ball being found in Box 2 at time t_2 , and $\neg 2_2$ represents Box 2 being empty at time t_2 ;
- N : Doing nothing. No box is checked. This allows Bob to verify Condition (I) holds.

After the Bob's measurement, the final shuffle is any unitary which satisfies

$$U_F \frac{1}{\sqrt{3}} (|1\rangle + |2\rangle - |3\rangle) = |3\rangle$$

Alice's final measurement at time t_3 , is a projective measurement, A , onto $|3\rangle\langle 3|, (|1\rangle\langle 1| + |2\rangle\langle 2|)$, with outcome 3_3 representing the ball being found in Box 3 at time t_3 , and $\neg 3_3$ representing Box 3 being empty (from Section 3.2 onwards, the analysis will be simplified by merging U_I into the pre-selection and U_F into A .)

The three sequences give the statistics:

$(B1, A)$	$P_{(B1,A)}(1_2, 3_3) = 1/9$	$P_{(B1,A)}(\neg 1_2, 3_3) = 0$
	$P_{(B1,A)}(1_2, \neg 3_3) = 2/9$	$P_{(B1,A)}(\neg 1_2, \neg 3_3) = 2/3$
$(B2, A)$	$P_{(B2,A)}(2_2, 3_3) = 1/9$	$P_{(B2,A)}(\neg 2_2, 3_3) = 0$
	$P_{(B2,A)}(2_2, \neg 3_3) = 2/9$	$P_{(B2,A)}(\neg 2_2, \neg 3_3) = 2/3$
(N, A)	$P_{(N,A)}(3_3) = 1/9$	

Table 1: Three Box Paradox Statistics

Using the Aharonov-Bergmann-Lebowitz rule³, with a pre-selection of 3_1 represented by $|Q_1\rangle = \frac{1}{\sqrt{3}} (|1\rangle + |2\rangle + |3\rangle)$, and a post-selection of 3_3 represented by $|Q_3\rangle = \frac{1}{\sqrt{3}} (|1\rangle + |2\rangle - |3\rangle)$, setting the intermediate outcome to $|Q_2\rangle = |1\rangle$ gives $P_{(B1,A)}(1_2|3_1, 3_3) = 1$ while setting the intermediate outcome to $|Q_2\rangle = |2\rangle$ gives $P_{(B2,A)}(2_2|3_1, 3_3) = 1$. Whenever Alice sees the ball in Box 3, she knows with certainty that Bob saw the ball, whether or not Bob looked in Box 1 or Box 2.

What seems paradoxical about these statistics is that, if Bob was able to perform his measurements without disturbing the system, both of Alice's pre- and post-selective inferences should be

³When using non-selective outcomes at the intermediate time, such as $(|2\rangle\langle 2| + |3\rangle\langle 3|)$ or $(|1\rangle\langle 1| + |3\rangle\langle 3|)$, Equation 4 must use these projectors for the intermediate outcomes, replacing $|\langle Q_1|Q_2\rangle|^2 |\langle Q_2|Q_3\rangle|^2$ with $|\langle Q_1|Q_2|Q_3\rangle|^2$.

simultaneously valid. This means the ball had to be in Box 1 with certainty at t_2 , and to be in Box 2 with certainty at t_2 .

It was to check this that Bob introduced his Condition (IV). The statistics Bob collects when he looks in both boxes is:

$(B1, B2)$	$P_{(B1,B2)}(1_2, 2_2) = 0$	$P_{(B1,B2)}(-1_2, 2_2) = 1/3$
	$P_{(B1,B2)}(1_2, -2_2) = 1/3$	$P_{(B1,B2)}(-1_2, -2_2) = 1/3$
$(B2, B1)$	$P_{(B2,B1)}(2_2, 1_2) = 0$	$P_{(B2,B1)}(-2_2, 1_2) = 1/3$
	$P_{(B2,B1)}(2_2, -1_2) = 1/3$	$P_{(B2,B1)}(-2_2, -1_2) = 1/3$

Table 2: Bob looks in both boxes.

Alice does not appear to be cheating: the ball is never found in both boxes simultaneously. Yet the paradox appears to be that the ball must have been in both boxes simultaneously, whenever Alice observes the post-selection 3_3 . Indeed, even if the pre- and post-selection provides a less than certain inference to Alice, any result where $P_{(B1,A)}(1_2|3_1, 3_3) + P_{(B2,A)}(2_2|3_1, 3_3) > 1$ seems to be paradoxical, and allows Alice to win the adversarial game on average, while offering seemingly fair odds to Bob.

What went wrong with the reasoning that led Bob to think the game was fair? What, if anything, is specifically *quantum* about the statistics of Tables 1 and 2? Ravon and Vaidman[11] argue that it is something to do with Bob believing that his measurement does not disturb the system. They argue that in all the attempts to model pre- and post-selection paradoxes within classical systems, the intervening measurement must leave a mark to make the post-selection impossible ie. when Box 1 is opened, if the ball is *not* observed, then there is some record left of this which prevents the ball ending up in Box 3 for Alice's measurement. In the case of the three box paradox, they argue there is no reason to suppose that a classical measurement can disturb the system in this way. To analyse this more carefully, I will now look at the role of quantum measurement disturbance in the three box paradox.

3. PPS Paradoxes and measurement disturbance

3.1. Operationally non-disturbing measurements

I will now highlight an essential element of a PPS paradox, absent from the discussion of [9, 10, 11]: that Bob's measurements should be operationally non-disturbing for Alice.

Bob's measurement B_i is an operationally non-disturbing measurement for Alice, if and only if, for each outcome Q_3 Alice can observe:

$$P_{(N,A)}(Q_3) = \sum_{Q_2} P_{(Bi,A)}(Q_2, Q_3) \quad (5)$$

When this condition holds, Alice can gain no information about what measurement Bob performed, or if Bob even performed a measurement, from the statistics of her measurement outcomes. This is Bob's Condition (I), and is clearly satisfied by Table 1.

This might seem a reasonable requirement, in itself. After all, in some of the adversarial games considered (including the three box paradox) it is not hard to see that if Alice were to have information about what measurement Bob performed, then she could improve her odds of winning at the adversarial game without needing to resort to quantum theory. Although such situations allow Alice to win, there is no special mystery how.

Perhaps more importantly, Sharp and Shanks[5], and Cohen[6] demonstrated that attempting to combine post-selective inferences when Bob's measurement changes the statistics of Alice's measurement results, will in general lead to inconsistent probabilistic predictions.

A simple example of the problem can be given. Suppose Bob performs a measurement B , which has a particular outcome Q_1 , and then Alice performs measurement A , with outcomes Q_2 and $\neg Q_2$. Alice can make post-selective inferences from the conditional probabilities $P_{(B,A)}(Q_1|Q_2) = P_{(B,A)}(Q_1, Q_2)/P_{(B,A)}(Q_2)$ and $P_{(B,A)}(Q_1|\neg Q_2) = P_{(B,A)}(Q_1, \neg Q_2)/P_{(B,A)}(\neg Q_2)$.

Suppose Alice assumes that these post-selective inferences are valid even if Bob did not make an intervening measurement. When Alice observes outcome Q_2 , she infers Bob would have observed Q_1 with probability $P_{(B,A)}(Q_1|Q_2)$, had he actually made the measurement. Similarly if she observes outcome $\neg Q_2$, she infers Bob would have observed Q_1 with probability $P_{(B,A)}(Q_1|\neg Q_2)$, had he actually made the measurement.

Now if these inferences are indeed valid when Bob does not make the measurement, and Alice observes her outcomes occurring with probabilities $P_{(N,A)}(Q_2)$ and $P_{(N,A)}(\neg Q_2)$, Alice is led to calculate that, had Bob actually made the measurement, he would have observed Q_1 with probability $P(Q_1) = P_{(B,A)}(Q_1|Q_2)P_{(N,A)}(Q_2) + P_{(B,A)}(Q_1|\neg Q_2)P_{(N,A)}(\neg Q_2)$. This is plainly inconsistent with Alice's knowledge that, had Bob actually made the measurement, he would have observed Q_1 with probability $P(Q_1) = P_{(B,A)}(Q_1|Q_2)P_{(B,A)}(Q_2) + P_{(B,A)}(Q_1|\neg Q_2)P_{(B,A)}(\neg Q_2)$. The only way the combination of inferences could be consistent with Alice's knowledge is if $P_{(N,A)}(Q_2) = P_{(B,A)}(Q_2)$ and $P_{(N,A)}(\neg Q_2) = P_{(B,A)}(\neg Q_2)$: in other words, if Bob's measurement is operationally non-disturbing for Alice.

