Supporting Information

George et al. 10.1073/pnas.1208374110

In this document, we detail the projections performed by Bob's measurements and provide a detailed account of the states prepared during the experiment. We detail how Alice is able to transform her initial state $|3\rangle$ into the state $|I\rangle$, and subsequently transform $|F\rangle$ back into state $|3\rangle$. We describe our notation for probabilities that allow us to describe the "three-box" game from both a classical and quantum perspective. We derive the Leggett–Garg function (1) for this system as calculated by an observer who assumes the system obeys the axioms of macrorealism (MR), namely, state definiteness and noninvasive measurability. We describe the sample fabrication and measurement setup and discuss the practicalities of the experimental measurements involving reading out the nuclear spin, and we discuss the significance of finite measurement precision and measurement errors.

I. Experimental Details

A. Projections Performed by Bob's Measurements. In the main text, we state that Bob finding a measurement result M_j-true prepares the state $|j\rangle$ by performing projector $\hat{P}_j$ on state $|I\rangle$, whereas Bob finding the result M_j-false prepares an orthogonal state $|\psi'_j\rangle$ by performing projector $\hat{P}_j^{\perp}$. Here, we give explicit vector representations of these states, and matrix representations of the projectors $\hat{P}_j$ and $\hat{P}_j^{\perp}$, to aid understanding. We can write a column vector to represent the general state $|\psi\rangle$ of the three-box problem as:

$$|\psi\rangle = a|1\rangle + b|2\rangle + c|3\rangle = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \quad \textbf{[S1]}$$

The initial and final states used by Alice are then written as:

$$|I\rangle = \frac{1}{\sqrt{3}}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \qquad \langle F| = \frac{1}{\sqrt{3}}\begin{pmatrix} 1 & 1 & -1 \end{pmatrix} \quad \textbf{[S2]}$$

We write the identity matrix as:

$$1 = \sum_j |j\rangle\langle j| = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \textbf{[S3]}$$

and we write the projectors $\hat{P}_j$ and $\hat{P}_j^{\perp}$ explicitly as:

$$\hat{P}_1 = |1\rangle\langle 1| = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\hat{P}_1^{\perp} = 1 - \hat{P}_1 = |2\rangle\langle 2| + |3\rangle\langle 3| = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \textbf{[S4]}$$

$$\hat{P}_2 = |2\rangle\langle 2| = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\hat{P}_2^{\perp} = 1 - \hat{P}_2 = |1\rangle\langle 1| + |3\rangle\langle 3| = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \textbf{[S5]}$$

Using this representation, it is straightforward to verify the claims in the main text, that:

$$P_{M_1}(A \cap B) = |\langle F|\hat{P}_1|I\rangle|^2 = P_{M_2}(A \cap B) = |\langle F|\hat{P}_2|I\rangle|^2 = 1/9 \quad \textbf{[S6]}$$

$$P_{M_1}(A \cap \neg B) = |\langle F|\hat{P}_1^{\perp}|I\rangle|^2 = P_{M_2}(A \cap \neg B) = |\langle F|\hat{P}_2^{\perp}|I\rangle|^2 = 0 \quad \textbf{[S7]}$$

These expressions describe Alice's ability to win $\gg$50% of rounds in the quantum version of the game.

B. Alice's Unitary Operations. ***1. Preparing the initial state.*** Alice would like to measure $|I\rangle$ and $|F\rangle$ but only has access to M_3. She performs effective M_I and M_F measurements by performing unitaries that map $|I\rangle \rightarrow |3\rangle$ and $|F\rangle \rightarrow |3\rangle$, followed by M_3 measurement. We define the unitary operation applied by Alice to transform between the states $|3\rangle$ and $|I\rangle$ in terms of its ability to split a population initially prepared in level $|3\rangle$ into an equal superposition of the states $|1\rangle$, $|2\rangle$, and $|3\rangle$. We construct $\hat{U}_I$ by concatenating two unitaries that can be implemented as rf pulses. The first step in performing $\hat{U}_I$ represents a rotation through angle θ in the $\{|3\rangle, |2\rangle\}$ plane, and the second step represents a rotation through angle $90° = \pi/2$ in the $\{|3\rangle, |1\rangle\}$ plane.

The first rotation (θ in the $\{|3\rangle, |1\rangle\}$ plane) must transfer one-third of the population from state $|3\rangle$ to state $|1\rangle$, leaving two-thirds of the population in state $|3\rangle$. The subsequent rotation must split the population in level $|3\rangle$ equally between $|3\rangle$ and $|2\rangle$, producing an equal population of one-third in each of the three $|j\rangle$ states.

Considering the coherent rotation in the $\{|3\rangle, |1\rangle\}$ plane, a rotation through θ transfers a fraction $\sin^2(\theta/2)$ into state $|1\rangle$, while leaving a fraction $\cos^2(\theta/2)$ in state $|3\rangle$; thus, to place one-third of the population in state $|1\rangle$, we have:

$$\sin^2(\theta/2) = 1/3 \qquad \cos^2(\theta/2) = 2/3 \quad \textbf{[S8]}$$

implying that:

$$\sin(\theta/2) = \sqrt{1/3} \qquad \cos(\theta/2) = \sqrt{2/3} \quad \textbf{[S9]}$$

$$\tan(\theta/2) = \sqrt{1/2} \qquad \theta = 2\tan^{-1}\left(\sqrt{1/2}\right) \quad \textbf{[S10]}$$

with the result that:

$$\theta = 70.6°(= 1.23 \text{ radians}) \quad \textbf{[S11]}$$

Alice prepares state $|3\rangle$ and performs $\hat{U}_I$ as two rotations: $\theta = 70.6°$ in the $\{|3\rangle, |1\rangle\}$ plane and $\pi/2 = 90°$ in the $\{|3\rangle, |2\rangle\}$ plane.

2. Alice's measurement of $|F\rangle$. The states $|I\rangle$ and $|F\rangle$ are defined in Vaidman's paper (2) as:

$$|I\rangle = \frac{|1\rangle + |2\rangle + |3\rangle}{\sqrt{3}} \qquad |F\rangle = \frac{|1\rangle + |2\rangle - |3\rangle}{\sqrt{3}} \quad \textbf{[S12]}$$

Because quantum states are defined only up to an overall multiplicative scalar (states such as $|F\rangle$ are rays in the Hilbert space), we can choose to write:

$$|F\rangle = \frac{-|1\rangle - |2\rangle + |3\rangle}{\sqrt{3}} \quad \textbf{[S13]}$$

A rotation through 2π radians introduces a sign change, such that there are two combined rotations through 2π, first on the $\{|3\rangle, |1\rangle\}$ level and then on the $\{|3\rangle, |2\rangle\}$ level. We have:

$$\begin{pmatrix} |1'\rangle \\ |3'\rangle \\ |2'\rangle \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}\begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{\hat{U}_{IF}} \begin{pmatrix} |1\rangle \\ |3\rangle \\ |2\rangle \end{pmatrix} \quad \textbf{[S14]}$$

The two rotations, each through 2π, have the combined effect of flipping the signs of the states $|1\rangle$ and $|2\rangle$ relative to state $|3\rangle$; specifically, we have:

$$|F\rangle = \hat{U}_{IF}|I\rangle \quad \textbf{[S15]}$$

Therefore, by applying these two rotations, Alice can map $|F\rangle \rightarrow |I\rangle \rightarrow |3\rangle$ and measure M_3 as per the main text.

