SUPPLEMENTARY INFORMATION

doi:10.1038/nature15759

# Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres

## SUPPLEMENTARY INFORMATION

B. Hensen,[1,2] H. Bernien,[1,2,*] A.E. Dréau,[1,2] A. Reiserer,[1,2] N. Kalb,[1,2] M.S. Blok,[1,2] J. Ruitenberg,[1,2] R.F.L. Vermeulen,[1,2] R.N. Schouten,[1,2] C. Abellán,[3] W. Amaya,[3] V. Pruneri,[3,4] M.W. Mitchell,[3,4] M. Markham,[5] D.J. Twitchen,[5] D. Elkouss,[1] S. Wehner,[1] T.H. Taminiau,[1,2] and R. Hanson[1,2,†]

[1] *QuTech, Delft University of Technology, P.O. Box 5046, 2600 GA Delft, The Netherlands*
[2] *Kavli Institute of Nanoscience Delft, Delft University of Technology, P.O. Box 5046, 2600 GA Delft, The Netherlands*
[3] *ICFO-Institut de Ciencies Fotoniques, The Barcelona Institute of Science and Technology, 08860 Castelldefels (Barcelona), Spain.*
[4] *ICREA-Institució Catalana de Recerca i Estudis Avanats, Lluis Companys 23, 08010 Barcelona, Spain.*
[5] *Element Six Innovation, Fermi Avenue, Harwell Oxford, Didcot, Oxfordshire OX110QR, United Kingdom.*

## EXPERIMENTAL

### A. Experimental setup

The experiments are performed on individual NV centres that are naturally present in high purity type IIa chemical-vapor deposition diamond samples (Element Six), with a $\langle 111 \rangle$ crystal orientation. In its negative charge state, the NV centre ground state is an electronic spin triplet with total spin 1. The zero-field splitting separates the $|m_s = 0\rangle$ and $|\mathrm{m}_s = \pm 1\rangle$ levels by 2.88 GHz. Additionally we split the $|\mathrm{m}_s = \pm 1\rangle$ levels by 0.14 GHz using a static magnetic field applied along the defect axis. We use the two levels $|m_s = 0\rangle$ and $|m_s = -1\rangle$, denoted as $|\uparrow\rangle$ and $|\downarrow\rangle$, respectively. We use microwave control pulses, applied via a gold stripline deposited on the sample surface, to rotate the electronic ground state spin. The carrier frequency of the pulses is resonant with the $|\mathrm{m}_s = 0\rangle$ to $|\mathrm{m}_s = -1\rangle$ ground state transition. The electronic spin resonance spectrum of both NV centres is split into three lines by the hyperfine interaction with the host nitrogen nuclear spin of the defect. Therefore, we chose a Hermite pulse envelope shape, which provides a broad and flat spectral distribution. In this way, we achieve a $\pi$ rotation within 180 ns with a fidelity exceeding 99.8% without initializing the nitrogen nuclear spin state.

The samples are kept at a temperature of 4 K in closed-cycle cryostats (Montana Instruments) to enable resonant optical excitation of spin-dependent transitions. This enables fast single-shot readout and high-fidelity optical initialization of the electronic spin state. The design of the electronics and optical setup is described in detail in previous work [29],[33]. To initialize the electronic spin into $|\downarrow\rangle$, we apply resonant excitation on the $|\mathrm{m}_s = \pm 1\rangle \leftrightarrow |E'\rangle$ optical transition, achieving an initialization fidelity of more than 99.8% within 5 $\mu$s. Electronic spin readout is accomplished by resonant excitation of the $E_x(E_y)$ transition for the sample in setup A (B). To guarantee a high readout fidelity by minimizing the spin mixing in the excited state, we chose samples with low transversal strain splitting and additionally use d.c. Stark tuning by applying d.c. electric fields to the on-chip electrodes (Fig. 1c inset). The average strain splitting during the experiment was 1.8 (2.3) GHz in sample A (B).

A high photon collection efficiency is a prerequisite for high fidelity optical single-shot readout and high-efficiency entanglement. To overcome the limitation of total internal reflection of the emitted photons, we fabricate solid immersion lenses in the diamond surface around preselected NVs (oriented along the $\langle 111 \rangle$ crystal direction) using a focused ion beam. In addition, we deposit a single-layer aluminium oxide anti-reflection coating. We select devices with a particularly high phonon-sideband (PSB, wavelength > 640 nm) photon collection efficiency in our home-built

* Present address: Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA
† r.hanson@tudelft.nl

confocal microscope setups, which we measure to be $(13 \pm 2)\%$. Note that this value includes the 80% quantum efficiency of the detectors. The zero-phonon line (ZPL) photon emission of each NV is separated from the PSB using a dichroic long-pass and an additional tunable bandpass filter. Per optical excitation, we collect on average $2.9 \cdot 10^{-4}$ ($1.7 \cdot 10^{-4}$) ZPL photons at location C after propagation through a single-mode fibre from location A (B). Measured fiber losses are 6 dB from A to C and 9 dB from B to C. On setup A, we use adaptive optics to couple the ZPL emission into the single-mode fiber, increasing the efficiency by about a factor of two, compared to using standard optics (Section M). We use two polarisers at C to block unwanted reflections from the excitation pulse. To compensate for polarization drift inside the fibers, an automated polarization control feedback loop adjusts two waveplates in front of the fiber couplers at A and B.

Synchronization of the individual setups is achieved via separate glass fiber connections. Distances between locations are determined using cartographic and altimetric data (Sections G, H).

One of the main challenges in performing the Bell experiment, is to maintain the required stability of all the involved NV centres, instruments, lasers and detection optics over a week time-scale. In Section J we list the various checks, optimisations and calibrations performed to achieve this goal.

### B. Two-photon quantum interference (Figure 3b)

Measurement of photon coincidences in different output ports of a beam splitter has become a standard tool to investigate the indistinguishability of single photons generated by different sources. To account for potentially unequal efficiency of the sources, one can compare the case where the photons are indistinguishable to one where they are made distinguishable on purpose. In our experimental system, using orthogonal polarization is prohibited by in-line fibre polarisers. We instead program the experimental sequence such that setup A generates photons in an 'early' time window and setup B in a 'late' time window which is delayed with respect to the early one by 300 ns. Since this is much longer than the optical lifetime of the NV centre, this allows us to clearly distinguish which of the setups has emitted a given detected photon. During the measurement this distinguishable setting is alternated (every 3750 excitation pulses) with the indistinguishable setting, in which both setups produce photons that arrive at the same time at the beamsplitter.

In the data analysis for the distinguishable setting the arrival times of the photons in the late time window (coming from setup B) are shifted back by 300 ns in order to overlap them with those from setup A. In this way, we can present the data in a more familiar way (Figure 3b). Note also that the data in Figure 3b is taken on a different NV pair than the one used in the other experiments; the two pairs are very similar and are therefore expected to yield equivalent results.

### C. Model of the entangled state (Figures 3c and 4a)

We estimate the readout fidelities of NV A and B from the daily calibration measurements during the recording of the XX and ZZ entanglement data (Figure 3c). We find: $F_0^A = 0.9536 \pm 0.0030$, $F_1^A = 0.9940 \pm 0.0011$, $F_0^B = 0.9390 \pm 0.0034$, $F_1^B = 0.9982 \pm 0.0006$. Using these values, we correct the spin-photon correlation data of Figure 3a for readout errors [33], to obtain the residual errors made in the spin-photon correlation, $e^{A,B}_{\text{early,late}}$ presented in the caption of Figure 3a.

To obtain an estimate for the fidelity of the generated entangled state, we model the state of the two NVs after a successful heralding event by a density matrix of the following form (See SI of [33]):

$$\rho = \frac{1}{2}\begin{pmatrix} 1-F_z & 0 & 0 & 0 \\ 0 & F_z & -V & 0 \\ 0 & -V & F_z & 0 \\ 0 & 0 & 0 & 1-F_z \end{pmatrix}. \tag{1}$$

Here we set

$$F_z = \frac{1}{2}\left((1-e^A_{\text{early}})(1-e^B_{\text{late}}) + (1-e^B_{\text{early}})(1-e^A_{\text{late}})\right), \tag{2}$$

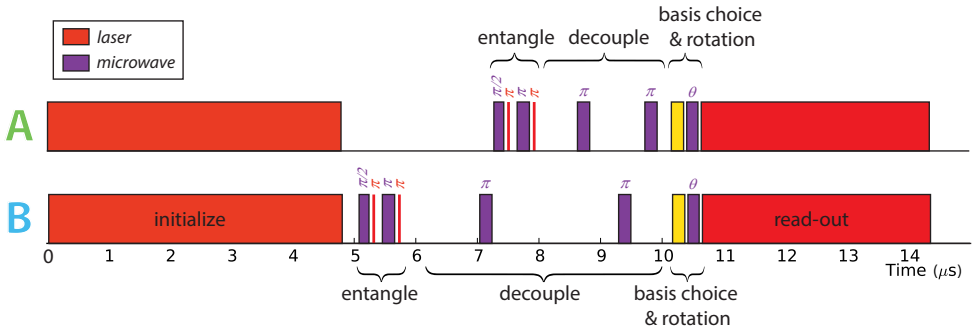


FIG. S1. Pulse sequence of the experiment. First, the spins at A (top) and B (bottom) are initialized by optical pumping. Then we perform the entangling sequence, consisting of two microwave and two optical pulses, followed by a dynamical decoupling sequence consisting of two microwave $\pi$-pulses that preserves the spin coherence. Finally, the readout basis is chosen and implemented and the spins are read out.