Clearly any counterfactual use of pre- and post-selection statistics is inconsistent if the intervening measurement is operationally disturbing. As is shown in the Appendices, *none* of the classical models of pre- and post-selection paradoxes presented in the literature[9, 10, 11, 12] involved operationally non-disturbing measurements. All involve intervening measurements which change the statistics of the final post-selection measurement.

This is not the case in the three box paradox. Alice's paradoxical inferences, that the ball is simultaneously in Box 1 and in Box 2, do not involve an operational disturbance by Bob. This leaves open the question as to whether a classical model can, in fact, reproduce the statistics of the three box paradox, in full, and leaves open whether the intervening measurements must disturb the system, even though they are not operationally disturbing.

3.2. *Optically non-invasive measurements*

The idea that a measurement introduces a disturbance into a system does not necessarily mean that the measurement is operationally disturbing. Equivalently, the fact that a measurement is operationally non-disturbing should not be taken to imply that no disturbance took place. To analyse this more carefully, it is helpful to introduce the ontological models framework[14, 17, 18, 19], which is a generalisation of Bell's notion of beables[20].

In this framework, any given experimental arrangement is characterised by a *preparation* process, E , and a *measurement* process, M , with distinct outcomes Q . Operationally, this is characterised by a probability $P_{(E,M)}(Q)$. In the case of quantum theory $P_{(E,M)}(Q) = |\langle Q|E\rangle|^2$.

The ontic state of the system represents the actual state of the world, between the preparation and the measurement. It is intended to represent all of the real physical properties the system possesses, independently of their observation or measurement. For example, in classical n -body statistical mechanics, the ontic state is a point in the n -body phase space. In wavefunction realist approaches to quantum theory the ontic state includes the wavefunction itself. Formally:

1. A preparation process E produces a probabilistic distribution $\mu_E(\lambda)$ over the ontic states λ . Any convex sum $\mu(\lambda) = \sum w_E \mu_E(\lambda)$ ($\sum w_E = 1$, $w_E \geq 0$), is also a valid preparation.
NB. $\int d\lambda \mu_E(\lambda) = 1$.
2. A measurement M is represented by a set of response functions, $\xi_M(Q|\lambda)$, each giving the probability of a different outcome Q occurring, conditional upon the actual ontic state of the system.
NB. $\sum_Q \xi_M(Q|\lambda) = 1$
3. The operational probabilities must be recovered through the formula:

$$P_{(E,M)}(Q) = \int d\lambda \mu_E(\lambda) \xi_M(Q|\lambda). \quad (6)$$

This encapsulates the idea that it is the properties of the ontic state which capture all the physical connections between the preparation process and the measurement outcomes.

If two different preparation procedures produce the same measurement statistics for all possible measurement processes, they are said to be *operationally equivalent* preparations. If

two different measurement procedures produce the same measurement statistics (up to a permutation of the outcomes) for all possible preparation processes, they are said to be *operationally equivalent* measurements.

4. The disturbance of the ontic state by the measurement M , when outcome Q occurs, is represented by the probability distribution⁴ of transitions to other ontic states: $\gamma_M(\lambda_2|Q, \lambda_1)$. NB. $\int d\lambda_2 \gamma_M(\lambda_2|Q, \lambda_1) = 1$

After the measurement, conditionalising on outcome Q having occurred is equivalent to a new preparation process:

$$\mu_Q(\lambda) = \frac{\int d\lambda_1 \mu_E(\lambda_1) \xi_M(Q|\lambda_1) \gamma_M(\lambda|Q, \lambda_1)}{\int d\lambda_1 \mu_E(\lambda_1) \xi_M(Q|\lambda_1)} \quad (7)$$

A measurement which has no effect on the ontic state is called *ontically non-invasive*. It is represented by

$$\gamma_M(\lambda_2|Q, \lambda_1) = \delta(\lambda_2 - \lambda_1) \quad (8)$$

It is trivial so show that an ontically non-invasive measurement is also operationally non-disturbing. The reverse is not necessarily true.

While Bob's measurements in the three box paradox are operationally non-disturbing, at least one must necessarily be ontically invasive. The application of the ontic model formalism quickly gives:

$$P_{(B1,A)}(1_2, 3_3) = \int d\lambda_2 d\lambda_1 \mu(\lambda_1) \xi_{B1}(1_2|\lambda_1) \gamma_{B1}(\lambda_2|1_2, \lambda_1) \xi_A(3_3|\lambda_2) \quad (9)$$

$$P_{(B2,A)}(2_2, 3_3) = \int d\lambda_2 d\lambda_1 \mu(\lambda_1) \xi_{B2}(2_2|\lambda_1) \gamma_{B2}(\lambda_2|2_2, \lambda_1) \xi_A(3_3|\lambda_2) \quad (10)$$

$$P_{(N,A)}(3_3) = \int d\lambda_1 \mu(\lambda_1) \xi_A(3_3|\lambda_1) \quad (11)$$

$$\begin{aligned} P_{(B1,A)}(1_2, 3_3) + P_{(B2,A)}(2_2, 3_3) &= \int d\lambda_2 d\lambda_1 \mu(\lambda_1) \xi_A(3_3|\lambda_2) [\xi_{B2}(2_2|\lambda_1) \gamma_{B2}(\lambda_2|2_2, \lambda_1) \\ &\quad + \xi_{B1}(1_2|\lambda_1) \gamma_{B1}(\lambda_2|1_2, \lambda_1)] \end{aligned} \quad (12)$$

Consider the overlap in the ontic state space between the support of the functions $\xi_{B1}(1_2|\lambda)$ and $\xi_{B2}(2_2|\lambda)$. Are there ontic states for which $\mu(\lambda) > 0$ and $\xi_{B1}(1_2|\lambda) \xi_{B2}(2_2|\lambda) \neq 0$?

If $B1$ is ontically non-invasive, then a measurement of $B1$ followed by $B2$ would give the result:

$$\begin{aligned} P_{(B2,B1)}(2_2, 1_2) &= \int d\lambda_2 d\lambda_1 \mu(\lambda_1) \xi_{B1}(1_2|\lambda_1) \gamma_{B1}(\lambda_2|1_2, \lambda_1) \xi_{B2}(2_2|\lambda_2) \\ &= \int d\lambda_1 \mu(\lambda_1) \xi_{B1}(1_2|\lambda_1) \xi_{B2}(2_2|\lambda_1). \end{aligned} \quad (13)$$

If there is a non-zero-measure overlap, then $P_{(B2,B1)}(2_2, 1_2) > 0$. But this means Bob could open Box 1, see a ball inside, then open Box 2 and see a second ball! Bob would clearly cry "foul" at this point! After all, Alice simply putting a ball in both boxes is a very easy way for her to win, and involves no paradox at all (see Appendix A.4). Bob's Condition (IV) rules this out: $P_{(B2,B1)}(2_2, 1_2) = 0$, as Table 2 shows. Hence the three box paradox requires that, if either $B1$ or $B2$ are ontically non-invasive, then $\xi_{B1}(1_2|\lambda) \xi_{B2}(2_2|\lambda) = 0$.

Now $\xi_{B1}(1_2|\lambda) \leq 1$ and $\xi_{B2}(2_2|\lambda) \leq 1$ which together with $\xi_{B1}(1_2|\lambda) \xi_{B2}(2_2|\lambda) = 0$ gives

$$\xi_{B2}(2_2|\lambda) + \xi_{B1}(1_2|\lambda) \leq 1. \quad (14)$$

It follows that if $B1$ and $B2$ are both ontically non-invasive, so $\gamma_{B1}(\lambda_2|1_2, \lambda_1) = \gamma_{B2}(\lambda_2|2_2, \lambda_1) = \delta(\lambda_2 - \lambda_1)$:

$$\begin{aligned} P_{(B1,A)}(1_2, 3_3) + P_{(B2,A)}(2_2, 3_3) &= \int d\lambda_1 \mu(\lambda_1) \xi_A(3_3|\lambda_1) [\xi_{B2}(2_2|\lambda_1) + \xi_{B1}(1_2|\lambda_1)] \\ &\leq \int d\lambda_1 \mu(\lambda_1) \xi_A(3_3|\lambda_1) = P_{(N,A)}(3_3) \end{aligned} \quad (15)$$

⁴Assume $\gamma_M(\lambda_2|Q, \lambda_1) = 0$ when $P(Q, \lambda_1) = 0$.

But the three box paradox occurs precisely because $P_{(B1,A)}(1_2, 3_3) + P_{(B2,A)}(2_2, 3_3) > P_{(N,A)}(3_3)$. As operational non-disturbance gives $P_{(N,A)}(3_3) = P_{(B1,A)}(3_3) = P_{(B2,A)}(3_3)$, it is simple to rewrite this as $P_{B1}(1_2|3_3) + P_{B2}(2_2|3_3) > 1$. This is precisely the condition that allows Alice to offer fifty-fifty odds to Bob, yet still expect to win the adversarial game on average.