C. Experimental Implementation of Nuclear Spin Readout. ***1. Sample.*** We use a naturally occurring nitrogen vacancy (NV^-) center in high-purity (spin-bearing impurities controlled below 1 part per billion) type IIa diamond grown by chemical vapor deposition, with a $\langle 111 \rangle$ crystal orientation obtained by cleaving a $\langle 100 \rangle$ substrate. We optimize the photon collection efficiency through use of a solid immersion lens deterministically fabricated by focused ion beam milling (3) to focus light onto the selected NV^- center. Microwave and rf pulses for the spin manipulation are applied through a lithographically defined strip-line adjacent to the solid immersion lens (3).

2. Measurement setup. We use a home-built, low-temperature confocal microscope that has been described in detail by Robledo et al. (4). All experiments are performed at a sample temperature of $T = 8.7$ K. A small magnetic field ($B \approx 5$ G, oriented along the NV^- symmetry axis) is applied by means of four permanent magnets arranged around the cryostat.

3. General. In the course of this experiment, we use different variations of single-shot nuclear spin readout, adapted to our specific purpose. In general, nuclear spin readout is implemented according to the following protocol (4):

i) Optional: Electron spin initialization by optical pumping into $m_S = 0$ (excitation of A_1 transition) or $m_S = \pm 1$ (excitation of E_x transition)
ii) Map nuclear spin onto electron spin: Selective microwave (MW) excitation of the hyperfine transition representing the state to be probed (in general, effecting a π rotation)
iii) Readout of the electron spin: Resonant optical excitation on E_x transition (for maximum contrast, $t_{ro} \approx 15$– 25 μs)
iv) Optional: Restore the electron spin state by optical pumping (for deterministic preparation of $m_S = +1$ or $m_S = -1$; optical pumping into $m_S = 0$, followed by a MW π-pulse)

If readout of the electron spin yields a result different from its initial state, we conclude that the nuclear spin occupies the probed state. The readout can be repeated using different MW frequencies, allowing us to perform population tomography on the full electron-nuclear spin state. We now outline the readout variations used.

4. Nuclear spin initialization. Initialization of the ^{14}N nuclear spin into $m_I = 0$ represents the first measurement of the Leggett–Garg test Q_1. This first measurement is probabilistic; we choose parameters that maximize the preparation fidelity with respect to the postmeasurement state, accepting a reduced preparation success probability:

The electron spin is initialized in the $m_S = \pm 1$ manifold by optical pumping, implemented by a pulse of 200 μs in duration, resonant with the E_x transition (fidelity $F = 99.4\%$). The initialization fidelity is further increased to $F > 99.9\%$ by post-selecting only experimental runs in which no photon is detected during the last 50 μs of the optical pumping pulse (avoiding accidental repopulation of $m_S = 0$).

We then apply a MW π-pulse resonant with the transition $|m_S = -1, m_I = 0\rangle \rightarrow |m_S = 0, m_I = 0\rangle$ with a state selectivity of ≈98%, limited by the proximity of other hyperfine transitions.

We probe successful initialization into $m_I = 0$ by requiring >0 detected photons during E_x excitation. To maximize fidelity, we keep the readout duration short (200 ns).

During the electron spin readout, there is a finite chance of optically induced electron spin flips. To ensure that the electron occupies the $m_S = -1$ state, we first optically pump it into $m_S = 0$ and then apply a selective MW π-pulse resonant with $|m_S = 0, m_I = 0\rangle \rightarrow |m_S = -1, m_I = 0\rangle$.

After successful initialization, we estimate an overlap with $|m_S = -1, m_I = 0\rangle$ of >95%.

All runs of the three-box experiment use this initialization step.

5. Three-box game: Bob's readout. The second readout (Bob's readout) consists of a selective MW π-pulse, resonant with:

$$|m_S = -1, m_I = -1\rangle \rightarrow |m_S = 0, m_I = -1\rangle \quad (M_1) \quad \textbf{[S16]}$$

$$|m_S = -1, m_I = +1\rangle \rightarrow |m_S = 0, m_I = +1\rangle \quad (M_2) \quad \textbf{[S17]}$$

$$|m_S = -1, m_I = 0\rangle \rightarrow |m_S = 0, m_I = 0\rangle \quad (M_3) \quad \textbf{[S18]}$$

depending on Bob's choice of measurement. Subsequently, the electron spin state is probed by a $t_{ro} = 20$-μs pulse resonant with E_x. This readout gives a large contrast ($F = 96\%$), but if $m_S = 0$ is detected, many excitation cycles may have occurred, and due to optically induced spin flips, the electron spin may be left in an undefined state. As a remedy, conditional on obtaining an $m_S = 0$ readout result, we restore the spin into $m_S = -1$ by optical pumping into $m_S = 0$, followed by a selective MW π-pulse, $|m_S = 0, m_I = +1(-1)\rangle \rightarrow |m_S = -1, m_I = +1(-1)\rangle$ ($M_{1(2)}$). This procedure ensures the electron is found deterministically in $m_S = -1$ after the readout, leaving nuclear spin coherence unaffected.

6. Three-box game: Alice's readout. Although for the last readout (Alice's readout), we could, in principle, apply the same protocol as in Bob's readout, we decided to read out all three nuclear spin states (box states) for each measurement, also allowing us to identify the few cases in which we do not find the ball in any of the boxes (e.g., to determine the finite detection efficiency of the nuclear spin readout).

For each probed nuclear spin state, we repeat two readout iterations consisting of a selective MW π-pulse and a 20-μs E_x readout pulse. This is repeated for the three hyperfine lines corresponding to the $m_S = -1$ manifold (implementing M_1, M_2, and M_3). The first probed state found to emit a photon is identified as the readout result; if no photon is detected, we consequently assign no result. To avoid a readout bias due to the order of the probed states, we permute the order between measurements.

7. Probing the initial state $|I\rangle$. To test successful generation of state $|I\rangle = \frac{1}{\sqrt{3}}(|1\rangle + |2\rangle + |3\rangle)$ (data in Fig. 3*A*, *i*), we show data from Bob's readout of the three-box game, with 1,200 repetitions of measuring each M_1, M_2, and M_3.

8. Probing repeatability. Data shown in Fig. 3*A*, *ii* and *iii* and in Fig. 3 *B* and *C* are obtained by correlating two successive readouts, implemented as Bob and Alice's readouts in the three-box game. However, here, we omit the NMR manipulation between Bob and Alice's readouts, such that both readout instances probe in the same basis. For each choice of Bob's measurement (M_1, M_2,

or M_3) in the following readout, we probe all nuclear spin states within the $m_S = -1$ manifold in the same measurement run.

II. Analysis of the Leggett–Garg Inequality

A. Leggett–Garg Function Is Satisfied for MR Systems. The Leggett–Garg function $\langle K\rangle$ is defined in terms of three sequential measurements Q_1, Q_2, and Q_3, having eigenvalues ± 1, as:

$$\langle K\rangle = \langle Q_1Q_2\rangle + \langle Q_2Q_3\rangle + \langle Q_1Q_3\rangle$$

If the observables Q_j are classical values that satisfy the assumptions of MR [e.g., they have a definite value (state definiteness (SD)] and measurement of one value does not change the subsequent values [nondisturbing measurability (NDM)], there are then eight possible combinations of $Q_1 \ldots Q_3$. The Leggett–Garg function can be thought of as the sum of three parity checks on three classical bits, of which two parity checks, at most, can be odd. Enumerating the combinations of Q_j shows that $\langle K\rangle$ will lie in the range $-1 \le \langle K\rangle \le 3$ in every case, such that the inequality is satisfied (Table S2).