$V$ can be estimated from the measured interference contrast in Figure 3b, in combination with the expected 3% reduction in phase coherence from the instability in excitation laser frequency (see section L below). The statistical uncertainty in the estimated $V$ is large because of the small number of events in the interference experiment. Our best estimate is $V \approx 0.9 * 0.97 = 0.873 \pm 0.060$. This yields an estimate for the state fidelity $\langle\Psi^-|\rho|\Psi^-\rangle = 0.92 \pm 0.03$. To get the expected correlation values for Figure 3c, we numerically perform the corresponding final basis rotations $U_{a,b}$ on $\rho$ and apply the expected readout-errors on the obtained density matrix. We use the same model to numerically find the best basis settings for the Bell experiment and to calculate the expected correlators in Figure 4a. This results in an expected S-parameter of $2.30 \pm 0.07$.

### D. Dynamical decoupling sequence

The coherence of the NV centre spins is limited by the interaction with a bath of $^{13}$C nuclear spins, resulting in a dephasing time $T_2^*$ of a few $\mu$s. To counteract this dephasing we apply a dynamical decoupling sequence that consists of two microwave (MW) $\pi$-pulses with appropriate spacing. The full experimental pulse sequence is shown in Fig. S1. To quantify the remaining detrimental effect of spin dephasing during the experiment, we omit the two optical $\pi$-pulses used to generate the spin-photon entanglement and replace the last MW rotation by a $\pi/2$ pulse. Ideally, this sequence should bring the spin to the state $|\downarrow\rangle$. The measured probability to end up in this state is above 99 % showing that decoherence is efficiently mitigated.

### E. Event-ready signal settings

The signal that heralds successful entanglement, recorded at location C, consists of one photon detection event in an early time window and one in a late time window, each in a different output port of the beam splitter. We have to define the start- and stop times of these $2 \times 2$ windows. Ideally the windows only contain detection events due to photons emitted by the NV centres at location A and B. The start time of the window should be late enough such that unwanted reflections from the laser excitation pulse are filtered out, while the stop time should be early enough such that the background count-rate is not dominating the detection probability.

We use the characterisation data of the ZZ and XX entanglement runs presented in Figure 3c of the main text to decide on time-window settings for the event-ready signal, prior to starting the Bell experiment. Figure S2 illustrates the method used to obtain a good window start-time and length, for each of the two channels, for the early time-window. For the late time-window we use the same settings, but shifted by the fixed time between the two laser excitation pulses in the entanglement protocol (250 ns).

Although the conservative settings thus decided on were used for the Bell test presented in the main text, the

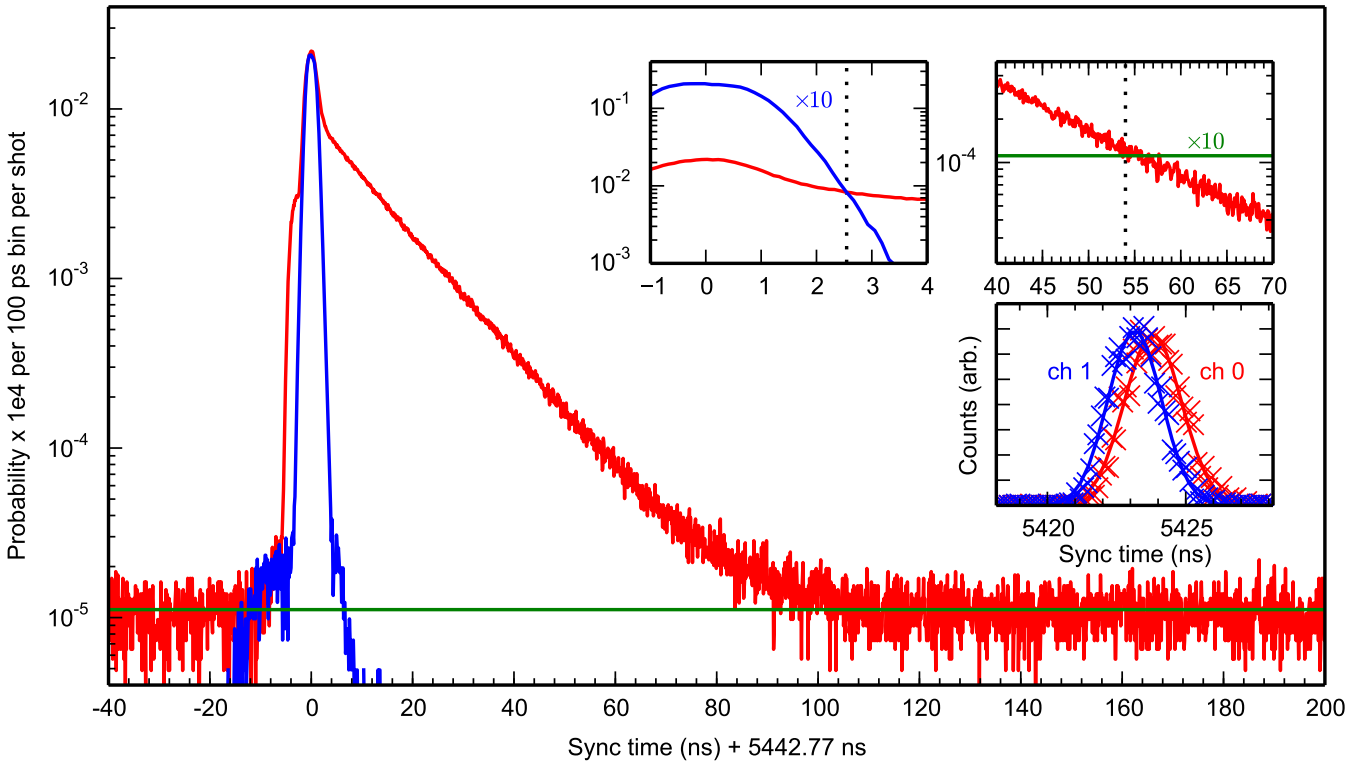


FIG. S2. Obtaining time-window settings for the event-ready signal. Red line: Histogram of the arrival times of the early time-bin photons from both setups together, during the XX entanglement characterisation, measured at location C. We can clearly distinguish the 2 ns, Gaussian shaped excitation pulse, followed by the exponential decaying emission from the NV's, and finally the constant background countrate. (The 1 ns plateau observed just before the laser pulse was caused by a reflection on an optical element in the excitation path of setup A). From simulations we expect to obtain $S > 2$ only when the probability to obtain a click from either NV is at least 10 times larger than the probability to obtain a spurious click from either the laser pulse or the background. Using the independently measured shape of the laser pulse (blue line), and the background count-rate fitted from the histogram after 120 ns (green line), we determine the start and stop of the filter window that satisfies this criterion (upper insets). We find a time window of $2.55 - 55$ ns after the centre of the laser-pulse. Before starting the Bell experiment, we measured the arrival of the laser pulse for the (different) detectors used, for each channel (lower inset). From the Gaussian fit, we find the centre of the laser pulse to be at $5423.80 \pm 0.01$ ns and $5423.15 \pm 0.01$ ns after the arrival of the sync pulse, for detectors 0 and 1 respectively. We then fixed the window settings for heralding events at C, to be used in the Bell test: the early window starts 5426.35 (5425.70) ns after the sync-pulse, and stops 5478.80 (5478.15) ns after the sync-pulse for channel 0 (channel 1).

dataset recorded during the Bell experiment contains all the photon detection times at location C. This allows us to investigate the effect of choosing different window-settings in post-processing. In Figure S3 we present the dependence of the recorded Bell violation $S$, and number of Bell trials $n$, if we offset the start of the windows.

### F. Bell violation for shorter readout duration

The data presented in the main text considered the maximum integration time allowed for the spin readout, given a 160 ns time for the generation of a random number, that still ensures the required locality conditions. By shortening the integration time-window in post-processing, we can explore testing additional models beyond those included in our

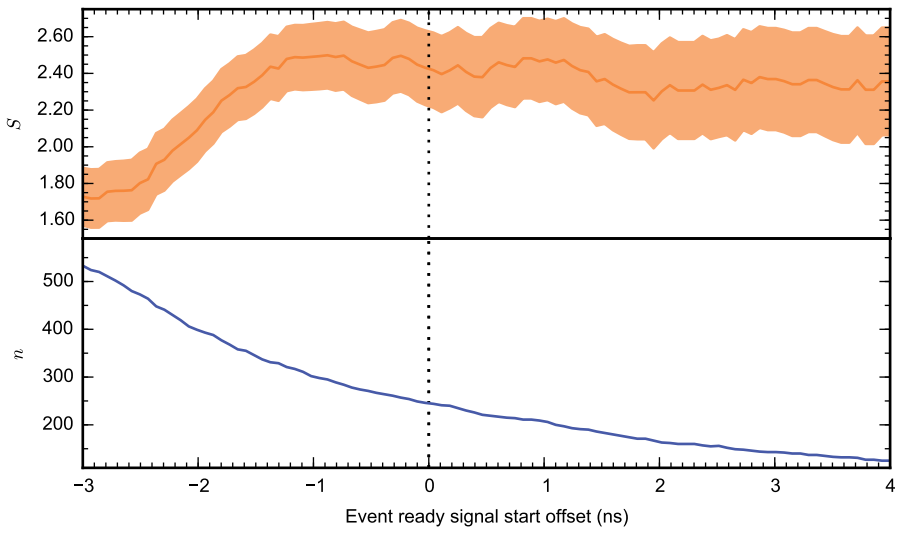


FIG. S3. CHSH parameter $S$ and number of Bell trials $n$ versus window start for the event-ready photon detections at location C. The time-offset shown is with respect to the windows given in Figure S2. Thus, the dotted line at zero denotes the settings as defined prior to starting the Bell experiment and used in the main text. Confidence region shown is one sigma, calculated according to the conventional analysis (see main text). Shifting the window back in time, the relative fraction of heralding events caused by clicks from the excitation laser reflections increases, thereby reducing the observed Bell violation.