Hence there are no possible ontic models for the three box paradox for which $B1$ and $B2$ are both ontically non-invasive, despite the fact that $B1$ and $B2$ are both operationally non-disturbing for Alice.

4. PPS Paradoxes and the Leggett-Garg Inequality

The paradox requires an ontically invasive measurement that is nevertheless operationally non-disturbing for Alice. This alone does not show that there is anything especially non-classical. I will now show that the three box paradox may be connected to violations of the Leggett-Garg Inequality[13], and that no classical model for the measurement disturbance can reproduce the paradox.

Two ideas seem necessary for Bob to believe that he has a fair chance at the adversarial game: that the ball is always in one and only one box; and that his measurement does not disturb the system. These assumptions are essentially the same assumptions that have been discussed extensively in the context of the Leggett-Garg Inequality, under the names *macrorealism* and *non-invasive measurability*:

- “1. Macrorealism *per se*. A macroscopic system which has available to it two or more macroscopically distinct states is at any given time in a definite one of those states.
2. Non-invasive measurability. It is possible in principle to determine which of these states the system is in without any effect on the state itself or on the subsequent system dynamics.” [13]

As in the case of PPS paradoxes, Leggett and Garg consider a sequence of three measurements, at times $t_1 < t_2 < t_3$, with a joint probability $P(Q_1, Q_2, Q_3)$ for a property of the system to possess the value Q_i at time t_i . They restrict the property to the values ± 1 and consider the expression $Q = Q_1Q_2 + Q_1Q_3 + Q_2Q_3$.

Q_1	Q_2	Q_3	Q
-1	-1	-1	+3
-1	-1	+1	-1
-1	+1	-1	-1
-1	+1	+1	-1
+1	-1	-1	-1
+1	-1	+1	-1
+1	+1	-1	-1
+1	+1	+1	+3

Table 3: The Leggett-Garg Function

Table 3 shows that for any probability distribution over these sequences of outcomes, $-1 \leq \langle Q \rangle \leq 3$. Leggett and Garg then consider a different arrangement, where on each run of the experiment, only two out of the three measurements are performed, giving $P_{(M_1, M_2)}(Q_1, Q_2)$, $P_{(M_1, M_3)}(Q_1, Q_3)$ and $P_{(M_2, M_3)}(Q_2, Q_3)$, from which the expression

$$Q_{LG} = \langle Q_1Q_2 \rangle_{(M_1, M_2)} + \langle Q_1Q_3 \rangle_{(M_1, M_3)} + \langle Q_2Q_3 \rangle_{(M_2, M_3)} \quad (16)$$

is calculated.

If the measurements were performed non-invasively, then clearly

$$\begin{aligned} P_{(M_1, M_2)}(Q_1, Q_2) &= \sum_{Q_3} P_{(M_1, M_2, M_3)}(Q_1, Q_2, Q_3) \\ P_{(M_1, M_3)}(Q_1, Q_3) &= \sum_{Q_2} P_{(M_1, M_2, M_3)}(Q_1, Q_2, Q_3) \end{aligned}$$

$$P_{(M_2, M_3)}(Q_2, Q_3) = \sum_{Q_1} P_{(M_1, M_2, M_3)}(Q_1, Q_2, Q_3) \quad (17)$$

and $-1 \leq Q_{LG} \leq 3$ would hold. This is the Leggett-Garg Inequality, and it can be violated by quantum theory.

As argued in [21], Equation 17 does not require ontic non-invasiveness, but does express the weaker condition that the measurements be operationally non-disturbing. Violation of the Leggett-Garg Inequality therefore requires operationally disturbing measurements. This would suggest that PPS paradoxes cannot be associated with Leggett-Garg Inequality violations. However, the Leggett-Garg Inequalities have been studied almost exclusively in the context of 2-dimensional Hilbert spaces. In the case of 3-dimensional Hilbert spaces, I will show it is possible to construct a Leggett-Garg Inequality which can be violated with operationally non-disturbing measurements. A PPS paradox is possible, and Alice can expect to win at the adversarial game, if and only if this Leggett-Garg Inequality is violated.

4.1. Macrorealism in the Three Box Paradox

In the context of the three box paradox, macrorealism is simply the claim that the ball is, at any time, in one box, and only in one box. There may be a probability distribution over which box the ball is in, but this must be understood strictly as some form of epistemic uncertainty. In the language of ontic models, this is expressed as saying that any preparation $\mu(\lambda)$ for the Three Box system is of the form

$$\mu(\lambda) = p_1\nu_1(\lambda) + p_2\nu_2(\lambda) + p_3\nu_3(\lambda) \quad (18)$$

where $\nu_i(\lambda) > 0$ only for ontic states where the ball is certain to be found in Box i whenever it is looked for, and similarly for $\nu_2(\lambda)$ and Box 2 etc., and the supports of the ν_i are disjoint. Such a model is non-contextually outcome definite⁵ for measurements of the location of the ball: $\xi_M(i|\lambda) = 1$ for all M and all ontic states in the support of $\nu_i(\lambda)$.

4.2. Non-invasive measurability in the Three Box Paradox

Leggett and Garg justified non-invasive measurability through introducing the idea of a *null-result* measurement: that some measurements only interact with the system if a particular outcome occurs. Leggett and Garg's definition was only appropriate for two outcome measurements. For the three box paradox it must be generalised to two distinct conditions, each weaker than ontic non-invasiveness, but either will lead to the Leggett-Garg Inequality. In the adversarial game, Bob's Condition (II) is an attempt to ensure his measurement is of this kind.

- A measurement M is *positive-result* non-invasive for some outcome Q if, and only if, for all λ_1 for which $\xi_M(Q|\lambda_1) > 0$, $\gamma_M(\lambda_2|Q, \lambda_1) = \delta(\lambda_2 - \lambda_1)$.

This kind of measurement represents the idea that it is possible to determine if the ball is in Box 1 without disturbing the ball when it is actually found in Box 1. Bob's test where he connects a spring to the top of the boxes, and gently applies a lifting force to one box slightly greater than the weight of the box, but well below that of the box and ball together, would be expected to be positive-result non-invasive.

If $B1$ is such a measurement, then observing that the ball is in the Box 1 leads to the post-measurement preparation state $\nu_1(\lambda)$, so

$$P_{(B1, A)}(3_3|1_2) = \int d\lambda \nu_1(\lambda) \xi_A(3_3|\lambda) \quad (19)$$

and similarly for $B2$ and Box 2.

- A measurement M is *negative-result* non-invasive for some outcome Q if, and only if, for all $P \neq Q$ and λ_1 such that $\xi_M(P|\lambda_1) > 0$, $\gamma_M(\lambda_2|P, \lambda_1) = \delta(\lambda_2 - \lambda_1)$.

This represents the idea that it is possible to determine if the ball is in Box 1 without disturbing the ball when it is *not* actually found in Box 1. Bob's test where he gently reaches

⁵It does not immediately run into problems with the Kochen Specker theorem, however, as it only requires this to hold for a single basis.

into an open box, without touching the sides, might be expected to be negative-result non-invasive, as Bob can verify the box is empty without disturbing anything.

If $B1$ is this kind of measurement, observing that the ball is not in Box 1, results in the post-measurement preparation state $\frac{p_2\nu_2(\lambda)+p_3\nu_3(\lambda)}{p_2+p_3}$, so

$$P_{(B1,A)}(3_3|-1_2) = \int d\lambda \frac{p_2\nu_2(\lambda) + p_3\nu_3(\lambda)}{p_2 + p_3} \xi_A(3_3|\lambda) \quad (20)$$

and similarly for $B2$ and Box 2.

When there are only two possible outcomes, positive- and negative-result non-invasiveness are equivalent conditions, and recover Leggett and Garg's original definition of a null-result measurement. When both conditions hold, this is equivalent to ontic non-invasiveness⁶. Either of these conditions individually lead to contradictions with the three box paradox.

4.3. The Leggett-Garg Inequality in the Three Box Paradox

The three box paradox can now be cast in terms of a Leggett-Garg Inequality violation. Assign the value $Q_i = -1$ when the ball is in Box 1 or Box 2, and $Q_i = +1$ when the ball is in Box 3. For the ball placed initially in Box 3, the possible sequences of values for $Q = Q_1Q_2 + Q_2Q_3 + Q_1Q_3$ are shown in Table 4.