B. Quantum Systems Can Violate the Leggett–Garg Inequality. The Leggett–Garg inequality is violated for quantum systems, however, if coherence persists between the times that different Q_j values are measured. Typically, a violation is observed by evaluating each term in the Leggett–Garg sum during separate runs of an experiment, as follows:

$$\langle K\rangle = \langle Q_1Q_2\rangle|_{\text{runs excluding } Q_3} + \langle Q_2Q_3\rangle|_{\text{runs excluding } Q_1} + \langle Q_1Q_3\rangle|_{\text{runs excluding } Q_2}$$

A macrorealist does not object to neglecting to measure Q_2 on a run that evaluates $\langle Q_1Q_3\rangle$ because he assumes that measurements do not disturb the state of the system (i.e., NDM holds); however, in the quantum case, measuring Q_2 can change the expectation of $\langle Q_1Q_3\rangle$.

For a system with two states $|1\rangle$ and $|2\rangle$, we can define $Q_j = +1$ if the system is found in the state $|1\rangle$ and $Q_j = -1$ if the system is found in the state $|2\rangle$. A spin-½ electron is an example of a physical system possessing two states: We can take spin up as $|1\rangle$, spin down as $|2\rangle$, and $Q_j = \sigma_z$ where σ_z is a Pauli matrix. Suppose that we prepare the state $|1\rangle\rangle$at time t_1 and that, during the interval $\Delta t = t_1 \ldots t_2$, we allow the system to evolve according to a unitary $U(\Delta t)$ acting as:

$$U(\Delta t) = U = \exp\left(\frac{2\pi i}{3}\frac{\sigma_y}{2}\right)$$

Writing in matrix notation, we have:

$$|1\rangle = \begin{pmatrix}1\\0\end{pmatrix} \quad |2\rangle = \begin{pmatrix}0\\1\end{pmatrix} \quad Q = \begin{pmatrix}1 & 0\\0 & -1\end{pmatrix}$$

and for the time evolutions:

$$U(\Delta t) = U = \exp\left(\frac{2\pi i}{3}\begin{pmatrix}0 & -i/2\\ i/2 & 0\end{pmatrix}\right) = \frac{1}{2}\begin{pmatrix}1 & \sqrt{3}\\ -\sqrt{3} & 1\end{pmatrix}$$

$$U(2\Delta t) = UU = \exp\left(\frac{4\pi i}{3}\begin{pmatrix}0 & -i/2\\ i/2 & 0\end{pmatrix}\right) = \frac{1}{2}\begin{pmatrix}-1 & -\sqrt{3}\\ \sqrt{3} & -1\end{pmatrix}$$

The initial state at time t_1 is represented by the density matrix ρ_1:

$$\rho_1 = |1\rangle\langle 1| = \begin{pmatrix}1 & 0\\0 & 0\end{pmatrix}$$

The Leggett–Garg correlators $\langle Q_1Q_2\rangle$ and $\langle Q_2Q_3\rangle$ are evaluated as:

$$\langle Q_1Q_2\rangle = \langle Q(t_1)Q(t_2)\rangle = \mathrm{Tr}\ \left((\rho_1)(Q)(UQU^\dagger)\right)$$
$$\langle Q_2Q_3\rangle = \langle Q(t_2)Q(t_3)\rangle = \mathrm{Tr}\ \left((U\rho_1U^\dagger)(UQU^\dagger)(UUQU^\dagger U^\dagger)\right)$$
$$= \mathrm{Tr}\ \left((\rho_1)(Q)(UQU^\dagger)\right) = \langle Q_1Q_2\rangle$$

The correlator $\langle Q_1Q_3\rangle$ is evaluated as:

$$\langle Q_1Q_3\rangle = \langle Q(t_1)Q(t_3)\rangle = \mathrm{Tr}\ \left((\rho_1)(Q)(UUQU^\dagger U^\dagger)\right)$$

Explicit calculation then yields:

$$\langle Q_1Q_2\rangle = \mathrm{Tr}\left(\begin{pmatrix}1 & 0\\0 & 0\end{pmatrix}\begin{pmatrix}1 & 0\\0 & -1\end{pmatrix}\frac{1}{2}\begin{pmatrix}1 & -\sqrt{3}\\ \sqrt{3} & 1\end{pmatrix}\begin{pmatrix}1 & 0\\0 & -1\end{pmatrix} \times \frac{1}{2}\begin{pmatrix}1 & \sqrt{3}\\ -\sqrt{3} & 1\end{pmatrix}\right) = -\frac{1}{2}$$

$$\langle Q_1Q_3\rangle = \mathrm{Tr}\left(\begin{pmatrix}1 & 0\\0 & 0\end{pmatrix}\begin{pmatrix}1 & 0\\0 & -1\end{pmatrix}\frac{1}{2}\begin{pmatrix}-1 & -\sqrt{3}\\ \sqrt{3} & -1\end{pmatrix}\begin{pmatrix}1 & 0\\0 & -1\end{pmatrix} \times \frac{1}{2}\begin{pmatrix}-1 & \sqrt{3}\\ -\sqrt{3} & -1\end{pmatrix}\right) = -\frac{1}{2}$$

and the Leggett–Garg function evaluates as:

$$\langle K\rangle = \langle Q_1Q_2\rangle + \langle Q_2Q_3\rangle + \langle Q_1Q_3\rangle = -\frac{3}{2},$$

which lies outside the range −1 to +3, violating the inequality.

III. Detectable Disturbance Is a Prerequisite to Violating the Leggett–Garg Inequality in 2D Hilbert Spaces

We are interested to know which of the hypotheses underpinning MR (SD or NDM) fails to hold when the Leggett–Garg inequality is violated. In a pre- and postselected system, we can define "detectable disturbance" as the change in the marginal statistics of the pre- and postselection, compared between the cases when an intervening measurement is performed or not. We demonstrate that for a two-level quantum system, violating the Leggett–Garg inequality requires detectable disturbance. We therefore show that the failure of NDM can explain all Leggett–Garg inequality violations in a two-level quantum system.

A. Leggett–Garg Inequality Specific to a Two-Level Quantum System. In a two-level quantum system, the Leggett–Garg function $\langle K\rangle$ is defined in terms of three measurements, Q_1, Q_2, and Q_3, with eigenvalues ± 1, taken sequentially at times t_1, t_2, and t_3. To observe a violation of the Leggett–Garg inequality, the averages $\langle Q_iQ_j\rangle$ must be performed using measurements at times t_i and t_j only. We write a subscript on the angle brackets to indicate the measurement times, such that:

$$\langle K\rangle = \langle Q_1Q_2\rangle_{(t_1,t_2)} + \langle Q_2Q_3\rangle_{(t_2,t_3)} + \langle Q_1Q_3\rangle_{(t_1,t_3)}$$

The Leggett–Garg function is then understood as the sum of results taken from three different ensemble averages in the quantum case. The hypotheses of MR (SD and NDM) lead a macrorealist to assume that each of the averages will be drawn from the same ensemble whenever a system obeying MR is measured.

1. Measurements of Q_j in the two-level case. A general measurement on a 2D quantum system is represented by a Pauli operator. We can assume that the two-level quantum system is degenerate and has no internal dynamics between the measurement times. If this

is not the case, we simply absorb the dynamics during intervals $t_1 \ldots t_2$ and $t_2 \ldots t_3$ into the definitions of the Q_j measurements.