.

null hypothesis. In particular, shortening the readout window while still observing a violation can test theories where the inputs are already determined earlier by some unknown physics. We can shift the proposed determination of the inputs further back in time, until the event-ready signal will no longer be space-like separated from the random input at location A (see Figure 2a). At this point we have 690 ns for the generation of a random bit. Shortening the readout to retain the required locality conditions given the earlier determined inputs, we find a violation of $S = 2.39 \pm 0.21$, which would correspond to a $P$-value of about 0.032 (0.054) for the conventional (complete) analysis. Note however that our null hypothesis formulated before the test is based on using a 3700 ns readout duration, and further analysis may be required to find the actual $P$-value for shorter readout durations. Alternatively, one can test theories that predict a maximum speed of physical influences beyond the speed of light. Figure S4 shows the dependence of the $S$ parameter on the readout integration time.

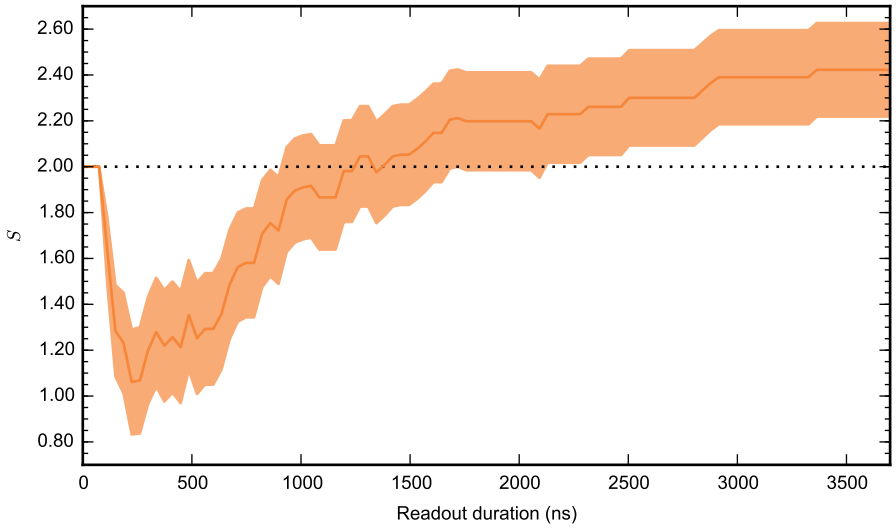


FIG. S4. CHSH parameter $S$ versus readout integration time at locations A and B. Confidence region shown is one sigma, calculated according to the conventional analysis (see main text).

### G. Location and distances

The experiment comprises three locations, A, B and C, defined as the position of the three time-tagging devices used. For locations A and B, the random number generators were located within 1 metre distance from the time-taggers, and the diamond samples and local photon detectors within 2 metre distance. For location C, the beam-splitter and photon detectors were located within 1 metre distance from the time-tagger. For each location X,Y and Z coordinates are determined in the following way:

First, a reference point on the outer wall of the building containing the setup is chosen. Then, the XYZ coordinates of the reference point are determined using the Large Scale Standard Map of The Netherlands (GBKN [34]), which has an accuracy better than 0.3 m, combined with the current elevation map of the Netherlands (AHN) [35], with a horizontal accuracy better than 0.5 m, and a vertical accuracy better then 0.1 m. Note that this accuracy is comparable to what can be obtained with GPS aided measurements. Finally, the relative position of the setup to the reference point is determined by manual distance measurements. The coordinates obtained in this manner are shown in Table S1. By converting to 3-dimensional Cartesian coordinates, the relative distances between the three locations is calculated with an uncertainty of less than $\pm 1.5$ m.

### H. Synchronisation of the experimental setups

Direct communication between the three labs A, B and C is implemented by means of electronic-optical converters (Highland Technology J720/J724/J730) and infrared optical fibres. To synchronize A and B during the Bell test, we send and record a common start reference time before every entanglement attempt. Towards this end, a trigger signal is sent by the arbitrary waveform generator (AWG) of B to both the time-tagger at C and the AWG at A. We calibrate the static trigger delays between A-B and B-C by measuring the round trip delays, including AWG trigger delay and local cable delays. The measured values are shown in Table S2.

As a cross-check, we use the independently measured trigger delay between A-C, and verify that the time delay between the routes B-C and A-B-C matches the expected value. Any potential instability in the trigger delays during

TABLE S1. Coordinates of the relevant locations. RD XY coordinates are Dutch RD coordinate system [36], Z coordinates are NAP according to the NLGEO2004 definition [36]. Latitude, longitude and height are calculated WGS84, using RDNAPTRANS$^{TM}$(2008) transformations.

| Location | RD X (m) | RD Y (m) | NAP height (m) |
|---|---|---|---|
| | latitude (°) | longitude (°) | geometric height (m) |
| A | 85441.0 | 446371.7 | -1.75 |
| | (N 52.001358) | (E 4.374233) | (41.7) |
| B | 85920.3 | 445185.0 | -3.43 |
| | (N 51.990753) | (E 4.381451) | (40.0) |
| C | 85490.9 | 445881.5 | -3.41 |
| | (N 51.996959) | (E 4.375059) | (40.1) |

the measurement would immediately be detected at the time-tagger at C by a shift in the arrival time of the laser pulse-photons coming from A and B.

TABLE S2. Trigger delays between lab locations. For A-C, the AWG trigger delay is included. The given uncertainty values assume a worst-case scenario, where all individual uncertainties add up linearly.

| Trigger delays | Time (ns) | Uncertainty (ns) |
|---|---|---|
| A - B | 9347 | 10 |
| B - C | 5526 | 8 |
| A - C | 3391 | 8 |

## I. Random number generation

We use two accelerated laser phase diffusion quantum random number generators [37] (QRNGs) of identical construction, designed and built at ICFO - The Institute of Photonic Sciences. The design, modelling, and testing of these devices is described in detail in [30]. As described there, each QRNG continually generates partially random "raw" bits at a rate of 200 MHz and performs a running parity calculation to output processed bits that aggregate the randomness from all previous raw bits. At the time an output bit is taken for use as a measurement setting, only the most recent $k$ raw bits will still be space-like separated from the distant measurement station, and thus only these $k$ bits contribute local randomness. Based on the timing diagram shown in Figure 2, and subtracting delays in the QRNG device (10 ns), internal delays of the sampling FPGA (30 ns), and cable delay to the time-taggers (10 ns), the window for generation of space-like separated raw bits is 160 ns, and thus $k = 32$.

The predictability $\mathcal{P}$ of the output (ideally $\mathcal{P} = 1/2$ for a perfect random bit source) is $\mathcal{P} \leq \frac{1}{2} + \tau^{(k)}$, where $\tau^{(k)}$ bounds the excess predictability of the extracted bit. Due to the parity calculation, $\tau^{(k)}$ decreases by roughly a factor of ten for each increment of $k$, reaching $\tau^{(k)} \leq 10^{-5}$ for $k \geq 6$. The model uses a $6\sigma$ bound on untrusted noise sources and "fully paranoid" assumptions about how these noises combine [30].

While the derived predictability errors $\mathcal{P}$ for $k = 32$ are too small to be tested (verifying this would require an impractical quantity of data), we can verify the validity of the predictions up to a given statistical significance. Prior to the experiment the QRNGs were tested at $k = 4$, achieved by keeping only every fourth output bit. Note that $k = 4$ output should, by the same model, have predictability errors $\leq 10^{-4}$. We applied the test suites NIST SP800-22 (1.5 Gb for each QRNG) and more extensively TestU01 Alphabit battery (210 Gb and 255 Gb), always finding results consistent with ideal randomness. The largest files tested contained $2^{33} \approx 8$ Gb and $2^{34} \approx 17$ Gb, respectively. Using the statistical uncertainty of a test of length $2^{33}$, we obtain a $2\sigma$ error bound of $\mathcal{P} < \frac{1}{2} + 2\frac{1}{2\sqrt{2^{33}}} = \frac{1}{2} + 1.08 \times 10^{-5}$. We use this latter number as the predictability error $\tau$ in the $P$-value calculation, but note that this is conservative as the theoretical model predicts that the $k = 32$ predictability error is smaller.

During the experiment, the continued performance of the QRNGs was monitored in two ways: First, the regulated reference voltage of the comparator after the pin photo-diode (see Figure 1 of [30]) was checked every 10 seconds to exclude failure of the phase-locked loop, laser, interferometer, and photo-detector. Second, we recorded a subset of $2^{23}$ random bits, generated during the course of the Bell experiment, observing means of $\frac{1}{2} + 5.8 \times 10^{-5}$ and $\frac{1}{2} + 1.3 \times 10^{-4}$, consistent with the zero bias of a properly functioning device, considering the statistical uncertainty of $\frac{1}{2\sqrt{2^{23}}} = 1.7 \times 10^{-4}$.