Box	Q_1	Box	Q_2	Box	Q_3	Q
3_1	$+1$	1_2 or 2_2	-1	1_3 or 2_3	-1	-1
				3_3	$+1$	-1
		3_2	$+1$	1_3 or 2_3	-1	-1
				3_3	$+1$	$+3$

Table 4: Three Box Paradox as a Leggett-Garg Inequality

Any probability distribution over these possible outcomes will show the mean value of Q is bounded by $-1 \leq \langle Q \rangle \leq 3$.

Although it is not always possible to read off the value of Q_2 from the measurement outcomes available to Alice and Bob, the assumptions of macrorealism and non-invasive measurability allows the inequality to be put into a form which is operationally well defined. According to macrorealism

$$P_{(N,A)}(3_3) = \int d\lambda (p_1\nu_1(\lambda) + p_2\nu_2(\lambda) + p_3\nu_3(\lambda)) \xi_A(3_3|\lambda) \quad (21)$$

Even when Bob performs no measurement, there was a matter of fact as to which Box was occupied:

$$P_{(N,A)}(i_2, 3_3) = p_i \int d\lambda \nu_i(\lambda) \xi_A(3_3|\lambda) \quad (22)$$

Given the pre-selection $P(3_1) = 1$, the expression for $\langle Q \rangle$ then becomes:

$$\langle Q \rangle = 3P_{(N,A)}(3_2, 3_3) - (1 - P_{(N,A)}(3_2, 3_3)) \quad (23)$$

which will be more conveniently expressed as

$$\langle Q \rangle = 4(P_{(N,A)}(3_3) - P_{(N,A)}(1_2, 3_3) - P_{(N,A)}(2_2, 3_3)) - 1 \quad (24)$$

The probabilities actually recorded by Alice and Bob's measurements are

$(B1, A)$	$P_{(B1,A)}(1_2, 3_3)$ $P_{(B1,A)}(1_2, \neg 3_3)$	$P_{(B1,A)}(\neg 1_2, 3_3)$ $P_{(B1,A)}(\neg 1_2, \neg 3_3)$
$(B2, A)$	$P_{(B2,A)}(2_2, 3_3)$ $P_{(B2,A)}(2_2, \neg 3_3)$	$P_{(B2,A)}(\neg 2_2, 3_3)$ $P_{(B2,A)}(\neg 2_2, \neg 3_3)$
(N, A)	$P_{(N,A)}(3_3)$	

⁶If there are two operationally equivalent measurements procedures available, one of which is positive-result non-invasive and the other negative-result non-invasive, these may be used together in a protocol for an ontically non-invasive measurement. The combined protocol involves randomly choosing which measurement procedure on each run of the experiment, and only keeping results from the runs where the non-invasive outcomes occur.

If Bob's measurements are assumed to be either positive- or negative-result non-invasive (or both), it becomes possible to calculate a value for $\langle Q \rangle$ from the observed data, to which a macrorealist would be committed.

If positive-result non-invasiveness is assumed to hold for Bi ,

$$P_{(Bi,A)}(i_2, 3_3) = p_i \int d\lambda \nu_i(\lambda) \xi_A(3_3|\lambda) = P_{(N,A)}(i_2, 3_3) \quad (25)$$

so

$$\langle Q \rangle_+ = 4 (P_{(N,A)}(3_3) - P_{(B1,A)}(1_2, 3_3) - P_{(B2,A)}(2_2, 3_3)) - 1 \quad (26)$$

Negative-result non-invasiveness implies

$$\begin{aligned} P_{(Bj,A)}(3_3) - P_{(Bj,A)}(j_2, 3_3) &= P_{(Bj,A)}(\neg j_2, 3_3) \\ &= \sum_{i \neq j} p_i \int d\lambda \nu_i(\lambda) \xi_A(3_3|\lambda) \\ &= \sum_{i \neq j} P_{(N,A)}(i_2, 3_3) = P_{(N,A)}(3_3) - P_{(N,A)}(j_2, 3_3) \end{aligned} \quad (27)$$

so

$$\langle Q \rangle_- = 4 (P_{(B1,A)}(3_3) - P_{(B1,A)}(1_2, 3_3) + P_{(B2,A)}(3_3) - P_{(B2,A)}(2_2, 3_3) - P_{(N,A)}(3_3)) - 1 \quad (28)$$

Given the operational non-disturbance requirement $P_{(B1,A)}(3_3) = P_{(B2,A)}(3_3) = P_{(N,A)}(3_3)$

$$\langle Q \rangle_+ = \langle Q \rangle_- = 4P_{(N,A)}(3_3) (1 - P_{(B1,A)}(1_2|3_3) - P_{(B2,A)}(2_2|3_3)) - 1 \quad (29)$$

Anyone believing in both macrorealism, and either the positive- or negative-result non-invasiveness of Bob's measurements, will be committed to believing this operationally well-defined value will satisfy the Leggett-Garg Inequality. Violation of the inequality occurs if, and only if,

$$P_{(B1,A)}(1_2|3_3) + P_{(B2,A)}(2_2|3_3) > 1 \quad (30)$$

Once again this is the condition for which Alice might offer reasonable seeming odds to Bob, and yet still be sure of winning on average. Alice can expect to win her adversarial game if, and only if, this Leggett-Garg inequality is violated.

The quantum three box paradox gives $P_{(B1,A)}(1_2|3_3) = P_{(B2,A)}(2_2|3_3) = 1$ and $P_{(N,A)}(3_3) = 1/9$, so

$$\langle Q \rangle = -\frac{13}{9} < -1$$

Experimental realisations of the three box paradox[4] have demonstrated this violation of the Leggett-Garg Inequality by over 7 standard deviations.

4.4. Operational eigenstates and the Three Box Paradox

Macrorealism, the idea that at all times the ball is in one, and only in one of the boxes, combined with either positive- or negative-result non-invasiveness cannot reproduce the three box paradox. Is macrorealism *per se* compatible with the paradox? While Bob's Condition (III) seems to suggest it is not, I will now show the paradox is only incompatible with a particular, if rather natural, kind of macrorealism.

Condition (III) involves Bob placing the ball in one of the boxes, performing either the $B1$ or $B2$ measurement, and then seeing if he can detect any observable change as a result. Suppose Bob chose to place the ball in Box i , so that the preparation is $\mu_i(\lambda)$, and then Bob performs the test Bj . The post-measurement preparation is

$$\mu_i^{(j)}(\lambda) = \int d\lambda_1 \mu_i(\lambda_1) \gamma_{Bj}(\lambda|k, \lambda_1) \quad (31)$$

where $k = j$ if $i = j$, but $k = \neg j$ if $i \neq j$. If Condition (III) is satisfied, then any subsequent measurement $\xi_M(q|\lambda)$ is unable to distinguish between $\mu_i(\lambda)$ and $\mu_i^{(j)}(\lambda)$, so

$$\int d\lambda \mu_i^{(j)}(\lambda) \xi_M(q|\lambda) = \int d\lambda \mu_i(\lambda) \xi_M(q|\lambda). \quad (32)$$

If the macrorealist is committed to the idea that any preparation $\mu(\lambda)$ is of the form

$$\mu(\lambda) = p_1\mu_1(\lambda) + p_2\mu_2(\lambda) + p_3\mu_3(\lambda) \quad (33)$$

then

$$\begin{aligned} P_{(Bi,A)}(i_2, 3_3) &= p_i \int d\lambda \mu_i^{(i)}(\lambda) \xi_A(3_3|\lambda) \\ &= p_i \int d\lambda \mu_i(\lambda) \xi_A(3_3|\lambda) \\ &= P_{(N,A)}(i_2, 3_3). \end{aligned} \quad (34)$$

This is the same as Equation 25, so the Leggett-Garg Inequality for Equation 29 follows as for positive-result non-invasiveness.

There is a crucial, if subtle, difference between Equation 18, to which a macrorealist is committed, and Equation 33, which leads to the Leggett-Garg Inequality. This difference was explored by [21], where three different forms of macrorealism were identified, only one of which is committed to Equation 33.

To understand the differences, it will be helpful to introduce the idea of an *operational eigenstate*. Suppose there is a set of measurements $\{M_\alpha\}$, each measurement has a particular outcome q_α , and these outcomes are operationally indistinguishable for any preparation $\mu(\lambda)$, so that:

$$\forall \alpha, \beta \quad \int d\lambda \mu(\lambda) \xi_{M_\alpha}(q_\alpha|\lambda) = \int d\lambda \mu(\lambda) \xi_{M_\beta}(q_\beta|\lambda). \quad (35)$$

The set $\tilde{q} = \{q_\alpha\}$ forms an equivalence class of outcomes, and may be thought of as measuring the value \tilde{q} of some property Q . A preparation $\mu_{\tilde{q}}(\lambda)$ is an operational eigenstate of Q if, and only if

$$\int d\lambda \mu_{\tilde{q}}(\lambda) \xi_{M_\alpha}(q_\alpha|\lambda) = 1. \quad (36)$$

Operational eigenstates of a property are defined as those preparations which determinately fix the value of the property. In the case of the three box Paradox, therefore, the operational eigenstates for the location of the ball are the preparations that involve determinately placing a ball in one of the boxes. Bob's Condition (III) states that his measurements of the location of the ball should be operationally non-disturbing for operational eigenstates of the location of the ball.