We then write Q_1, Q_2, and Q_3 as measurements along three directions $\hat{n}_1$, $\hat{n}_2$, and $\hat{n}_3$, such that $\hat{Q}_j = \hat{\sigma}\cdot\hat{n}_j$. For the purposes of evaluating the Leggett–Garg function, the important quantity is the inner product between the measurement directions. We define $\cos\theta_{ij} = \hat{n}_i\cdot\hat{n}_j$, and analysis shows that $\langle Q_iQ_j\rangle_{(t_i,t_j)} = \cos\theta_{ij}$. In terms of this, the Leggett–Garg function becomes:

$$\langle K\rangle = \cos\theta_{12} + \cos\theta_{23} + \cos\theta_{13} \qquad \textbf{[S19]}$$

In the quantum case, we can pick three directions for $\hat{n}_j$, such that $\theta_{12} = \theta_{23} = \theta_{13} = 120° = 2\pi/3$ radians. Because $\cos(2\pi/3) = -1/2$, this choice obtains $\langle K\rangle = -3/2$ when the quantum system is measured, violating the inequality.

2. Detectable disturbance during measurement. We define the detectable disturbance D as the difference in $\langle K\rangle$ induced by performing pairs of measurements, compared with performing all three measurements:

$$D = \langle Q_1Q_2\rangle_{(t_1,t_2)} - \langle Q_1Q_2\rangle_{(t_1,t_2,t_3)} + \langle Q_2Q_3\rangle_{(t_2,t_3)} - \langle Q_2Q_3\rangle_{(t_1,t_2,t_3)} + \langle Q_1Q_3\rangle_{(t_1,t_3)} - \langle Q_1Q_3\rangle_{(t_1,t_2,t_3)}$$

Analysis shows that the nonzero contributions to D arise from the $\langle Q_1Q_3\rangle$ terms; thus, D is a comparison between evaluating $\langle Q_1Q_3\rangle$ in an experiment in which the system is measured at times t_1 and t_3 only and an experiment in which the system is measured at each of the times t_1, t_2, and t_3, but with the Q_2 measurement result discarded. We have:

$$\begin{aligned} D = \langle Q_1Q_3\rangle_{(t_1,t_3)} - \langle Q_1Q_3\rangle_{(t_1,t_2,t_3)} &= \langle Q_1Q_3\rangle_{(t_1,t_3)} \\ &- P(Q_2=+1)\langle Q_1Q_3|Q_2=+1\rangle_{(t_1,t_2,t_3)} \\ &- P(Q_2=-1)\langle Q_1Q_3|Q_2=-1\rangle_{(t_1,t_2,t_3)} \end{aligned}$$

The difference between $\langle Q_1Q_3\rangle_{(t_1,t_3)}$ and $\langle Q_1Q_3\rangle_{(t_1,t_2,t_3)}$ arises because including a Q_2 measurement at time t_2 and discarding the result will nevertheless project the quantum system onto the eigenstates of $\hat{\sigma}\cdot\hat{n}_2$ before the Q_3 measurement. We find that:

$$\begin{aligned} &P(Q_2=+1)\langle Q_1Q_3|Q_2=+1\rangle_{(t_1,t_2,t_3)} \\ &+P(Q_2=-1)\langle Q_1Q_3|Q_2=-1\rangle_{(t_1,t_2,t_3)} \\ &=\cos^2(\theta_{12}/2)\cos\theta_{23} - \sin^2(\theta_{12}/2)\cos\theta_{23} \end{aligned}$$

The expression for the disturbance then becomes:

$$\begin{aligned} D &= \langle Q_1Q_3\rangle_{(t_1,t_3)} - P(Q_2=+1)\langle Q_1Q_3|Q_2=+1\rangle_{(t_1,t_2,t_3)} \\ &\quad - P(Q_2=-1)\langle Q_1Q_3|Q_2=-1\rangle_{(t_1,t_2,t_3)} \\ &= \cos\theta_{13} - \cos^2(\theta_{12}/2)\cos\theta_{23} + \sin^2(\theta_{12}/2)\cos\theta_{23} \\ &= \cos\theta_{13} - \left(\cos^2(\theta_{12}/2) - \sin^2(\theta_{12}/2)\right)\cos\theta_{23} \\ &= \cos\theta_{13} - \cos\theta_{12}\cos\theta_{23} \end{aligned} \qquad \textbf{[S20]}$$

The condition for obtaining no detectable disturbance ($D = 0$) is therefore:

$$\cos\theta_{13} = \cos\theta_{12}\cos\theta_{23} \qquad \textbf{[S21]}$$

If the condition in Eq. **S21** can be satisfied, all observers will agree that a nondetectable measurement of Q_2 has taken place, whereas if the condition in Eq. **S21** is violated, no one could believe that NDM has taken place.

B. Zero Detectable Disturbance Implies the Leggett–Garg Inequality Is Satisfied. We now show that a measurement sequence that has no detectable disturbance will necessarily satisfy the Leggett–Garg inequality in a two-level quantum system. Starting with Eq. **S19**, we substitute in the condition $\cos\theta_{13} = D + \cos\theta_{12}\cos\theta_{23}$ from Eq. **S21**, obtaining:

$$\langle K\rangle = \cos\theta_{12} + \cos\theta_{23} + \cos\theta_{13}$$

$$\langle K\rangle = \cos\theta_{12} + \cos\theta_{23} + \cos\theta_{12}\cos\theta_{23} + D$$

Rearranging and collecting terms, we have:

$$\langle K - D + 1\rangle = \cos\theta_{12}\cos\theta_{23} + \cos\theta_{12} + \cos\theta_{23} + 1$$

$$\langle K - D + 1\rangle = (\cos\theta_{12} + 1)(\cos\theta_{23} + 1)$$

We find $\langle K - D + 1\rangle$ consists of the product of two terms, each of which is in the range $0 \le \langle\cos\theta_{ij} + 1\rangle \le 2$, such that the whole Leggett–Garg function is in the range $D - 1 \le \langle K\rangle \le D + 3$, or:

$$D = 0 \Rightarrow -1 \le \langle K\rangle \le +3$$

The condition for zero-disturbance is therefore identical to the condition that the Leggett–Garg inequality is satisfied, and violation of the Leggett–Garg inequality must be accompanied by detectable disturbance in the two-level case.

No measurement of a two-level quantum system could convince a stubborn macrorealist that a failure of SD alone has taken place, because NDM will necessarily have failed in any successful Leggett–Garg inequality violation demonstrated in a two-level quantum system. This is in comparison to our work on a three-level quantum system, in which we show that NDM remains valid, whereas a Leggett–Garg inequality is violated.

1. Analysis of the detectable disturbance. Here, we derive some results asserted in the proof above. The detectable disturbance D is the change in $\langle K\rangle$ induced by measuring at pairs of times vs. measuring at all three times. D contains three terms. We have:

$$D = \underbrace{\langle Q_1Q_2\rangle_{(t_1,t_2)} - \langle Q_1Q_2\rangle_{(t_1,t_2,t_3)}}_{D_{12}} + \underbrace{\langle Q_2Q_3\rangle_{(t_2,t_3)} - \langle Q_2Q_3\rangle_{(t_1,t_2,t_3)}}_{D_{23}} + \underbrace{\langle Q_1Q_3\rangle_{(t_1,t_3)} - \langle Q_1Q_3\rangle_{(t_1,t_2,t_3)}}_{D_{13}}$$

where D_{ij} represents the change to the expected value of the two-measurement correlation $\langle Q_iQ_j\rangle$ induced by performing the third measurement $Q_{k\neq i,j}$ while ignoring the Q_k result.