### J. Experimental control and stability

A typical experimental run proceeds as follows (see Figure S5): Both Adwin controllers perform a charge-resonance check (CR check, see below). Adwin A sends a trigger signal to B when the CR-check is successful. When Adwin B is also ready, it triggers AWG B, which in turn sends out an initial 'sync-pulse' signal, to trigger AWG A. AWG B waits for the pre-calibrated A-B trigger delay (see Table S2), and then proceeds with 250 repetitions of the programmed pulse sequence (Figure S1). Similarly, AWG A runs 250 repetitions of its sequence, as soon as it receives the initial sync-pulse. When finished, both AWGs trigger their Adwin to start the next CR check. This cycle repeats itself until the run is quit manually, or until the 45 minute run time is over.

A major challenge for our experiment is to keep a complex setup with many critical components stable over the long integration time necessary to collect data, in order to ensure entanglement generation, spin rotations and spin readout with a high fidelity that is required to observe a violation of the Bell inequality. To maintain stable operation for a period of several weeks, we employ a combination of pre-selection, parameter feedback, periodic calibration and optimisation measurements. Additionally, we monitor various key experimental indicators while running the Bell experiment. If an indicator is outside of the pre-defined range, the current run is either stopped completely, or a signal is recorded in the data to indicate that the setup is in an invalid state (see section K). Note that if this happens, no data from previous trials or the current trial is discarded (only trials that were marked as invalid in advance are affected).

Below we list all aspects of the setups that were actively stabilized and monitored, from short to long time-scale.

- Every 250 entanglement attempts (3.75 ms), we ensure that both NVs are negatively charged and on resonance with the excitation lasers (CR-check, see [27]). We excite both the $|\mathrm{m}_s = \pm 1\rangle \leftrightarrow |E'\rangle$ and the $|\mathrm{m}_s = 0\rangle \leftrightarrow |E_x(E_y)\rangle$ optical transitions during 50 $\mu$s for setups A (B), while counting the number of detected photons. If no photon is detected, the NV centre is assumed to be in its neutral charge state $\mathrm{NV}^0$ and a yellow laser, resonant to the $\mathrm{NV}^0$ zero-phonon line, is applied to transfer the centre back to $\mathrm{NV}^-$. If, however, the number of photons exceeds a pre-set threshold (between 30-40 counts), the experiment proceeds with the entanglement sequence. Otherwise, the experiment repeats the check, until the threshold is met.

- An automated feedback loop stabilizes the NV excitation and emission frequency to the two red excitation lasers, by tuning the d.c. electric field applied to the gold gate-electrodes. Before each CR check, the arrival of photon detections are correlated with a 0.1 V, 50 $\mu$s period modulation of the gate electrode. This signal is averaged over about 1000 CR checks (few seconds), and serves as the error signal for the feedback. In a similar manner the laser frequency of the yellow laser is stabilized to the $\mathrm{NV}^0$ zero-phonon line, using a modulation of the AOM driving frequency, while counting photon detections during the yellow repump phase.

- Every 10 seconds, both setups check proper functioning of a few key devices (AWG, wavemeters, laser stabilisation). If a check fails, the run is stopped and the setup waits for human intervention. Additionally, the control program monitors:

  - The average counts during the CR-check. If too low, we assume that the automated gate voltage feedback has failed. A slow scan of the gate voltage, spanning a larger range than the feedback loop, is made to find back the optimum point. During the scan, the recorded attempts are marked invalid. If still too low, or too high, we assume that the NV dipole polarization has drifted. This requires the excitation polarisation to be adjusted manually.

  - The average counts obtained during the yellow repump phase. If too low, we assume the automated yellow laser frequency feedback has failed. A slow sweep of the yellow laser frequency is made, and the optimum point is set. During the scan, the recorded attempts are marked invalid.
  - The strain splitting, calculated from the current red laser frequencies. If this splitting is not within the predefined range (1.5 - 1.8 GHz for NV A, 1.5 - 2.4 GHz for setup B), the subsequent data is marked invalid until the strain splitting is within range again. This is typically achieved by manually tuning the readout laser excitation frequency.
  - The amount of unwanted reflections of the excitation light from the surface of the diamond. If this is above a predefined threshold, the setup performs an automated scan of the waveplate angles in the ZPL emission path (see Figure 1c) to improve the polarisation rejection of the excitation laser. The reflections are measured for both setups independently by integrating a specific time-window of the arrival of photons at location C. During the spin-initialisation part of the protocol, we identify parts in the histogram for which only photons from one of the setups arrive. If the amount of reflections integrated in a time-window corresponding to the arrival of the 2 nanosecond laser excitation pulse is higher than the integrated NV emission, the subsequent attempts are marked as invalid.

- Every 5 minutes, we scan the spin-pumping laser, and set it to the point producing optimal CR-counts. The data recorded during the scan is marked invalid.

- After each 45-minute run, the spatial positions of both objectives are optimized to ensure reproducible NV excitation and photon collection. Furthermore, the laser powers are recalibrated.

- Every day, we perform a spin-photon correlation measurement as presented in Figure 3a. If the obtained correlations show errors higher than expected, the microwave pulses for spin rotations are recalibrated. We also check the performance of the final basis rotation (fast microwave switch), spin-coherence and spin readout. Finally, we check the photon detection probabilities per optical excitation at location C, from both setups. If the detection probability drops below $2.3 \times 10^{-4}$ ($1.4 \times 10^{-4}$) for setup A (B), we re-align the detection optics of the ZPL emission.

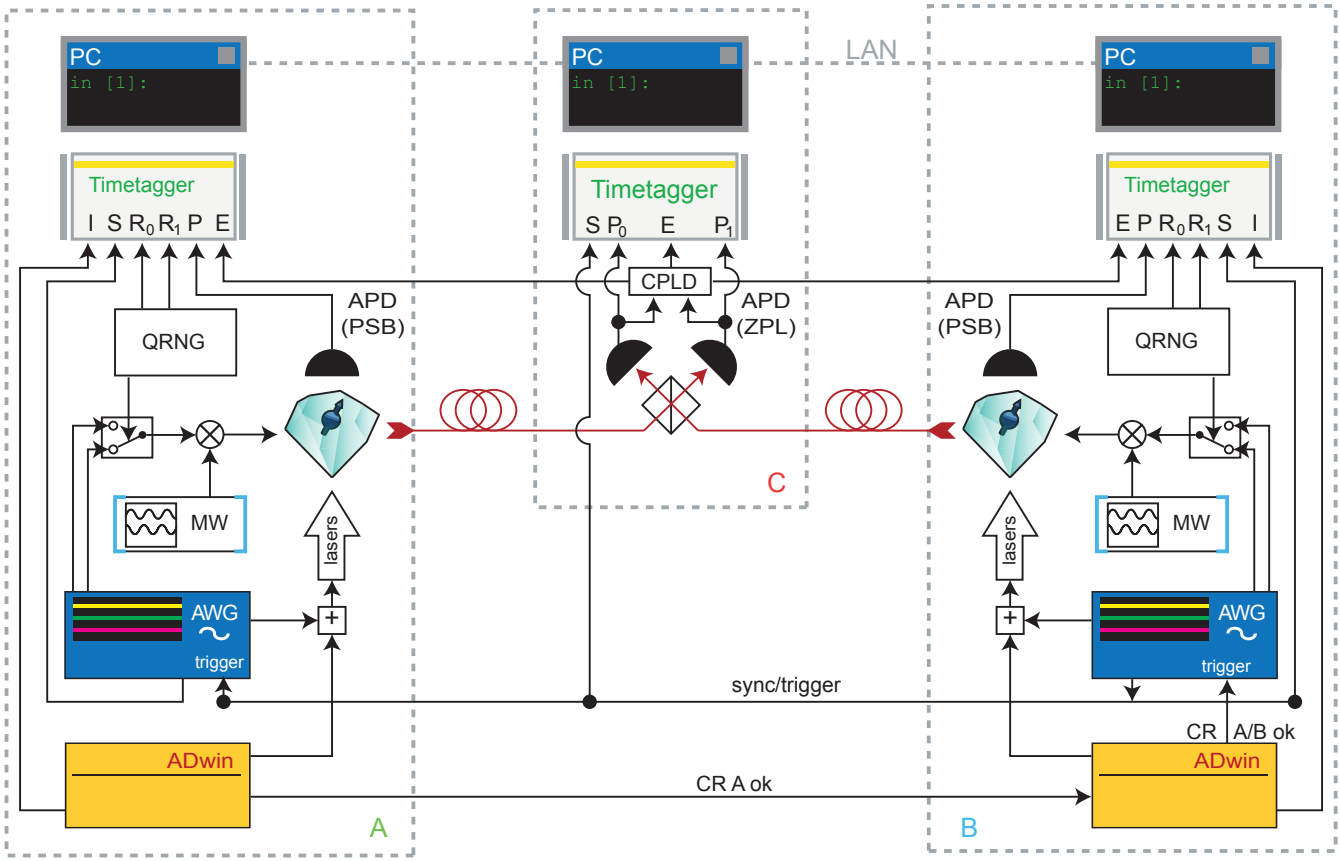


FIG. S5. Control schematic of the experiment. The experiment is performed in three separate laboratories A, B and C (grey dashed lines). The experiments at A and B use a real-time control unit (Jäger ADWIN ProII). Each repetition starts by repeatedly checking whether the NV emission frequency is on resonance with the excitation lasers (CR). Once this check is successful on setup A, its ADWIN passes a signal to setup B. When B has also successfully passed the test, ADWIN B triggers AWG B, which in turn triggers AWG A. Both AWGs then output a synchronization signal (S) which is recorded by their respective time taggers, whose signal is read out in blocks using a PC. The AWGs also output the fast control sequence that is applied to the NV centres. Towards this end, they control the laser pulses and output two different microwave (MW) pulses that perform the basis rotations in the Bell experiment. A microwave switch controls which of these two pulses is applied to the NV centre, depending on the output of the QRNG. The value of the generated random bit (R0 or R1) is recorded by time-tagging devices, alongside with the NV spin state readout signal (P) that is detected by single photon detectors (APD). At location C, two APDs detect upcoming ZPL photons, whose arrival time is recorded by another time-tagger. Upon consecutive detection of two photons, a programmable logic device (CPLD) outputs a signal that is used to mark these instances in the time-tagging data (input E). This allows the PCs at A and B to directly discard experimental runs in which no entanglement was generated, thus keeping the amount of stored data in reasonable bounds.