Three forms of macrorealism can now be identified:

- *Operational eigenstate mixture macrorealism* holds that any possible preparation $\mu(\lambda)$ of the system is represented by a statistical mixture of operational eigenstates of Q and so is of the form:

$$\mu(\lambda) = \sum_i p_i \mu_{\tilde{i}}(\lambda) \quad (37)$$

For the three box paradox, this is Equation 33. If Condition (III) holds, this cannot violate the Leggett-Garg inequality, Equation 29, and is incompatible with the statistics of Table 1.

- *Operational eigenstate support macrorealism* holds that the set of all ontic states in the support of preparation states, is identical to the set of all ontic states in the supports of operational eigenstates of Q . For any λ for which there is a $\mu(\lambda) > 0$, there must exist some operational eigenstate for which $\mu_{\tilde{i}}(\lambda) > 0$.

However, unlike operational eigenstate mixture macrorealism, while in general a preparation takes the form

$$\mu(\lambda) = \sum_i p_i \nu_i(\lambda)$$

there is no requirement that $\nu_i(\lambda)$ is also an operational eigenstate of \tilde{i} .

Condition (III) establishes that operational eigenstates could be measured in an operationally non-disturbing way, and that Equations 31 and 32 hold. But if $\nu_i(\lambda)$ is *not* an operational eigenstate, this provides no guarantee that

$$\int d\lambda_1 \nu_i(\lambda_1) \gamma_{Bj}(\lambda|k, \lambda_1) \quad (38)$$

is operationally indistinguishable from $\nu_i(\lambda)$.

Operational non-disturbance is not the same as ontic non-invasiveness. In the case of operational eigenstates, there is a disturbance by $\gamma_{Bj}(\lambda|k, \lambda_1)$, but the $\mu_i(\lambda)$ are affected in a manner analogous to equilibrium distributions. Condition (III) simply doesn't check for the different effect of $\gamma_{Bj}(\lambda|k, \lambda_1)$ on $\nu_i(\lambda)$, so the Leggett-Garg Inequality can be violated and the three box paradox can be reproduced. Appendix A.5 provides a constructive example.

- *Supra eigenstate support macrorealism* is the final type of macrorealism available. In common to all brands of macrorealism, this view holds that

$$\mu(\lambda) = \sum_i p_i \nu_i(\lambda)$$

and that $\xi_M(i|\lambda) = 1$ whenever $\nu_i(\lambda) > 0$. However, it allows that when $\nu_i(\lambda)$ is not an operational eigenstate, there may be λ such that $\nu_i(\lambda) > 0$ but which do not lie in the support of any operational eigenstate. These are novel ontic states which can only arise in preparations which have non-zero probabilities for at least two values for Q . Now Condition (III)'s failure to account for these is more direct: Bob's operational eigenstate preparations simply couldn't include these novel ontic states. The de Broglie-Bohm theory[22, 23] is a general ontological model for quantum theory which is of this kind.

In the case of the three box paradox, when a quantum state is prepared that involves a superposition of the ball being in one or another box, supra eigenstate support macrorealism holds that it is still the case that the ball really is determinately in one or the other box, and the probability of which is an expression of epistemic uncertainty. However, the ball may be in an ontic state that is not accessible by simply placing the ball in the box, and it may therefore behave differently to an operational eigenstate preparation. Appendix A.6 gives the smallest constructive example of such a model.

5. PPS Paradoxes and non-classicality

It is now possible to provide clear answers to the questions: Why would Bob have considered it reasonable to play the adversarial game? and What exactly is non-classical about the three box paradox?

The reasoning behind Bob's four conditions are:

- Condition (I) The intervening measurement must be operationally non-disturbing for Alice. In the context of the adversarial game, if Alice's final measurement conveys information about which box Bob opened, Alice can improve her odds of winning. In the context of considering PPS paradoxes in general, any attempt to combine inferences from different measurement contexts is simply inconsistent if operational non-disturbance does not hold.
- Condition (II) If Bob is convinced that his measurements cannot introduce any disturbance to the system, then he would believe ontic non-invasiveness holds, and would be willing to play the game. If he accepts any form of macrorealism, and simultaneously believes he can perform a measurement that is either positive- or negative-result non-invasive, then he would be willing to play the game.
- Condition (III) Bob can verify that his measurements are operationally non-disturbing for operational eigenstate preparations. If he believes in operational eigenstate mixture macrorealism, he would be willing to play the game.
- Condition (IV) Bob can verify that $P_{(B2,B1)}(2_2, 1_2) = P_{(B1,B2)}(1_2, 2_2) = 0$, so he is sure that Alice is not cheating by simply placing two balls in the boxes.

Condition (I) and Condition (IV) are straightforward: without them there is simply no reason to suppose there is anything either paradoxical or non-classical in the first place. If there is anything non-classical, if there is anything wrong with Bob's reasoning, it must be found in Condition (II) and Condition (III).

The idea that a given measurement might affect the system being measured is not in itself either paradoxical or non-classical. Even classically, the best one might hope for is to get arbitrarily close to non-disturbance for some very careful measurements. However, it is the type of disturbance required by the three box paradox that reveals the quantum nature of the situation. The failure of both positive- and negative-result non-invasiveness for any implementation of Bob’s measurement should seem surprising: a spring failing to gently lift a box containing a ball would not be expected to disturb the ball, nor would simply looking into a box, or gently reaching a hand into an empty box. Yet if the three box paradox is to hold, all such procedures must affect the quantum system, and must affect it in the same manner.

Whether one accepts macrorealism about the location of the ball or not, if Bob’s measurement is ontically non-invasive, the three box paradox cannot hold. Maintaining macrorealism is not as straightforward as simply dropping ontic non-invasiveness, however. An intuition lurking alongside the idea that the ball is always in one, and only in one, of the boxes, is that whenever the ball is in a given box, it behaves exactly as it appears to behave when it is *observed* to be in that box. This runs into difficulties, for when the ball’s location is observed, it is in an operational eigenstate. This rather natural idea of macrorealism would lead to operational eigenstate mixture macrorealism, and for that the three box paradox could not hold.

Operational eigenstate support macrorealism maintains part of the intuition, by only allowing ontic states that appear in the support of operational eigenstates. The unobserved ball’s ontic state is always one that can occur when the ball is being observed. However, the price is that those ontic states must now be behaving differently to their appearances. Neither positive- nor negative-result non-invasiveness will be possible, even for operational eigenstates. While the observed behaviour of the ball, determinately placed in one box while Bob checks Condition (III), is showing no detectable disturbance, something must nevertheless be undergoing change, below the level of appearances, as a result of Bob’s measurements. This change takes place even when Bob is only interacting with a different box: placing the ball in Box 1, then opening the empty Box 2, somehow disturbs the ball in Box 1 in an unobservable way. But when the system is prepared as in a quantum superposition, and the ball is not being directly observed, these same disturbances emerge and lead to observable consequences.

Supra eigenstate support macrorealism takes the opposite route. Operational eigenstates do not appear to be disturbed by Bob’s measurements, and it may be maintained that the ontic states in their support are not, in fact disturbed. However, when the ball is prepared through a quantum superposition, it may now be in an ontic state that does not appear in any operational eigenstate. When it is not being observed, the ball can behave differently.

Neither possibility corresponds to a *classical* model for a PPS paradox. Yet the fact that some form of macrorealism remains possible demonstrates that, at least for the measurements under consideration, non-contextual outcome definiteness is not ruled out by the three box paradox. Appendix A.5 and Appendix A.6 give constructive examples showing this. It is only the nature of the measurement disturbance that is a specifically quantum feature of the paradox.

6. Conclusion

Nothing paradoxical occurs when applying pre- and post-selection statistics to observed sequences of properties, whether classical or quantum. There is still nothing paradoxical when some of those properties are unobserved. But assuming the statistics should be the same in both cases assumes that the process of observation had no effect. There is an implicit counterfactual: the unobserved statistics would still have been recorded had the properties been observed. Paradoxical inferences in pre- and post-selection statistics only occur if they involve combinations from different choices of intervening measurements.