2. Evaluating D_{12} and D_{23}. Clearly, $D_{12} = 0$, because the measurement at time t_3 would otherwise influence the result of past events at time t_1 or t_2. We might suspect by symmetry that $D_{23} = \langle Q_2Q_3\rangle_{(t_2,t_3)} - \langle Q_2Q_3\rangle_{(t_1,t_2,t_3)}$ will also be zero, and we can show this explicitly. The overlaps of the Pauli operators $\hat{Q}_2 = \hat{\sigma}\cdot\hat{n}_2$ and $\hat{Q}_3 = \hat{\sigma}\cdot\hat{n}_3$ yield:

$$P(Q_3 = \pm 1|Q_2 = \pm 1) = \cos^2(\theta_{23}/2)$$

$$P(Q_3 = \pm 1|Q_2 = \mp 1) = \sin^2(\theta_{23}/2)$$

and:

$$\begin{aligned} \langle Q_2Q_3\rangle_{(t_2,t_3)} &= P(Q_3 = \pm 1|Q_2 = \pm 1) - P(Q_3 = \pm 1|Q_2 = \mp 1) \\ &= \cos^2(\theta_{23}/2) - \sin^2(\theta_{23}/2) = \cos\theta_{23} \end{aligned}$$

We now need to evaluate $\langle Q_2Q_3\rangle_{(t_1,t_2,t_3)}$. Suppose that the Q_1 preparation followed by Q_2 measurements yields $Q_2 =$

+1 with probability p and $Q_2 = -1$ with probability $1 - p$. We have:

$$\begin{aligned}\langle Q_2Q_3\rangle_{(t_1,t_2,t_3)} &= p[P(Q_3=+1|Q_2=+1)-P(Q_3=-1|Q_2=+1)]\\ &\quad+(1-p)[P(Q_3=-1|Q_2=-1)-P(Q_3=+1|Q_2=-1)]\\ &= p\cos^2(\theta_{23}/2)-p\sin^2(\theta_{23}/2)+(1-p)\cos^2(\theta_{23}/2)\\ &\quad-(1-p)\sin^2(\theta_{23}/2)=\cos^2(\theta_{23}/2)-\sin^2(\theta_{23}/2)=\cos\theta_{23}\end{aligned}$$

from which the influence of the Q_1 measurement represented by p cancels, implying $D_{23} = 0$.

3. Evaluating D_{13}. We can see that $\langle Q_1Q_3\rangle_{(t_1,t_3)}$ is insensitive to the state before Q_1, by substituting $t_1 \to t_2$ and following a similar argument as for $\langle Q_2Q_3\rangle_{(t_2,t_3)}$ above. We show that $\langle Q_1Q_3\rangle_{(t_1,t_2,t_3)}$ is also insensitive to the initial state by assuming that the state before Q_1 measurement yields $Q_1 = +1$ with probability q and $Q_1 = -1$ with probability $1 - q$. We have:

$$\begin{aligned}\langle Q_1Q_3\rangle_{(t_1,t_2,t_3)} &= P(Q_2=+1)\langle Q_1Q_3|Q_2=+1\rangle\\ &\quad+P(Q_2=-1)\langle Q_1Q_3|Q_2=-1\rangle\\ &= qP(Q_2=+1|Q_1=+1)[P(Q_3=+1|Q_2=+1)\\ &\quad-P(Q_3=-1|Q_2=+1)]+qP(Q_2=-1|Q_1=+1)\\ &\quad\times[P(Q_3=+1|Q_2=-1)-P(Q_3=-1|Q_2=-1)]\\ &\quad+(1-q)P(Q_2=+1|Q_1=-1)[P(Q_3=+1|Q_2=+1)\\ &\quad-P(Q_3=-1|Q_2=+1)]+(1-q)P(Q_2=-1|Q_1=-1)\\ &\quad\times[P(Q_3=+1|Q_2=-1)-P(Q_3=-1|Q_2=-1)]\end{aligned}$$

This expression contains 16 terms, yielding:

$$\begin{aligned}\langle Q_1Q_3\rangle_{(t_1,t_2,t_3)} &= q\cos^2(\theta_{12}/2)\left[\cos^2(\theta_{23}/2)-\sin^2(\theta_{23}/2)\right]\\ &\quad+q\sin^2(\theta_{12}/2)\left[\sin^2(\theta_{23}/2)-\cos^2(\theta_{23}/2)\right]\\ &\quad+(1-q)\sin^2(\theta_{12}/2)\left[-\cos^2(\theta_{23}/2)+\sin^2(\theta_{23}/2)\right]\\ &\quad+(1-q)\cos^2(\theta_{12}/2)\left[-\sin^2(\theta_{23}/2)+\cos^2(\theta_{23}/2)\right]\end{aligned}$$

The initial preparation represented by q again cancels:

$$\begin{aligned}\langle Q_1Q_3\rangle_{(t_1,t_2,t_3)} &= \cos^2(\theta_{12}/2)\left[\cos^2(\theta_{23}/2)-\sin^2(\theta_{23}/2)\right]\\ &\quad-\sin^2(\theta_{12}/2)\left[\cos^2(\theta_{23}/2)-\sin^2(\theta_{23}/2)\right]\\ &= \left[\cos^2(\theta_{12}/2)-\sin^2(\theta_{12}/2)\right]\left[\cos^2(\theta_{23}/2)\right.\\ &\quad\left.-\sin^2(\theta_{23}/2)\right]=\cos\theta_{12}\cos\theta_{23}\end{aligned}$$

The total expression for the detectable disturbance is then:

$$\begin{aligned}D &= \underbrace{D_{12}}_{0}+\underbrace{D_{23}}_{0}+D_{13}=\langle Q_1Q_3\rangle_{(t_1,t_3)}-\langle Q_1Q_3\rangle_{(t_1,t_2,t_3)}\\ &= \cos\theta_{13}-\cos\theta_{12}\cos\theta_{23}\end{aligned}$$

This is Eq. **S21** in the text above. Detectable disturbance is therefore a necessary consequence of violating the Leggett–Garg inequality in a two-level quantum system.

IV. Classical Models of the Three-Box Problem

A classical probability model of the three-box problem can be constructed using Kolmogorov's axioms by assuming that the system of three boxes exists at all times in a definite state of having one box occupied and the other two empty. States of the system are conventionally labeled "λ" in this context. The simplest model of three boxes sharing one ball assumes a one-to-one correspondence between the system states λ_j and the available boxes, such that if λ is known, all measurement results can be inferred with certainty; the system with three states λ_1, λ_2, and λ_3 behaves such that being in state λ_j corresponds to finding M_j-true and $M_{k\neq j}$-false. On any particular run of the experiment, the system is in a definite state λ, but we may not know what this state is. We can describe our knowledge of the state with the probability distribution over the λ, writing $P(\lambda_i)$ with $\sum_i P(\lambda_i) = 1$. We can also assume many equivalent microstates $\{\lambda_1, \lambda_1', \lambda''_2, \cdots\}$ exist, which produce identical results when studied with the M_j measurements. This situation is illustrated in Fig. S1.