### K. Data recording and processing

The time-tagging devices at locations A and B record the following events:

- $R_0$: a 'zero' was generated by the QRNG;
- $R_1$: a 'one' was generated by the QRNG;
- $P$: a click was detected in the local phonon-side band emission;
- $E$: an entanglement heralding signal was received from the CPLD at location C;
- $I$: One of the benchmark indicators (section J) is outside of the pre-defined range, which causes the Adwin to send a pulse to this channel every 250 entanglement attempts.

For each of these recorded events, the total number of sync-pulses ($S$) detected since the start of the run is saved, as well as the time passed since the last sync-pulse. Every few hundred milliseconds, the recorded events are transferred to the PC. During the experiment, about 2 megabyte of data is generated every second. To keep the size of the generated data-set manageable, blocks of about 100000 events are saved to the hard drive only if an entanglement heralding event ($E$) is present in that block.

The time-tagging device at location C records the following events:

- $P_0$: a click was detected in output port 0 of the beam-splitter;
- $P_1$: a click was detected in output port 1 of the beam-splitter;
- $E$: an entanglement heralding signal was received from the CPLD;

For each of these events, the current number of detected sync-pulses ($S$) since the start of the run is recorded, as well as the time passed since the last sync-pulse and the total time passed since the start of the run. As the data-rate is orders of magnitude lower here, all data is saved to the hard-drive.

For each run, we first extract the data of those attempts where entanglement was successfully generated, i.e. all data that has the same sync-pulse number as an entanglement heralding signal $E$. We then determine if this is a valid trial of the Bell experiment, determined by following criteria:

1. At location C, the recorded photon arrival times were within the bounds of the pre-defined filters (see section E).
2. At location A and B, there was no invalid marker $I$ recorded in one of the previous 250 attempts.
3. At location A and B, no local photon $P$ was recorded in a time-window of 200 ns after the two optical excitations for entanglement generation.

We emphasize that all events that determine whether or not an iteration is a valid Bell trial either are recorded in the past light-cone of the random basis choice (2. and 3.) or are space-like separated from the random basis choice (1.) of both A and B.

All rounds that fulfil the above criteria are a Bell trial, corresponding to $t = 1$, in the language of the statistical analysis presented in Section N below. For these events, we extract the generated random numbers and readout results of both setups A and B. For setup A, an event on channel $R_0$ corresponds to $a = 0$, an event on $R_1$ indicates $a = 1$. The readout result $x = +1$ is assigned if within the readout integration time-window (as defined by the spacetime analysis presented in the main text), there is at least one readout detection event (channel $P$) recorded and $x = -1$ otherwise. For setup B, $b$ and $y$ are assigned analogous.

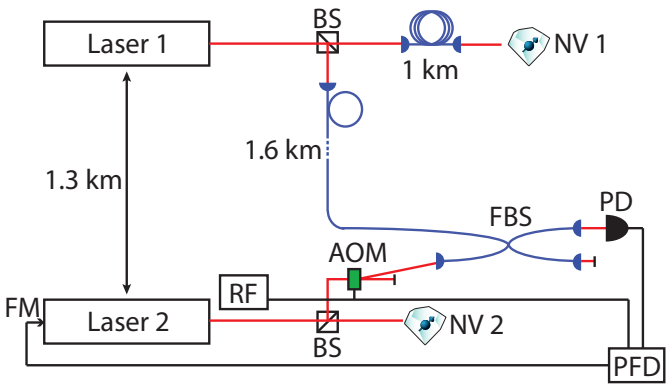


FIG. S6. Schematic of the relative laser frequency stabilization. FM: Laser frequency modulation input. RF: Radiofrequency source. BS: Beam splitter. FBS: Fibre beam splitter. PD: Photodiode. PFD: Phase-frequency detector. NV: Nitrogen-vacancy centre. AOM: Acousto-optical modulator.

### L. Stabilization of the excitation laser frequency

In our previous demonstration of entanglement between remote spins separated by 3 meters [29], both NV defects were excited by the same laser, and the difference in optical path length from this laser to either of the NV centres was much smaller than the coherence length of the laser. Therefore, the relative phase between early and late time-bin of the photonic wavepacket could be described in a common rotating frame - that of the laser source. In the present experiment, the large separation between the two setups requires excitation by two independent laser systems. Retaining the common reference frame upon interference of the time-bin encoded photonic qubits requires that the source lasers in both setups imprint the same phase difference between early and late time bin. To accomplish this, a digital feedback loop is used to stabilize the frequency of one laser to that of the other.

Fig. S6 shows a schematic of this feedback loop. A small fraction of the excitation laser of setup A is frequency-shifted by 135 MHz by an acousto-optical modulator. Using a fibre-based beamsplitter, this laser field is interfered with a part of the light of laser B, which has been transmitted to laboratory A using an additional optical fibre of 1.6 km length.

To stabilize the relative frequency of the lasers, the interference signal is recorded using a fast amplified photodetector (Thorlabs PDA10). The beat frequency is compared to the RF source signal that drives the AOM using a digital phase-frequency detector (Menlo systems DXD200). The output of this device is applied to the current of laser diode A, shifting its frequency and thus closing the feedback loop. The remaining 200 kHz (FWHM) relative frequency deviation is an effect of the limited short-term stability of the used diode lasers (Toptica TA-SHG pro). An additional fibre delay line of 1 km length reduces the effect of drifts in the frequency of laser B on the relative frequency difference of the pulses that excite the NV centres.

In this configuration, we measured the fluctuations in the phase of the driving lasers using an analog phase detector to be about $0.2\pi$ FWHM at the 250 ns time difference between the two excitation pulses. According to our calculations, this should lead to a fidelity reduction of the XX correlations of about 3%. Further improvement would be possible by reducing the laser linewidth, e.g. with an external reference cavity, or by decreasing the temporal separation between the two time-bins.

### M. Adaptive optics

We use a deformable mirror (Boston Micromachines) in the ZPL detection path of setup A to compensate for optical aberrations due to fabrication imperfections of the solid immersion lens. To optimize the mirror surface, the NV centre is off-resonantly excited by a 532-nm green laser and the collected fluorescence signal is optimized by step-wise varying the mirror surface according to Zernike polynomials up to eleventh order [38]. The optimized shape of the mirror

surface is represented in the inset of Figure S7. The increase in collection efficiency is characterized by comparing saturation curves measured locally on the setup A for a flat and optimized mirror configuration, as shown on Figure S7. The evolution of the ZPL count rates versus the excitation power $P_{\mathrm{e}}$ is fit to the function :

$$A_{sat}\frac{P_{\mathrm{e}}}{P_{\mathrm{e}}+P_{\mathrm{sat}}} \tag{3}$$

with the maximum count rate $A_{sat}$ and the saturation power $P_{\mathrm{sat}}$ as free parameters. We find $A_{sat} = 18(1)$ kcts when the mirror surface is flat, and $A_{sat} = 35(2)$ kcts after optimizing. Thus, for this particular solid immersion lens, the use of adaptive optics allows for an increase in collection efficiency by a factor of 1.9(1).

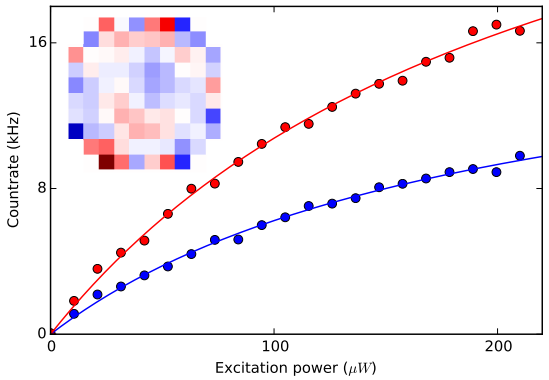


FIG. S7. Saturation curve. The NV is excited with a green laser of varying power and the resulting emission into the zero-phonon line is measured with a flat (blue data) and with an optimized (red data) mirror surface. From the fit curves (red and blue lines), we deduce an increase in collection efficiency by a factor of 1.9. The inset shows a heat map of the actuator voltages, where red (blue) rectangles denote positive (negative) excursion.

## STATISTICAL ANALYSIS

Local hidden variable models (LHVM) predict concrete limitations on the statistics that can be observed in a Bell experiment. These are typically phrased in terms of probabilities or expectation values. Naturally, however, in any experiment we can only observe a finite number of events, and not probabilities. We thus need to quantify the statistical evidence against an LHVM given a finite number of events. While we will focus on LHVM below, we remark that we obtain the same statistical evidence against any theory of nature for which a bound as stated in Lemma 1 (in essence, the CHSH inequality) holds for our experiment.