When the intervening measurements are operationally disturbing, this is simply inconsistent. Such situations can easily arise with classical measurement disturbances, and involve no paradox. Under such circumstances the most committed macrorealist should feel under no obligation to play an adversarial game, or to believe that a Leggett-Garg Inequality could not be violated. It is not sufficient to violate a Leggett-Garg Inequality, or to present an adversarial game with losing odds, to show a paradox: Bob must first have some reason to believe the game is fair.

With Bob’s conditions I have attempted to make precise the ideas on which such a belief could rest. If these conditions are not met, then classical toy models (Appendix A.1 to Appendix A.4)

are easy to produce which mimic the paradoxes, but there is simply no reason why Bob would have been willing to play one of these adversarial games in the first place.

When the conditions are met, the occurrence of the paradox should still be straightforwardly read as showing that the counterfactual is not valid. The conclusion of this paper is that the reason for the failure of the counterfactual, in these cases, is due to the intrinsically quantum nature of the measurement disturbance, and not some other quantum effect such as contextuality or a failure of macrorealism.

This paper has been concerned almost exclusively with a detailed analysis of the three box paradox, as the simplest, paradigm example of a quantum PPS paradox. To what extent can its conclusions be generalised to all quantum PPS paradoxes? It is straightforward that both ontic non-invasiveness and operational eigenstate mixture macrorealism must fail in such cases. Since this paper was first circulated, Allen[24, 25] has responded with a proof suggesting that operational eigenstate support macrorealism cannot hold in general in Hilbert spaces of dimension greater than 3. While the existence of de Broglie-Bohm theory proves that supra eigenstate support macrorealism is always tenable, Pusey and Leifer[26] have argued that non-contextuality may still be incompatible with PPS paradoxes in general.

There remains the problem of a general definition of when exactly a quantum PPS paradox can be said to occur. Leifer and Spekkens[8] clarified the algebraic conditions on the intermediate measurement outcomes. I have argued here that these must be supplemented, at the very least, by the requirement that the intermediate measurements be operationally non-disturbing to the statistics of the post-selection measurement, or else the combination of inferences is simply inconsistent. It is interesting to note that Bob's verification that $P(1_2 \wedge 2_2) = 0$ does not satisfy this stricter requirement, and without this verification it is trivial for Alice to cheat. The implications of adopting this stricter definition of a PPS paradox will be explored in a future paper.

Acknowledgements I would like to thank Chris Timpson, Richard George, Erik Gauger, Andrew Briggs, John-Mark Allen, Matt Pusey, Ronnie Hermens and the members of the Wolfson College Academic Writing group 'Ta da! The Teddy Bears (of Doom)'. This research has been supported in part by the John Templeton Foundation and the Templeton World Charity Foundation.

Appendix A. Ontic Models for PPS Games

In this Appendix I will rewrite the classical models for PPS paradoxes considered in [9, 10, 11] in the ontic model formalism, to show how all involve operationally disturbing measurements. Three other ontic models are considered which do not involve operationally disturbing measurements: a model which demonstrates the importance of Bob's Condition (IV) and constructive examples of the macrorealist options discussed in Section 4.4.

Appendix A.1. Kirkpatrick's Card Game

In Kirkpatrick's card game[9], the ontic state is represented by two piles of cards, {Active, Passive}, containing the following cards {Jack of Spades, Queen of Spades, Jack of Diamonds, Queen of Diamonds, King of Hearts, King of Hearts}. Five ontic states represent the different allowed combinations of cards in the two piles:

λ_a : Face=Queen. Active={QS,QD}, Passive={JS,JD,2KH}

λ_b : Suit=Spades. Active={JS,QS}, Passive={JD,QD,2KH}

λ_c : Suit=not-Spades. Active={JD,QD,2KH}, Passive={JS,QS}

λ_d : Suit=Diamond. Active={JD,QD}, Passive={JS,QS,2KH}

λ_e : Suit=not-Diamond. Active={JS,QS,2KH}, Passive={JD,QD}

Measurements involve picking a card at random from one of the piles, and asking if the face or the suit has a specific value (e.g. 'Is the Face a Queen?' or 'Is the Suit Spades?'). If the measurement changes from a Suit to Face question, or vice versa, the card is picked from the Passive pile, otherwise the card is picked from the Active pile. If the card was taken from the Active pile, then

after the result of the measurement the card is simply restored to its pile. If the card was taken from the Passive pile, a new ontic state is prepared according to the outcome of the measurement.

The outcomes and updates⁷ are given in the form (probability of outcome; post measurement preparation state):

	B1		B2		A	
	1 ₂	¬1 ₂	2 ₂	¬2 ₂	3 ₃	¬3 ₃
λ _a	(1/4; λ _b)	(3/4; λ _c)	(1/4; λ _d)	(3/4; λ _e)	0	1
λ _b	(1; λ _b)	0	0	(1; λ _b)	1/2	1/2
λ _c	0	(1; λ _c)	(1/2; λ _c)	(1/2; λ _c)	0	1
λ _d	0	(1; λ _d)	(1; λ _d)	0	1/2	1/2
λ _e	(1/2; λ _e)	(1/2; λ _e)	0	(1; λ _e)	0	1

For the three box Paradox, the initial pre-selection asks if the Face is Queen, preparing λ_a. Opening Box 1 is represented by asking if the Suit is Spade. Opening Box 2 is represented by asking if the Suit is Diamond. Alice's post-selection asks if the Face is the King. This gives the probabilities:

(B1, A)	$P_{(B1,A)}(1_2, 3_3) = 1/8$	$P_{(B1,A)}(\neg 1_2, 3_3) = 0$
	$P_{(B1,A)}(1_2, \neg 3_3) = 1/8$	$P_{(B1,A)}(\neg 1_2, \neg 3_3) = 3/4$
(B2, A)	$P_{(B2,A)}(2_2, 3_3) = 1/8$	$P_{(B2,A)}(\neg 2_2, 3_3) = 0$
	$P_{(B2,A)}(2_2, \neg 3_3) = 1/8$	$P_{(B2,A)}(\neg 2_2, \neg 3_3) = 3/4$
(N, A)	$P_{(N,A)}(3_3) = 0$	

While this successfully reproduces the result $P_{(B1,A)}(1_2|3_3) = P_{(B2,A)}(2_2|3_3) = 1$, it fails to be a true PPS paradox as $P_{(B1,A)}(3_3) = P_{(B2,A)}(3_3) = 1/8$ but $P_{(N,A)}(3_3) = 0$.

Kirkpatrick's game is neither non-invasive (ontic state λ_a) nor macrorealist (ontic states λ_a, λ_c and λ_e) in the sense used in this paper. It is effectively the failure of macrorealism that Kirkpatrick argues accounts for the quantum properties. It is possible that more complex choices of ontic states could better reproduce the three box paradox statistics. Kirkpatrick[12] does suggest such a modification, in response to Ravon and Vaidman, so that $P_{(N,A)}(3_3) \neq 0$.

Appendix A.2. Ravon and Vaidman's Card Game

Ravon and Vaidman[11] present a simplified card game, based on Kirkpatrick's. The number of cards are reduced, by removing the queens and a king.

- λ_a: Face. Active={}, Passive={JS,JD,KH}
- λ_b: Suit=S. Active={JS}, Passive={JD,KH}
- λ_c: Suit=not-S. Active={JD,KH}, Passive={JS}
- λ_d: Suit=D. Active={JD}, Passive={JS,KH}
- λ_e: Suit=not-D. Active={JS,KH}, Passive={JD}

The outcomes and updates⁸ are:

	B1		B2		A	
	1 ₂	¬1 ₂	2 ₂	¬2 ₂	3 ₃	¬3 ₃
λ _a	(1/3; λ _b)	(2/3; λ _c)	(1/3; λ _d)	(2/3; λ _e)	0	1
λ _b	(1; λ _b)	0	0	(1; λ _b)	1/2	1/2
λ _c	0	(1; λ _c)	(1/2; λ _c)	(1/2; λ _c)	0	1
λ _d	0	(1; λ _d)	(1; λ _d)	0	1/2	1/2
λ _e	(1/2; λ _e)	(1/2; λ _e)	0	(1; λ _e)	0	1

⁷The updates for measuring B1 on λ_e and B2 on λ_c follow the rules exactly as stated by Kirkpatrick. However, it does not seem these will produce repeatable measurement outcomes, which would require an update to λ_b or λ_c for B1 and λ_d or λ_e for B2, depending on the outcome.