We can consider a continuous set of states $\{\lambda\}$ and a probability distribution over these as $\mu(\lambda)$, such that:

$$\mu(\lambda)\in\mathcal{R}\qquad \mu(\lambda)\geq 0\;\forall\;\lambda\qquad \int d\lambda\;\mu(\lambda)=1$$

Given a probability distribution $\mu(\lambda)$, we can define a measurement function $\xi_{\text{context}}(\text{result}|\lambda)$ that describes how each state λ will respond when measured. The probability of observing a particular result is $P_{\text{context}}(\text{result})$; thus, for instance, if Alice is measuring M_1, we write $P_{M_1}(A)$ as the probability that Alice finds M_1-true, and $P_{M_1}(\neg A)$ as the probability that Alice finds M_1-false:

$$P_{M_1}(A)=\int d\lambda\;\xi_{M_1}(A|\lambda)\mu(\lambda)\qquad P_{M_1}(\neg A)=\int d\lambda\;\xi_{M_1}(\neg A|\lambda)\mu(\lambda)$$

The key difference between such a "classical" model and the quantum picture of the experiment is that quantum interference is not possible because $\mu(\lambda) \geq 0$. These concepts are explored in related work (5).

V. Deriving a Leggett–Garg Function for the Three-Box Game

For a two-level quantum system, we have shown that it is necessary to perform the measurements of $\langle Q_1Q_2\rangle$, $\langle Q_2Q_3\rangle$, and $\langle Q_1Q_3\rangle$ independently to obtain a violation of the Leggett–Garg inequality. In the three-level case, a measurement result such as $\neg M_1$ can preserve the coherence between states $|2\rangle$ and $|3\rangle$, and it is therefore possible to violate a Leggett–Garg inequality while making three sequential measurements during the same run of an experiment. To derive the Leggett–Garg function specific to the three-level case, we must understand that a macrorealist can make counterfactual inferences that differ from the inferences that are valid for a quantum system, for example:

$$\neg M_1 \Rightarrow (M_2\wedge\neg M_3)\vee(\neg M_2\wedge M_3) \qquad \textbf{[S22]}$$

Seeing box 1 empty means either that the ball is in box 2 and not in box 3 or that the ball is in box 3 and not in box 2. Superpositions between boxes 2 and 3 are not allowed in the MR picture, due to SD, but superpositions are allowed before measurement in the quantum case and can survive following a partial measurement in the unobserved states. We can exploit the difference between the MR and quantum mechanics (QM) positions to derive an expression for $\langle K\rangle$ using counterfactual inferences, which would be valid if MR (and specifically SD) holds; we can then look for a violation of the Leggett–Garg inequality using an expression derived using counterfactual inferences.

A basic property of the three-box problem is that Alice is unable to detect any effect of measurements made by Bob; by definition, a successful demonstration of the three-box problem uses nondisturbing measurements. If we can derive a Leggett–Garg function specific to the three-box case and show that the inequality is violated by an experiment that successfully implements the three-box problem, the assumption of SD is shown to be invalid.

We have shown that all successful violations of the Leggett–Garg inequality in a two-level quantum system must involve measurements that cause disturbance; therefore, by extending the Leggett–Garg inequality to a three-level system, we can show

that the absence of state definiteness alone is responsible for the Leggett–Garg inequality violation. In this way, we go beyond existing studies in the literature. We now derive the Legett–Garg inequality for the three-level system.

A. Probability Notation. In the macrorealist picture, finding the system in state j corresponds to finding a macroscopic object, such as a hidden ball, in a location, such as box j. We write probabilities $P_{M_j}(B)$ to indicate the chance that when Bob performs measurement M_j, he sees a full box (finds state j) and $P_{M_j}(\neg B)$ as the probability that he finds box j is empty (measures "not state j"). The probability of the combined event, where both Bob and Alice see full boxes (Bob finds state j, followed by Alice finding M_3-true on her final measurement), is $P_{M_j}(B \cap A) = P_{M_j}(B|A)P_{M_j}(A)$, whereas Alice's probability of finding M_3-true when Bob has made no intervening measurement is written as $P_N(A)$.

The probabilities $P_{M_j}(\cdots)$ and $P_N(\ldots)$ are well-defined in both quantum and macrorealist theories, but our objective is to highlight the differences between these two theoretical descriptions. A macrorealist further believes that "counterfactual" expressions take a definite value. He defines quantities, such as $\tilde{P}_{\tilde{M}_j}(B)$, that give the probability for Bob to have found the object had he performed M_j. This allows a macrorealist to insert a resolution of the identity into his expressions for probabilities as:

$$\tilde{P}_{\tilde{M}_1}(B) + \tilde{P}_{\tilde{M}_2}(B) + \tilde{P}_{\tilde{M}_3}(B) = 1 \qquad \textbf{[S23]}$$

wherever he chooses. (We track quantities that are undefined in quantum mechanics with tilde symbols.)

B. Leggett–Garg Function for the Three-Level System. Given the definition of:

$$\langle K \rangle = \langle Q_1 Q_2 \rangle + \langle Q_2 Q_3 \rangle + \langle Q_1 Q_3 \rangle \qquad \textbf{[S24]}$$

we apply the Leggett–Garg analysis to our system as follows. Our experiment uses measurement-based initialization (4) to prepare the initial state for Alice, and we therefore take $Q_1 = +1$ in all cases. We assign $Q_2 = -1$ whenever Bob observes the object in box 1 or box 2 and $Q_2 = +1$ whenever Bob should infer the object is in box 3. We assign $Q_3 = +1$ whenever Alice's final M_3 result is true and assign $Q_3 = -1$ whenever the M_3 result is false.

If the macrorealist framework is applicable, one of six possible measurement histories (*a*–*f*) must account for each particular run of the experiment (Table S1). To assign the probabilities that a given history occurred, the macrorealist must calculate the unobserved quantities $\tilde{P}_{\tilde{M}_3}(B \cap A)$ and $\tilde{P}_{\tilde{M}_3}(B \cap \neg A)$. If the measurements are operationally nondisturbing (a property that we check experimentally), it is possible to substitute $\tilde{P}_{\tilde{M}_1}(B|A) \Rightarrow P_{M_1}(B|A)$ and $\tilde{P}_{\tilde{M}_2}(B|A) \Rightarrow P_{M_2}(B|A)$ (etc.) for the measurements that are made, such that:

$$\tilde{P}_{\tilde{M}_3}(B) = 1 - P_{M_1}(B) - P_{M_2}(B) \qquad \textbf{[S25]}$$

$$\tilde{P}_{\tilde{M}_3}(B \cap A) = P_N(A) - P_{M_1}(B \cap A) - P_{M_2}(B \cap A) \qquad \textbf{[S26]}$$

$$\tilde{P}_{\tilde{M}_3}(B \cap \neg A) = P_N(\neg A) - P_{M_1}(B \cap \neg A) - P_{M_2}(B \cap \neg A) \qquad \textbf{[S27]}$$