A common way to analyse statistics in Bell experiments is to compute the number of standard deviations that separate the observed data from the best LHVM. However, this method has well known flaws [16],[21],[39],[40] (see [40] for a detailed discussion). In particular, we would have to assume Gaussian statistics and independence between subsequent attempts, allowing for the memory loophole. Fortunately, it is possible to rigorously analyze the statistics even when allowing for memory as was first done by Gill [41]. Instead of the standard deviation, intuitively, one bounds the probability of observing the experimental data, if nature was indeed governed by an LHVM. In the language of hypothesis testing, this is known as the $P$-value, where the null hypothesis is that the experiment can be modelled as an LHVM (see e.g. [42]). Informally, we thus have

$$P\text{-value} = \max_{\mathrm{LHVM}} \Pr[\text{data at least as extreme as observed} \mid \text{experiment is governed by LHVM}] \ . \tag{4}$$

A small $P$-value can be interpreted as strong evidence against the null hypothesis, that is, that the experiment was governed by an arbitrary LHVM. There is an extensive literature regarding methods for evaluating the $P$-value in Bell experiments [16],[21],[42],[39],[40, 41, 43–51] and discussions regarding the analysis of concrete experiments and loopholes [12],[52–58].

Here, we focus on the case of the CHSH inequality as relevant for our experiment. For the simple case of the CHSH inequality, it has been shown how to derive tight bounds on the $P$-value if the settings are chosen uniformly. Such a bound was first informally derived in [16], and later rigorously developed by Bierhorst [39] whose approach we will follow closely with two modifications: First, the analysis of [39] was done for uniform and independent choices of measurement settings. We extend the approach of [39] to the case where A and B use a partially predictable RNG [30]. Second, although our experiment can readily be seen to correspond to an event-ready scheme in the sense of Bell, we formally include the event-ready procedure in our analysis.

## N. How to compute the $P$-value

See Figure 1a; an event-ready Bell experiment consisted of $m$ entanglement attempts. Let $\mathbf{t}^m = (t_i)_{i=1}^m$ denote the output signal of the "event-ready"-box, where the tag $t_i = 0$ corresponds to a failure (no, not ready) event, and $t_i = 1$ to a successful preparation (yes, ready) of the boxes A and B [59]. We will reserve the word *trial* for the attempts that correspond to a successful preparation. Throughout, we use superscripts $m$ to remind ourselves a sequence $\mathbf{t}^m$ has $m$ elements. Let $\mathbf{a}^m = (a_i)_{i=1}^m, \mathbf{b}^m = (b_i)_{i=1}^m$ denote the inputs to boxes A and B in Figure 1a, where $a_i, b_i \in \{0, 1\}$. Furthermore, let $\mathbf{x}^m = (x_i)_{i=1}^m, \mathbf{y}^m = (y_i)_{i=1}^m$ with $x_i, y_i \in \{\pm 1\}$ denote the output of boxes A and B.

We denote by $|\mathbf{t}^m|$ the number of ones in the binary sequence $\mathbf{t}^m$. Since we will use only the attempts where $t_i = 1$, let us define

$$n := |\mathbf{t}^m| = \sum_{i=1}^{m} t_i \tag{5}$$

to be the number of attempts in which $t_i = 1$, where $n$ is fixed as discussed below. Given the observed values $\mathbf{a}^m, \mathbf{b}^m, \mathbf{x}^m, \mathbf{y}^m$, and $\mathbf{t}^m$ recorded in the experiment, we can compute the CHSH function

$$k := \sum_{i=1}^{m} t_i \cdot \frac{(-1)^{a_i b_i} x_i y_i + 1}{2} \ . \tag{6}$$

Note that $k$ is the number of times that $(-1)^{a_i b_i} x_i y_i = 1$. When viewing CHSH as a non-local game [13], $k$ is thus the number of times that Alice and Bob win the CHSH game. For large $n$, and uniform distribution of the inputs, we have $k \approx n(S + 4)/8$, where $S$ is the value of the familiar CHSH correlator. We prove that

$$P\text{-value} \leq P_{n,k}(\mathbb{B}_\xi) = \sum_{j=k}^{n} \binom{n}{j} \xi^j (1 - \xi)^{n-j} \ , \tag{7}$$

where $P_{n,k}(\mathbb{B}_\xi)$ is the probability that $n$ i.i.d. (independent and identically distributed) Bernoulli trials with probability

$$\xi = 3/4 + 3(\tau + \tau^2) \tag{8}$$

have at least $k$ successes. Here $\tau$ denotes a partial predictability of the inputs given the history of the experiment, defined in (11) and (12) below, where we take $\tau = \max\{\tau_A, \tau_B\}$. The value for $\tau$ used in the experiment is given in Section I. We remark that $n$ is fixed in this analysis, i.e., we stop the experiment if a certain number $n$ of trials have been collected, where $n$ was decided independently of the data observed.

Even though we allowed that the LHVM could depend on previous attempts, thus making no extra assumptions on the memory of the devices, the upper bound is the tail probability of an i.i.d. distribution. This is not at all uncommon for sums of random variables, and there are many other examples where such a simplification occurs (see for instance [60] and [61]).

## O. Properties of the tested models

We introduce the sequence of random variables $(A_i, B_i, X_i, Y_i, T_i, H_i)_{i=1}^m$ in correspondence with the concrete outcomes as described above, where $i$ is used to label the $i$-th element: let $\mathbf{A}^m = (A_i)_{i=1}^m, \mathbf{B}^m = (B_i)_{i=1}^m$ denote the inputs

to the boxes, $\mathbf{X}^m = (X_i)_{i=1}^m, \mathbf{Y}^m = (Y_i)_{i=1}^m$ the outputs of the boxes, $\mathbf{T}^m = (T_i)_{i=1}^m$ the sequence of event-ready signals, and $\mathbf{H}^m = (H_i)_{i=1}^m$ the histories of attempts previous to the $i$-th attempt. We make no assumptions regarding the statistics of the event-ready procedure, which may be under full control of the LHVM, and can depend arbitrarily on the history of the experiment. The random variable $H_i$ models the state of the experiment prior to the measurement. As such, $H_i$ includes any hidden variables, usually denoted using the $\lambda$ symbol [13]. It also includes the history of all possible configurations of inputs and outputs of the prior attempts $(A_j, B_j, X_j, Y_j, T_j)_{j=1}^{i-1}$.

We consider all models that restrict the random variables in the following way:

1. *Local randomness generation.* Conditioned on the history of the experiment the inputs $A_i, B_i$ are independent of each other

$$\forall i, A_i \perp\!\!\!\perp B_i \mid H_i \ , \tag{9}$$

and of the output of the event-ready signal

$$\forall i, A_i \perp\!\!\!\perp T_i, B_i \perp\!\!\!\perp T_i \mid H_i \ . \tag{10}$$

We allow $A_i$ and $B_i$ to be partially predictable given the history of the experiment:

$$\forall (i, a_i, h_i), \Pr\left(A_i = a_i | H_i = h_i\right) \leq \frac{1}{2} + \tau_A \ , \tag{11}$$

$$\forall (i, b_i, h_i), \Pr\left(B_i = b_i | H_i = h_i\right) \leq \frac{1}{2} + \tau_B \ . \tag{12}$$

2. *Locality.* The outputs $x_i$ and $y_i$ only depend on the local input settings and history: they are independent of each other and of the input setting at the other side, conditioned on the previous history and the current event-ready signal:

$$\forall i, (X_i, A_i) \perp\!\!\!\perp (Y_i, B_i) | H_i, T_i \ . \tag{13}$$

3. *Sequentiality of the experiments.* Every one of the $m$ attempts takes place sequentially such that any possible signalling between different attempts beyond the previous conditions is prevented [16].

Except for these conditions the variables might be correlated in any possible way.

Given the experimental setup and sequence as described in the main text and Figures 1 and 2, all local realist theories that predict that the number generators produce a free random bit in a timely manner, and that the output is final once recorded in the electronics, satisfy these restrictions [1].

## P. Proof outline

Let us first provide a sketch of the main steps before giving detailed proofs, where we will use the notation and definitions introduced above. Let $\Delta_i^{a,b,x,y,t}$ be an indicator function

$$\Delta_i^{a,b,x,y,t} = \mathbb{1}\{A_i = a, B_i = b, X_i = x, Y_i = y, T_i = t\} \ . \tag{14}$$

That is, $\mathbb{1}\{A_i = a, B_i = b, X_i = x, Y_i = y, T_i = t\}$ is itself a random variable that is a function of the random variables $A_i$,$B_i$,$X_i$,$Y_i$ and $T_i$. It takes on the value 1 if all equalities are satisfied for a particular choice of $a, b, x, y, t$, and 0 otherwise. Next, we define the Bell random variable associated with the $i$-th attempt as

$$C_i = \sum_{a,b,x,y,t} \Delta_i^{a,b,x,y,t} \cdot t \cdot \frac{(-1)^{ab} xy + 1}{2} \ . \tag{15}$$

The concrete instance of the $i$-th attempt we denote by

$$c_i = c_i(a_i, b_i, x_i, y_i, t_i) = t_i \cdot \frac{(-1)^{a_i b_i} x_i y_i + 1}{2}, \tag{16}$$

which is a function that can be computed using inputs and outputs of the three boxes in Figure 1a, in the $i$-th attempt. As indicated, we will often drop the dependence on $(a_i, b_i, x_i, y_i, t_i)$ by writing $c_i$ in order to lighten the notation. The term concrete instance means that $c_i$ can be computed from the data observed in the experiment.

The main steps of the arguments are as follows. First we need to introduce some notation. We define the random variable

$$Z^m = \sum_{i=1}^{m} C_i \ . \tag{17}$$

Note that $k = z^m$ is a concrete instance of $Z^m$. As above, we will often drop the dependence on $\mathbf{A}^m$,$\mathbf{B}^m$, $\mathbf{X}^m$, $\mathbf{Y}^m$, $\mathbf{T}^m$ and simply write $Z^m$.