⁸Ravon and Vaidman only specify the updates for particular measurements, instead of providing a complete set of rules. Assuming that the rest follow the structure of Kirkpatrick's game gives the rules stated. This assumption does not impact on the measurement statistics for the actual sequences considered. One modification of these rules allows $P_N(A) = 1/3$, but still would not create a true PPS paradox.

the initial preparation state is λ_a , giving:

$(B1, A)$	$P_{(B1,A)}(1_2, 3_3) = 1/6$	$P_{(B1,A)}(\neg 1_2, 3_3) = 0$
	$P_{(B1,A)}(1_2, \neg 3_3) = 1/6$	$P_{(B1,A)}(\neg 1_2, \neg 3_3) = 2/3$
$(B2, A)$	$P_{(B2,A)}(2_2, 3_3) = 1/6$	$P_{(B2,A)}(\neg 2_2, 3_3) = 0$
	$P_{(B2,A)}(2_2, \neg 3_3) = 1/6$	$P_{(B2,A)}(\neg 2_2, \neg 3_3) = 2/3$
(N, A)	$P_{(N,A)}(3_3) = 0$	

Again, the PPS paradox occurs $P_{(B1,A)}(1_2|3_3) = P_{(B2,A)}(2_2|3_3) = 1$, but it fails to be a true PPS paradox as $P_{(B1,A)}(3_3) = P_{(B2,A)}(3_3) = 1/6$ but $P_{(N,A)}(3_3) = 0$.

Appendix A.3. Leifer and Spekkens's Ball Game

Leifer and Spekkens[10] consider a ball within a square box. The ball may be in one of four positions: top left; top right; bottom left; and bottom right. The box may be divided into two compartments: either top-bottom or left-right. The location of the ball may only be measured by dividing the box into two compartments and shaking one of the compartments. A rattling sound indicates the ball is present but disturbs it. No rattle indicates the ball is in the other compartment but does not disturb it. They consider preparing the ball to be in the bottom. Then a measurement is made of either the left ($B1$) or right ($B2$) compartment, finally followed by a post-selection on a successful top measurement (A).

λ_a : Bottom left

λ_b : Bottom right

λ_c : Top left

λ_d : Top right

The outcomes and updates are:

	$B1$		$B2$		A	
	1_2	$\neg 1_2$	2_2	$\neg 2_2$	3_3	$\neg 3_3$
λ_a	$(1; 1/2(\delta_{(\lambda, \lambda_a)} + \delta_{(\lambda, \lambda_c)}))$	0	0	$(1; \lambda_a)$	0	$(1; \lambda_a)$
λ_b	0	$(1; \lambda_b)$	$(1; 1/2(\delta_{(\lambda, \lambda_b)} + \delta_{(\lambda, \lambda_d)}))$	0	0	$(1; \lambda_b)$
λ_c	$(1; 1/2(\delta_{(\lambda, \lambda_a)} + \delta_{(\lambda, \lambda_c)}))$	0	0	$(1; \lambda_c)$	$(1; 1/2(\delta_{(\lambda, \lambda_c)} + \delta_{(\lambda, \lambda_d)}))$	0
λ_d	0	$(1; \lambda_d)$	$(1; 1/2(\delta_{(\lambda, \lambda_b)} + \delta_{(\lambda, \lambda_d)}))$	0	$(1; 1/2(\delta_{(\lambda, \lambda_c)} + \delta_{(\lambda, \lambda_d)}))$	0

The initial preparation state of the system is $1/2(\delta_{(\lambda, \lambda_a)} + \delta_{(\lambda, \lambda_b)})$. The probabilities are now:

$(B1, A)$	$P_{(B1,A)}(1_2, 3_3) = 1/4$	$P_{(B1,A)}(\neg 1_2, 3_3) = 0$
	$P_{(B1,A)}(1_2, \neg 3_3) = 1/4$	$P_{(B1,A)}(\neg 1_2, \neg 3_3) = 1/2$
$(B2, A)$	$P_{(B2,A)}(2_2, 3_3) = 1/4$	$P_{(B2,A)}(\neg 2_2, 3_3) = 0$
	$P_{(B2,A)}(2_2, \neg 3_3) = 1/4$	$P_{(B2,A)}(\neg 2_2, \neg 3_3) = 1/2$
(N, A)	$P_{(N,A)}(3_3) = 0$	

The PPS paradox occurs as $P_{(B1,A)}(1_2|3_3) = P_{(B2,A)}(2_2|3_3) = 1$, but it fails to be a true PPS paradox as $P_{(B1,A)}(3_3) = P_{(B2,A)}(3_3) = 1/4$ but $P_{(N,A)}(3_3) = 0$. Unlike Kirkpatrick's model, this satisfies both operational eigenstate mixture macrorealism, and negative-result non-invasiveness. It follows that no possible modification of this classical model could simulate the three box paradox or violate the Leggett-Garg Inequality. Furthermore, the protocol detailed in Footnote 6 can easily create an optically non-invasive measurement for this model, for which $P_{(B1,A)}(1_2|3_3) = P_{(B2,A)}(2_2|3_3) = 0$.

Appendix A.4. Alice's Easy Cheat

The simplest way to reproduce the statistics of Table 1 is for Alice to cheat. Alice places a single ball in 8 cases out of 9, and then ensures it is shuffled away from Box 3 before the final measurement. The remainder of the times she places a ball in Box 1 and a ball in Box 2. She knows in these cases Bob must see a ball, and shuffles one of them to Box 3 for her final measurement.

The required ontic states and their measurement responses are:

	<i>B1</i>		<i>B2</i>		<i>A</i>	
	1 ₂	¬1 ₂	2 ₂	¬2 ₂	3 ₃	¬3 ₃
λ _a	0	1	0	1	1	0
λ _b	0	1	1	0	0	1
λ _c	1	0	0	1	0	1
λ _d	1	0	1	0	0	1
λ _e	1	0	0	1	1	0
λ _f	0	1	1	0	1	0

All measurements are ontically non-invasive. Alice's shuffle can be represented by the transition probabilities $\gamma_C(\lambda_3|\lambda_2)$:

$$\begin{aligned}
\gamma_C(\lambda|\lambda_a) &= \frac{1}{2} (\delta_{(\lambda,\lambda_b)} + \delta_{(\lambda,\lambda_c)}) \\
\gamma_C(\lambda|\lambda_b) &= \delta_{(\lambda,\lambda_b)} \\
\gamma_C(\lambda|\lambda_c) &= \delta_{(\lambda,\lambda_c)} \\
\gamma_C(\lambda|\lambda_d) &= \frac{1}{2} (\delta_{(\lambda,\lambda_e)} + \delta_{(\lambda,\lambda_f)})
\end{aligned} \tag{A.1}$$

With the initial preparation:

$$\mu_C(\lambda) = \frac{1}{9} (4\delta_{(\lambda,\lambda_a)} + 2\delta_{(\lambda,\lambda_b)} + 2\delta_{(\lambda,\lambda_c)} + \delta_{(\lambda,\lambda_d)}) \tag{A.2}$$

this model successfully reproduces Table 1. However, it fails to reproduce Table 2. Bob's Condition (IV) is an essential test if there is supposed to be something puzzling about the three box paradox.

Appendix A.5. Operational eigenstate support macrorealism

The following ontic model is non-contextually value definite for all the measurements under consideration, all of the ontic states appear in the support of operational eigenstates, and it reproduces exactly Tables 1 and 2. There are 16 ontic states λ_a to λ_p . Q_2 gives the location of the ball at the time of Bob's measurement. $B1$ and $B2$ give the change in the ontic state as a result of Bob's measurement. Q_3 gives the effect of the shuffle, prior to Alice's opening Box 3.