Using these definitions, the macrorealist framework deduces the expression for $\langle K \rangle$ (Table S1) as:

$$\begin{aligned}\langle K \rangle = &- P_{M_1}(B \cap A) - P_{M_2}(B \cap A) + 3\tilde{P}_{\tilde{M}_3}(B \cap A) \\ &- P_{M_1}(B \cap \neg A) - P_{M_2}(B \cap \neg A) - \tilde{P}_{\tilde{M}_3}(B \cap \neg A)\end{aligned} \qquad \textbf{[S28]}$$

which, in terms of observable quantities, is:

$$\begin{aligned}\langle K \rangle = &- P_{M_1}(B \cap A) - P_{M_1}(B \cap \neg A) - P_{M_2}(B \cap A) - P_{M_2}(B \cap \neg A) \\ &+ 3P_N(A) - 3P_{M_1}(B \cap A) - 3P_{M_2}(B \cap A) \\ &- P_N(\neg A) + P_{M_1}(B \cap \neg A) + P_{M_2}(B \cap \neg A)\end{aligned} \qquad \textbf{[S29]}$$

This expression simplifies to:

$$\langle K \rangle = 4P_N(A) - 4P_{M_1}(B \cap A) - 4P_{M_2}(B \cap A) - 1 \qquad \textbf{[S30]}$$

We know that the quantum expressions for these occurrences are:

$$P_N(A) = |\langle F|I\rangle|^2 = 1/9 \qquad \textbf{[S31]}$$

$$P_{M_1}(B \cap A) = |\langle F|\hat{P}_1|I\rangle|^2 = 1/9 \qquad \textbf{[S32]}$$

$$P_{M_2}(B \cap A) = |\langle F|\hat{P}_2|I\rangle|^2 = 1/9 \qquad \textbf{[S33]}$$

We observe two points here, given the probabilities above. The first is that Alice is unable to determine whether Bob has chosen to perform measurement M_1, measurement M_2, or neither measurement (N), and Alice's result is independent of measurement context. We have:

$$P_{M_1}(A) = P_{M_2}(A) = P_N(A) \qquad \textbf{[S34]}$$

The second probability is that:

$$P_{M_1}(B\neg \cap A) = P_{M_2}(B\neg \cap A) = 0 \qquad \textbf{[S35]}$$

Alice will never find her M_3 result true when Bob has found his M_j result false, which is the key feature that enables Alice to win the three-box game. This implies:

$$P_{M_1}(B \cap A) = P_{M_1}(A) \qquad \textbf{[S36]}$$

$$P_{M_2}(B \cap A) = P_{M_2}(A) \qquad \textbf{[S37]}$$

We can extract the probability of Alice's measurement from each term via Bayes theorem:

$$P((B \cap A)) = P(B|A)P(A) \qquad \textbf{[S38]}$$

yielding:

$$\begin{aligned}\langle K \rangle &= 4P(A)(1 - P_{M_1}(B|A) - P_{M_2}(B|A)) - 1 \\ &= \frac{4}{9}(1 - P_{M_1}(B|A) - P_{M_2}(B|A)) - 1\end{aligned} \qquad \textbf{[S39]}$$

Given the macrorealist's hypothesis, the events under M_1 and M_2 should be mutually exclusive, and sums of events under these cases will obey an inequality:

$$P_{M_1}(\cdots) + P_{M_2}(\cdots) \leq 1 \qquad \textbf{[S40]}$$

The equality holds when the Leggett–Garg function in Eq. **S39** takes as its limiting value $\langle K \rangle = -1$. In the quantum case, meanwhile, $P_{M_1}(\cdots)$ and $P_{M_2}(\cdots)$ are independent, and we have:

$$P_{M_1}(\cdots) + P_{M_2}(\cdots) \leq 2 \qquad \textbf{[S41]}$$

allowing the Leggett–Garg function to reach a value of:

$$\langle K \rangle = -\frac{13}{9} = -1.4\dot{4} \qquad \textbf{[S42]}$$

This is outside the range $-1 \leq \langle K \rangle \leq 3$, providing an opportunity to detect an inconsistency with MR.

VI. Error Analysis of the Experimental Results

We find small deviations from the values expected from an ideal implementation. In the following, we give a brief description of the origin of these discrepancies and discuss their consequences on the macrorealist's possible conclusions.

i) We find $\sum_j P_{M_j}(B) < 1$ (i.e., there is not always a ball found in all the boxes). This is a consequence of a smaller than unity probability of correctly identifying the electronic $m_S = 0$ state, resulting in an effective detection efficiency of $P_{det} \approx 90\%$. Although the macrorealist might conclude that there is not always an object hidden in the boxes, he still finds an unbiased initial state (within statistical uncertainty). Therefore, he cannot expect Alice to take advantage of this discrepancy. Based on his secret choice of M_1 or M_2 and the reduced probability of finding an object, he expects a maximum probability of $\frac{P_{det}}{2} \leq \frac{1}{2}$ of Alice predicting his positive measurement outcome correctly; thus, the macrorealist finds an even stronger violation of his expectations.

ii) For sequential measurements i, $i+1$, we find both $P(M_{j,i+1}|M_{j,i})$ and $P(\neg M_{j,i+1}|\neg M_{j,i}) < 100\%$ (Fig. 3*B*) (i.e., after measuring its position, with a small probability, the object is moved to a different box). This finding could indeed explain correlations between Bob and Alice's measurements: As a worst-case scenario, Bob could assume a hidden mechanism in the game whereby his successful measurement "moves" the object, deterministically storing it in the box Alice is probing and maximizing her conditional probability $P_{M_j}(B|A)$. He would deduce an upper limit for her success probability of $P_{M_j}(B|A) \leq \frac{1}{3} + P(\text{object moves})$. Taking into account all "Changed" and "Undetermined" events (Fig. 3*B*), he finds $P(\text{object moves}) \leq 28\%$ and $P_{M_j}(B|A) \leq 61\%$, clearly violated by the experimental findings.

iii) We find $P(A) \approx 14\% > 1/9$ (Fig. 4*A*). However, from the QM description, we expect:

$$P_N(A) = |\langle F|I\rangle|^2 = 1/9 \quad \textbf{[S43]}$$

$$P_{M_1}(B \cap A) = |\langle F|\hat{P}_1|I\rangle|^2 = 1/9 \quad \textbf{[S44]}$$

$$P_{M_2}(B \cap A) = |\langle F|\hat{P}_2|I\rangle|^2 = 1/9 \quad \textbf{[S45]}$$

In our implementation, between measurement A and B, we apply the transformation $|F\rangle \rightarrow |I\rangle \rightarrow |3\rangle$, consisting of NMR pulses of a total duration of ≈750 μs. The rf-induced heating of the sample and nuclear spin dephasing limit the fidelity of this operation, leading to an increased probability $P(A)$. In the QM picture, Alice detects more positive results than she should (unconditional on measurement B); thus, her conditional probability $P_{M_j}(B|A)$ to predict Bob's measurement outcome correctly must drop below the theoretical maximum of 100% (Fig. 4*B*).

A. Statistical Error Analysis. For each particular run of our experiment, we can either count one or more photons ($n \geq 1$) or no photons ($n = 0$), inferring that the electron is in the $m_S = 0$ or $m_S = \pm 1$ state. A detailed analysis of the inferences between photon number and spin state was presented by Robledo et al. (4) as a combination of geometric distribution (accounting for the spin flip rate), binomial distribution (accounting for photon detector efficiency), and the poissonian background rate. For the purposes of our analysis, we define a variable $P = |m_S|$, which is the value of Bob's or Alice's M_j result on any particular round of the experiment. We assign $P = 0$ when we count $n = 0$ photons, and we assign $P = 1$ when we count $n \geq 1$ photons. We then define the probability p of finding $m_S = 0$ during a particular shot of the experiment as f, so that during N trials of the experiment, we expect to observe statistics:

$$\text{Mean}[p] = Nf \quad \textbf{[S46]}$$

$$\text{Var}[p] = \sigma^2(p) = Nf(1-f) \quad \textbf{[S47]}$$

$$\text{Standard Deviation}[p] = \sigma(p) = \sqrt{Nf(1-f)} \quad \textbf{[S48]}$$

We use this to calculate the statistical significance of our results (e.g., the chance that the Leggett–Garg function we measured is compatible with MR and that counting statistics have produced a violation by chance).