The $P$-value is the probability that an any LHVM can produce a result at least as extreme as the data observed, that is, that it produces a value equal or larger than $k$ over $n$ trials, i.e., $Z^m \geq k$. Since we run the experiment until a particular number of trials $n$ are recorded, and the LHVM has control over the event-ready signal, this means that we also have to consider the probability that the LHVM produces such $k$ for a larger (or smaller) number of attempts than actually observed in the experiment. To formally model this, note that stopping when $n$ trials (i.e., successful heralding attempts) have occurred is equivalent to saying we ignore all future attempts. That is, our implementation sets $t_j = 0$ for all $j > \ell$, where $\ell$ denotes the attempt in which we observed the $n$-th trial. Nevertheless, the total number of attempts $m \geq n$ that an LHVM can use to produce at least $k$ in $n$ trials is allowed to be arbitrarily large. Below we will establish the following upper bound on the $P$-value in a series of steps, where the maximization is taken over all possible LHVM (including $m$). The first equality (18) is just the definition of the $P$-value.

$$P\text{-value} = \max_{\text{LHVM}} \sum_{\substack{\mathbf{t}^m \in \{0,1\}^m \\ |\mathbf{t}^m| = n}} \Pr\left(\mathbf{T}^m = \mathbf{t}^m\right) \Pr\left(Z^m \geq k \mid \mathbf{T}^m = \mathbf{t}^m\right) \tag{18}$$

$$= \max_{\text{LHVM}} \sum_{\substack{\mathbf{t}^m \in \{0,1\}^m \\ |\mathbf{t}^m| = n}} \Pr\left(\mathbf{T}^m = \mathbf{t}^m\right) \Pr\left(\text{number of 1's in } \ (C_1 \cdot t_1, \ldots, C_m \cdot t_m) \geq k \mid \mathbf{T}^m = \mathbf{t}^m\right) \tag{19}$$

$$\leq P_{n,k}(\mathbb{B}_\xi) \ . \tag{20}$$

Recall that $n$ is fixed in this analysis, i.e., we stop the experiment [62], if a certain number $n$ of trials have been collected, where $n$ has to be decided independently of the data observed. Note that we cannot stop the experiment if the $P$-value drops below a desired limit, or even perform the experiment multiple times until reaching a desired $P$-value. Otherwise, the $P$-value would have to be computed in a completely different form.

Viewing CHSH from the perspective of non-local games makes our proof very intuitive: $k$ can be understood as the number of times that Alice and Bob win the CHSH game using a local-hidden variable strategy [13]. The $P$-value is then the probability that they win the game at least $k$ times if we perform an experiment in which we wait until $n$ successful heralding attempts have occurred, maximized over all possible local-hidden variable strategies. In our proof, we first bound the probability that they win in round (attempt) $i$, conditioned on the entire history leading up to this round, by a number $\xi$. Finally, we bound the probability that they win at least $k$ times on $n$ rounds by an inductive argument: intuitively, we peel off round by round of the game, where at each peel we apply the analysis for round $i$. In other words, the high-level overview of the steps to obtain the bound (20) from (19) is as follows:

*Step 1: Bounding the probability of one attempt given the past history.*
The first step is to bound $\Pr\left(C_i = 1 \mid H_i = h_i, T_i = t_i\right)$ for attempt $i$, for any possible history $h_i$ and event-ready output $t_i$. Intuitively, this can be understood as a derivation of the CHSH inequality, using the properties of local randomness generation, locality and sequentiality. We first expand the desired term using the definition of $C_i$ in (15)

as

$$\Pr\left(C_i=1|H_i=h_i,T_i=t_i\right)=\sum_{\substack{z\in\{-1,1\}\\ a,b\in\{0,1\}\\ (a,b)\neq(1,1)}}\Pr\left(X_i=Y_i=z,A_i=a,B_i=b|H_i=h_i,T_i=t_i\right)$$
$$+\sum_{z\in\{-1,1\}}\Pr\left(-X_i=Y_i=z,A_i=1,B_i=1|H_i=h_i,T_i=t_i\right)\ . \tag{21}$$

The crucial step is now to use the condition of local randomness generation and locality of measurement outcomes to break these probabilities into simpler terms

$$\begin{aligned}&\Pr\big(X_i=x,Y_i=y,A_i=a,B_i=b|H_i=h_i,T_i=t_i\big)=\\&\quad\Pr\left(A_i=a|H_i=h_i,T_i=t_i\right)\Pr\left(B_i=b|H_i=h_i,T_i=t_i\right)\\&\quad\cdot\Pr\left(X_i=x|A_i=a,H_i=h_i,T_i=t_i\right)\Pr\left(Y_i=y|B_i=b,H_i=h_i,T_i=t_i\right)\ .\end{aligned}$$

Plugging this decomposition into (21) and solving the resulting optimization problem yields

$$\Pr\left(C_i=1|H_i=h_i,T_i=t_i\right)\leq\xi\ , \tag{22}$$

for $\xi=3/4+3(\tau+\tau^2)$.

*Step 2: Replacing the history with the recorded values of* $\mathbf{C}^{i-1}$ *and* $\mathbf{T}^i$. Above, we allowed for arbitrary histories $h_i$ that include any state of the experiment. The next step is to replace the conditioning on the history by a conditioning on the recorded sequence up to the $i$-th attempt $\mathbf{C}^{i-1}=(C_i)_{j=1}^{i-1}$, and the event-ready output $\mathbf{T}^i=(T_j)_{j=1}^{i}$, instead of the entire history, giving us

$$\Pr\left(C_i=1|\mathbf{T}^i=\mathbf{t}^i,\mathbf{C}^{i-1}=\mathbf{c}^{i-1}\right)\leq\xi\ . \tag{23}$$

This is a consequence of (22) together with the law of total probability, which in its discrete form says that the probability of any event $Z=z$ can be written as $\Pr\left(Z=z\right)=\sum_{z'}\Pr\left(Z'=z'\right)\Pr\left(Z=z\mid Z'=z'\right)$ for any distribution $\Pr\left(Z'=z'\right)$. We remark that steps 2 and 3 only use (22), and do not depend on the form of the null hypothesis.

*Step 3: Going from one to many attempts* The final step is to use this bound, to bound the $P$-value for $m$ attempts with $|\mathbf{t}^m|=n$, where $m$ can be arbitrarily large. This follows from an inductive argument as in Bierhorst [39]. When performing the induction step, we bound the next step using (23).

## Q. Technical details

*Step 1: Bounding the probability of one attempt given the past history (CHSH inequality).*

First, in Lemma 1, we prove that if the experiment is ruled by an LHVM, then the probability that $C_i$ takes the value one given the state of the experiment is bounded from above by some $\xi$ for any possible history $h_i$ and event-ready attempt $t_i$. This bound is basically a rederivation of CHSH with imperfect random number generators.

**Lemma 1.** *Let $m\in\mathbb{N}$, and let the sequence $(\mathbf{A}^m,\mathbf{B}^m,\mathbf{X}^m,\mathbf{Y}^m,\mathbf{H}^m,\mathbf{T}^m)$ be defined as in Section O. Suppose that the null hypothesis holds, i.e., nature is governed by an LHVM. Given that the predictability of the RNG is $\tau$, we have for all $i\in\mathbb{N}$ with $i\leq m$, any possible history $H_i=h_i$ of the experiment, and all $T_i=t_i$ that the probability of $C_i=1$ is upper bounded by*

$$\Pr\left(C_i=1|H_i=h_i,T_i=t_i\right)\leq\xi\ , \tag{24}$$

*where $\xi=3/4+3(\tau+\tau^2)$.*

*Proof.* We first expand the desired term using the definition of $C_i$ as

$$\begin{aligned}\Pr\left(C_i=1|H_i=h_i,T_i=t_i\right)=&\sum_{\substack{z\in\{-1,1\}\\a,b\in\{0,1\}\\(a,b)\neq(1,1)}}\Pr\left(X_i=Y_i=z,A_i=a,B_i=b|H_i=h_i,T_i=t_i\right)\\&+\sum_{z\in\{-1,1\}}\Pr\left(-X_i=Y_i=z,A_i=1,B_i=1|H_i=h_i,T_i=t_i\right)\ .\end{aligned}\tag{25}$$

We can break these probabilities into simpler terms

$$\begin{aligned}&\Pr\big(X_i=x,Y_i=y,A_i=a,B_i=b|H_i=h_i,T_i=t_i\big)\\&\quad=\Pr\left(X_i=x,A_i=a|H_i=h_i,T_i=t_i\right)\\&\qquad\cdot\Pr\left(Y_i=y,B_i=b|H_i=h_i,T_i=t_i\right)&(26)\\&\quad=\Pr\left(A_i=a|H_i=h_i,T_i=t_i\right)\Pr\left(X_i=x|A_i=a,H_i=h_i,T_i=t_i\right)\\&\qquad\cdot\Pr\left(B_i=b|H_i=h_i,T_i=t_i\right)\Pr\left(Y_i=y|B_i=b,H_i=h_i,T_i=t_i\right)\ .&(27)\end{aligned}$$