	λ _a	λ _b	λ _c	λ _d	λ _e	λ _f	λ _g	λ _h	λ _i	λ _j	λ _k	λ _l	λ _m	λ _n	λ _o	λ _p
<i>Q</i> ₂	1 ₂ , ¬2 ₂				¬1 ₂ , 2 ₂				¬1 ₂ , ¬2 ₂							
<i>B1</i>	λ _a	λ _b	λ _a	λ _b	λ _e	λ _f	λ _g	λ _h	λ _i	λ _j	λ _i	λ _j	λ _m	λ _n	λ _m	λ _n
<i>B2</i>	λ _a	λ _b	λ _c	λ _d	λ _e	λ _f	λ _e	λ _f	λ _i	λ _j	λ _k	λ _l	λ _i	λ _j	λ _k	λ _l
<i>Q</i> ₃	¬3 ₃	3 ₃	3 ₃	¬3 ₃	¬3 ₃	3 ₃	3 ₃	¬3 ₃	¬3 ₃	3 ₃	3 ₃	¬3 ₃	3 ₃	¬3 ₃	¬3 ₃	3 ₃

The preparations

$$\begin{aligned}
\mu_{1,\alpha}(\lambda) &= \frac{2}{3}\delta_{(\lambda,\lambda_a)} + \frac{1}{3}\delta_{(\lambda,\lambda_b)} \\
\mu_{1,\beta}(\lambda) &= \frac{1}{3} (\delta_{(\lambda,\lambda_a)} + \delta_{(\lambda,\lambda_c)} + \delta_{(\lambda,\lambda_d)}) \\
\mu_{2,\alpha}(\lambda) &= \frac{2}{3}\delta_{(\lambda,\lambda_e)} + \frac{1}{3}\delta_{(\lambda,\lambda_f)} \\
\mu_{2,\beta}(\lambda) &= \frac{1}{3} (\delta_{(\lambda,\lambda_e)} + \delta_{(\lambda,\lambda_g)} + \delta_{(\lambda,\lambda_h)}) \\
\mu_{3,\alpha}(\lambda) &= \frac{2}{3}\delta_{(\lambda,\lambda_i)} + \frac{1}{3}\delta_{(\lambda,\lambda_j)} \\
\mu_{3,\beta}(\lambda) &= \frac{1}{3} (\delta_{(\lambda,\lambda_i)} + \delta_{(\lambda,\lambda_k)} + \delta_{(\lambda,\lambda_l)}) \\
\mu_{3,\gamma}(\lambda) &= \frac{1}{3} (\delta_{(\lambda,\lambda_i)} + \delta_{(\lambda,\lambda_m)} + \delta_{(\lambda,\lambda_n)}) \\
\mu_{3,\delta}(\lambda) &= \frac{1}{3} (\delta_{(\lambda,\lambda_i)} + \delta_{(\lambda,\lambda_o)} + \delta_{(\lambda,\lambda_p)})
\end{aligned} \tag{A.3}$$

form convex mixtures to give operational eigenstate preparations

$$\begin{aligned}
\mu_{|1\rangle}(\lambda) &= (1 - p_1)\mu_{1,\alpha}(\lambda) + p_1\mu_{1,\beta}(\lambda) \\
\mu_{|2\rangle}(\lambda) &= (1 - p_2)\mu_{2,\alpha}(\lambda) + p_2\mu_{2,\beta}(\lambda) \\
\mu_{|3\rangle}(\lambda) &= (1 - p_\beta - p_\gamma - p_\delta)\mu_{3,\alpha}(\lambda) + p_\beta\mu_{3,\beta}(\lambda) + p_\gamma\mu_{3,\gamma}(\lambda) + p_\delta\mu_{3,\delta}(\lambda)
\end{aligned} \tag{A.4}$$

which include all ontic states in their support. The remaining quantum state preparations are:

$$\begin{aligned}
\mu_{|1+2+3\rangle}(\lambda) &= \frac{1}{9}(2\delta_{(\lambda,\lambda_a)} + \delta_{(\lambda,\lambda_d)} + 2\delta_{(\lambda,\lambda_e)} + \delta_{(\lambda,\lambda_h)} + 2\delta_{(\lambda,\lambda_i)} + \delta_{(\lambda,\lambda_p)}) \\
\mu_{|1+3\rangle}(\lambda) &= \frac{1}{6}(2\delta_{(\lambda,\lambda_a)} + \delta_{(\lambda,\lambda_d)} + 2\delta_{(\lambda,\lambda_i)} + \delta_{(\lambda,\lambda_l)}) \\
\mu_{|2+3\rangle}(\lambda) &= \frac{1}{6}(2\delta_{(\lambda,\lambda_e)} + \delta_{(\lambda,\lambda_h)} + 2\delta_{(\lambda,\lambda_i)} + \delta_{(\lambda,\lambda_n)})
\end{aligned} \tag{A.5}$$

This ontic model reproduces Tables 1 and 2. It therefore reproduces all of the relevant statistics for the three box paradox.

The ontic states $\lambda_c, \lambda_d, \lambda_g, \lambda_h, \lambda_k - \lambda_p$ contain structure that allows the system to change state when Bob performs one of his tests. This disturbance is a necessary feature of any ontic model that hopes to reproduce a PPS paradox. It should be noted that while Bob's measurement can change the state of the ball, it does not cause the ball to change boxes. The ball is always in one, and only one, of the boxes.

Appendix A.6. Supra eigenstate support macrorealism

The operational eigenstate support macrorealist model in Appendix A.5 contains redundant states in its definition of the operational eigenstates. The only operational eigenstate preparations that are needed to show the three box paradox are $\mu_{1,\alpha}(\lambda)$, $\mu_{2,\alpha}(\lambda)$ and $\mu_{3,\alpha}(\lambda)$. This reduced model is a supra eigenstate support macrorealist theory: the ontic states $\lambda_d, \lambda_h, \lambda_l, \lambda_n, \lambda_p$ appear in the superposition preparations but do not appear in the reduced model operational eigenstate preparations.

- [1] Y. Aharonov and L. Vaidman, *Journal of Physics A* **24**, 2315 (1991).
- [2] K. J. Resch, J. S. Lundeen, and A. M. Steinberg, *Physics Letters A* **324**, 125 (2004).
- [3] P. Kolenderski, U. Sinha, L. Youning, T. Zhao, M. Volpini, A. Cabello, R. Laflamme, and T. Jennewein, *ArXiv e-print service* (2011), [arXiv.org://quant-ph/1107.5828](https://arxiv.org/abs/1107.5828).
- [4] R. E. George, L. Robledo, O. J. E. Maroney, M. S. Blok, H. Bernien, M. L. Markham, D. J. Twitchen, J. J. L. Morton, G. A. D. Briggs, and R. Hanson, *Proceedings of the National Academy of Sciences* **110**, 3777 (2013), [arXiv.org://1205.2594](https://arxiv.org/abs/1205.2594).
- [5] W. D. Sharp and N. Shanks, *Philosophy of Science* **60**, 488 (1993).
- [6] O. Cohen, *Physical Review A* **51**, 4373 (1995).
- [7] R. E. Kastner, *Philosophy of Science* **70**, 145 (2003).
- [8] M. S. Leifer and R. W. Spekkens, *Physical Review Letters* **95**, 200405 (2005).
- [9] K. A. Kirkpatrick, *Journal of Physics A* **36**, 4891 (2003).
- [10] M. S. Leifer and R. W. Spekkens, *International Journal of Theoretical Physics* **44**, 1977 (2005), [arxiv.org:quant-ph/0412179](https://arxiv.org/abs/quant-ph/0412179).
- [11] T. Ravon and L. Vaidman, *Journal of Physics A* **40**, 2873 (2007).
- [12] K. A. Kirkpatrick, *Journal of Physics A* **40**, 2883 (2007).
- [13] A. J. Leggett and A. Garg, *Physical Review Letters* **54**, 857 (1985).
- [14] R. W. Spekkens, *Physical Review A* **71**, 052108 (2005).

- [15] Y. Aharonov, P. G. Bergmann, and J. L. Lebowitz, *Physical Review B* **134**, 1410 (1964).
- [16] N. Aharon and L. Vaidman, *Physical Review A* **77**, 052310 (2008).
- [17] T. Rudolph, ArXiv e-print service (2006), [arxiv.org://quant-ph/0608.120](https://arxiv.org/abs/quant-ph/0608120).
- [18] N. Harrigan and R. W. Spekkens, *Foundations of Physics* **40**, 125 (2010), [arXiv:0706.266](https://arxiv.org/abs/0706.266).
- [19] N. Harrigan and T. Rudolph, ArXiv e-print service (2007), [arxiv.org://quant-ph/0709.4266](https://arxiv.org/abs/quant-ph/0709.4266).
- [20] J. S. Bell, *Epistemological Letters* (1976), reprinted in [27].
- [21] O. J. E. Maroney and C. G. Timpson, *British Journal for the Philosophy of Science* (**forthcoming**) (2017), [arXiv:1412.6139](https://arxiv.org/abs/1412.6139).
- [22] P. R. Holland, *The Quantum Theory of Motion* (Cambridge, 1993).
- [23] D. Bohm and B. J. Hiley, *The Undivided Universe* (Routledge, 1993).
- [24] J. M. A. Allen, *Quantum Studies: Mathematics and Foundations* **3**, 161 (2016).
- [25] J. M. A. Allen, O. J. E. Maroney, and S. Gogioso, ArXiv e-print service (2016), [arXiv:1610.00022](https://arxiv.org/abs/1610.00022).
- [26] M. L. Pusey and M. S. Leifer, ArXiv e-print service (2015), [arXiv:1506.07850](https://arxiv.org/abs/1506.07850).
- [27] J. S. Bell, *Speakable and unspeakable in quantum mechanics* (Cambridge University Press, 1987).