B. Fair Sampling vs. Adversarial Macrorealist Positions. In our experiment, we have the option to measure either the population in electron spin sublevel $m_S = -1$ or in the electron spin sublevels $m_S = -1$ and $m_S = +1$ when performing Bob and Alice's measurements M_j. Measuring the $m_S = -1$ populations only, we have the possibility of obtaining "undetermined" outcomes in which the population branches from $m_S = -1$ to the uninspected $m_S = +1$ levels during measurement, whereas by measuring the $m_S = -1$ and $m_S = +1$ levels, we minimize these undetermined events while increasing the number of Δm_I nuclear spin flips, which corresponds to Bob measuring that the state has definitely changed between subsequent measurements.

There are two approaches that we could use to interpret the undetermined outcomes. The default assumption is that the unmeasured values are distributed fairly and will follow the same distribution as the measured values, whereas the more extreme assumption is that each unmeasured value somehow represents Alice "cheating" by hiding values that favor the macrorealist hypothesis. If we take this extreme position, it is interesting to know whether a result compatible with MR could be recovered by allowing Bob to assign a value to each undetermined result as he pleases (6). We then define quantities such as:

$$P(A \cap B)^{\min}_{M1} = \frac{N_{M1}(B \cap A)}{N_{M1}(B \cap A) + N_{M1}(\neg B \cap A) + N_{M_1}(U)} \quad \textbf{[S49]}$$

$$P(A \cap B)^{\text{fair}}_{M1} = \frac{N_{M1}(B \cap A)}{N_{M1}(B \cap A) + N_{M1}(\neg B \cap A)} \quad \textbf{[S50]}$$

$$P(A \cap B)^{\max}_{M1} = \frac{N_{M1}(B \cap A) + N_{M_1}(U)}{N_{M1}(B \cap A) + N_{M1}(\neg B \cap A) + N_{M_1}(U)} \quad \textbf{[S51]}$$

where $N_{M_1}(U)$ is the number of undetermined measurement readings, given that Bob has performed M_1. This bounds the possibilities for Bob to reassign undetermined readings. In fact, both in the case in which we assume fair sampling and without, we find that $\langle K \rangle \leq -1$ and that our results are therefore incompatible with MR. We calculate each case and include errors as per our statistical analysis above. In the case that we include only the $m_S = -1$ readout, we find:

$$K^{\min}_{|m_S=-1} = -1.2026 \quad \sigma^{\min}_{|m_S=-1} = 0.0259 \quad (7.81\sigma \text{ violation}) \quad \textbf{[S52]}$$

$$K^{\text{fair}}_{|m_S=-1} = -1.2647 \quad \sigma^{\text{fair}}_{|m_S=-1} = 0.0234 \quad (11.29\sigma \text{ violation}) \quad \textbf{[S53]}$$

$$K^{\max}_{|m_S=-1} = -1.3494 \quad \sigma^{\max}_{|m_S=-1} = 0.0173 \quad (20.19\sigma \text{ violation}) \quad \textbf{[S54]}$$

When using the complete register readout on $m_S = -1$ and $m_S = +1$, we have:

$$K^{\min}_{|m_S=\pm1} = -1.1373 \quad \sigma^{\min}_{|m_S=\pm1} = 0.0252 \quad (5.46\sigma \text{ violation}) \quad \textbf{[S55]}$$

$$K^{\text{fair}}_{|m_S=\pm1} = -1.1833 \quad \sigma^{\text{fair}}_{|m_S=\pm1} = 0.0241 \quad (7.60\sigma \text{ violation}) \quad \textbf{[S56]}$$

$$K^{\max}_{|m_S=\pm1} = -1.2531 \quad \sigma^{\max}_{|m_S=\pm1} = 0.0210 \quad (12.07\sigma \text{ violation}) \quad \textbf{[S57]}$$

In the event, we found that the undetermined measurement outcomes do not give Bob sufficient leeway to explain the discrepancy of our result from the range predicted by MR, even when taking the most adversarial position permissible with respect to our data.

1. Leggett AJ, Garg A (1985) Quantum mechanics versus macroscopic realism: Is the flux there when nobody looks? *Phys Rev Lett* 54(9):857–860.
2. Aharon N, Vaidman L (2008) Quantum advantages in classically defined tasks. *Phys Rev A* 77(5):052310.
3. Bernien H, et al. (2012) Two-photon quantum interference from separate nitrogen vacancy centers in diamond. *Phys Rev Lett* 108(4):043604.
4. Robledo L, et al. (2011) High-fidelity projective read-out of a solid-state spin quantum register. *Nature* 477(7366):574–578.
5. Maroney OJE (2012) Detectability, invasiveness and the quantum three box paradox. *arXiv quant-ph*, 1207.3114v1.
6. Knee GC, et al. (2012) Violation of a Leggett-Garg inequality with ideal non-invasive measurements. *Nat Commun* 3:606.

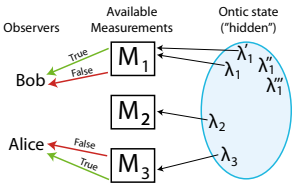

Fig. S1. Classical model of the three-box problem. In a simple classical model, the system is assumed to exist in a definite state λ_j. The specific state λ_j then determines how the system will respond to each measurement M_j.

Table S1. Assignment Q_j values for each run of the experiment

Case	Q_1	Bob measures	Q_2	Alice measures	Q_3	K	Probability
a	+1	M_1	−1	M_3	+1	−1	$P_{M_1}(B \cap A)$
b	+1	M_2	−1	M_3	+1	−1	$P_{M_2}(B \cap A)$
c	+1	Infers M_3	+1	M_3	+1	+3	$\tilde{P}_{\tilde{M}_3}(B \cap A)$
d	+1	M_1	−1	$\neg M_3$	−1	−1	$P_{M_1}(B \cap \neg A)$
e	+1	M_2	−1	$\neg M_3$	−1	−1	$P_{M_2}(B \cap \neg A)$
f	+1	Infers M_3	+1	$\neg M_3$	−1	−1	$\tilde{P}_{\tilde{M}_3}(B \cap \neg A)$

According to the MR picture, one of the six cases above must account for each run of the experiment. The measured probabilities P_{M_1} and P_{M_2} and inferred (counterfactual) probabilities $\tilde{P}_{\tilde{M}_3}$ that weight the value of K corresponding to each history are listed in the table.

Table S2. Enumerating values of Q_1, Q_2, and Q_3 for the Leggett–Garg function in a classical system

Measurements			Parity checks/correlators			Leggett–Garg function
Q_1	Q_2	Q_3	Q_1Q_2	Q_2Q_3	Q_1Q_3	K
+1	+1	+1	+1	+1	+1	3
−1	−1	−1	+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	−1
−1	+1	−1	−1	−1	+1	−1
+1	−1	−1	−1	+1	−1	−1

Each combination of Q_j yields a K value between −1 and +3.

Other Supporting Information Files

Dataset S1 (PDF)
Dataset S2 (PDF)