The first equality followed by the locality condition, the second one simply by the definition of conditional probability. With this decomposition, we can express (25) as

$$\begin{aligned}\Pr\left(C_i=1|H_i=h_i,T_i=T_i\right)=&\sum_{\substack{a,b\in\{0,1\}\\(a,b)\neq(1,1)}}\alpha_a\beta_b\left(\chi_a\gamma_b+(1-\chi_a)(1-\gamma_b)\right)\\&+\alpha_1\beta_1\left(\chi_1(1-\gamma_1)+(1-\chi_1)\gamma_1\right)&(28)\\=&\sum_{a,b\in\{0,1\}}\alpha_a\beta_b f_{a,b}&(29)\\\leq&\left(\frac{1}{2}+\tau\right)^2\sum_{a,b\in\{0,1\}}f_{a,b}\ ,&(30)\end{aligned}$$

where we have used the shorthands

$$\begin{aligned}\chi_a&:=\Pr\left(X_i=1|A_i=a,H_i=h_i,T_i=t_i\right)\ ,&(31)\\\gamma_b&:=\Pr\left(Y_i=1|B_i=b,H_i=h_i,T_i=t_i\right),&(32)\\\alpha_a&:=\Pr\left(A_i=a|H_i=h_i,T_i=t_i\right)\ ,&(33)\\\beta_{b'}&:=\Pr\left(B_i=b|H_i=h_i,T_i=t_i\right)\ ,&(34)\\f_{a,b}&:=\begin{cases}\chi_a\gamma_b+(1-\chi_a)(1-\gamma_b) & \text{if } (a,b)\neq(1,1)\ ,\\ \chi_a(1-\gamma_b)+(1-\chi_a)\gamma_b & \text{otherwise.}\end{cases}&(35)\end{aligned}$$

It thus remains to bound the sum of $f_{a,b}$. Note that we can write

$$\begin{aligned}\sum_{a,b\in\{0,1\}}f_{a,b}=&\left(\chi_0\gamma_0+(1-\chi_0)(1-\gamma_0)\right)+\left(\chi_0\gamma_1+(1-\chi_0)(1-\gamma_1)\right)\\&+\left(\chi_1\gamma_0+(1-\chi_1)(1-\gamma_0)\right)+\left(\chi_1(1-\gamma_1)+(1-\chi_1)\gamma_1\right)&(36)\\=&\,\chi_0\left(\gamma_0+\gamma_1\right)+\left(1-\chi_0\right)\left(2-\gamma_0-\gamma_1\right)\\&+\chi_1\left(\gamma_0+1-\gamma_1\right)+\left(1-\chi_1\right)\left(1-\gamma_0+\gamma_1\right)\ .&(37)\end{aligned}$$

Since (37) is a sum of two convex combinations, it must take its maximum value at one of the extreme points, that is

with $\chi_0 \in \{0,1\}$ and $\chi_1 \in \{0,1\}$. We can thus consider all four combinations of values for $\chi_0$ and $\chi_1$ given by

$$\sum_{a,b\in\{0,1\}} f_{a,b} = \begin{cases} 3-2\gamma_0 \text{ if } (\chi_0,\chi_1)=(0,0)\ , \\ 3-2\gamma_1 \text{ if } (\chi_0,\chi_1)=(0,1)\ , \\ 1+2\gamma_1 \text{ if } (\chi_0,\chi_1)=(1,0)\ , \\ 1+2\gamma_0 \text{ if } (\chi_0,\chi_1)=(1,1)\ . \end{cases} \tag{38}$$

Since $0 \leq \gamma_0, \gamma_1 \leq 1$, we have in all cases that the sum is upper bounded by 3. Using (30) we thus have

$$\Pr\left(C_i = 1 | H_i = h_i, T_i = t_i\right) \leq \left(\frac{1}{2}+\tau\right)^2 \cdot 3 = \frac{3}{4} + 3\left(\tau+\tau^2\right)\ . \tag{39}$$

□

*Step 2: Replacing the history with the recorded values of $\mathbf{C}^{i-1}$ and $\mathbf{T}^i$.*

Now, building on top of Lemma 1, we prove that the probability that $C_i$ takes the value one given not the entire history, but only the event-ready attempts and the prior sequence, is bounded from above by some $\xi$. While the two statements look very similar, the main difference is that while in Lemma 1 we condition on the entire history $H_i = h_i$, in Lemma 2 we condition on the event-ready successes $\mathbf{T}^i = \mathbf{t}^i$, and the prior sequence $\mathbf{C}^{i-1} = (C_j)_{j=1}^{i-1}$ of data that can actually be observed [63]. Although both statements are similar, it is Lemma 2 that we can easily use in the proof of Lemma 3 to bound the $P$-value by the survival function of a Binomial.

We will need Proposition 1, which is a basic probabilistic statement necessary for Lemma 2. In essence, it is just the measure theoretic version of

$$\Pr\left(A=a\right) = \sum_b \Pr\left(A=a|B=b\right)\Pr\left(B=b\right)\ . \tag{40}$$

We state it for completeness, with the purpose of having the derivation of the bound on the $P$-value as self contained as possible.

**Proposition 1** (Law of total probability)**.** *Let $A, B$ be two random variables on the same probability space $\Omega$ with $\mathbb{E}(|A|) < \infty$. Then the probability of an event $A = a$ admits the following integral form*

$$\Pr\left(A=a\right) = \int_\Omega \Pr\left(A=a|B=b\right) d\mu(b)\ , \tag{41}$$

*for some measure $d\mu$ on $\Omega$.*

**Lemma 2.** *Let $m, n \in \mathbb{N}$ and let the sequence $(\mathbf{A}^m, \mathbf{B}^m, \mathbf{X}^m, \mathbf{Y}^m, \mathbf{H}^m, \mathbf{T}^m)$ be defined as in Section O. Suppose that the null hypothesis holds, i.e., nature is governed by an LHVM. Given that the predictability of the RNG is $\tau$, we have that for all $i \in \mathbb{N}$ with $m \geq n$ and $i \leq m$, and all sequences $\mathbf{t}^i \in \{0,1\}^i$ and $\mathbf{c}^{i-1} \in \{0,1\}^{i-1}$, that the probability that $C_i$ takes the value one satisfies*

$$\Pr\left(C_i = 1 | \mathbf{T}^i = \mathbf{t}^i, \mathbf{C}^{i-1} = \mathbf{c}^{i-1}\right) \leq \xi\ , \tag{42}$$

*where $\xi = 3/4 + 3(\tau+\tau^2)$.*

*Proof.* The following equalities hold from the definition of conditional probability and Proposition 1

$$\begin{aligned} &\Pr\left(C_i = 1 | \mathbf{T}^i = \mathbf{t}^i, \mathbf{C}^{i-1} = \mathbf{c}^{i-1}\right)\Pr\left(\mathbf{T}^i = \mathbf{t}^i, \mathbf{C}^{i-1} = \mathbf{c}^{i-1}\right) \\ &\quad = \Pr\left(C_i = 1, \mathbf{T}^i = \mathbf{t}^i, \mathbf{C}^{i-1} = \mathbf{c}^{i-1}\right) &(43) \\ &\quad = \int_\Omega \Pr\left(C_i = 1, \mathbf{T}^i = \mathbf{t}^i, \mathbf{C}^{i-1} = \mathbf{c}^{i-1} | H_i = h_i\right) d\mu(h_i)\ . &(44) \end{aligned}$$

Let us bound the integrand in the previous equation. We have

$$\begin{aligned}\Pr\left(C_i=1,\mathbf{T}^i=\mathbf{t}^i,\mathbf{C}^{i-1}=\mathbf{c}^{i-1}|H_i=h_i\right)&=\Pr\left(C_i=1,T_i=t_i|H_i=h_i\right)\cdot\delta &(45)\\&=\Pr\left(C_i=1|H_i=h_i,T_i=t_i\right)\Pr\left(T_i=t_i|H_i=h_i\right)\cdot\delta &(46)\\&\leq\xi\Pr\left(T_i=t_i|H_i=h_i\right)\cdot\delta &(47)\\&\leq\xi\Pr\left(\mathbf{T}^i=\mathbf{t}^i,\mathbf{C}^{i-1}=\mathbf{c}^{i-1}|H_i=h_i\right)\ , &(48)\end{aligned}$$

where $\delta$ is a shorthand for

$$\delta=\Pr\left(\mathbb{1}\{\mathbf{T}^{i-1}=\mathbf{t}^{i-1},\mathbf{C}^{i-1}=\mathbf{c}^{i-1}\}=1|H_i=h_i\right)\ . \tag{49}$$

The first equality (45) follows from the fact that $\mathbf{t}^{i-1}$ and $\mathbf{c}^{i-1}$ are events either compatible or incompatible with $h_i$, the second one (46) from the definition of conditional probability, and the inequality (47) from Lemma 1. We now introduce (48) back into (44) to obtain

$$\begin{aligned}&\Pr\left(C_i=1|\mathbf{T}^i=\mathbf{t}^i,\mathbf{C}^{i-1}=\mathbf{c}^{i-1}\right)\Pr\left(\mathbf{T}^i=\mathbf{t}^i,\mathbf{C}^{i-1}=\mathbf{c}^{i-1}\right)\\&\quad\leq\xi\int_\Omega\Pr\left(\mathbf{T}^i=\mathbf{t}^i,\mathbf{C}^{i-1}=\mathbf{c}^{i-1}|H_i=h_i\right)d\mu(h_i) &(50)\\&\quad=\xi\Pr\left(\mathbf{T}^i=\mathbf{t}^i,\mathbf{C}^{i-1}=\mathbf{c}^{i-1}\right)\ , &(51)\end{aligned}$$

where the equality (51) follows from Proposition 1. We complete the proof by cancelling the terms $\Pr\left(\mathbf{T}^i=\mathbf{t}^i,\mathbf{C}^{i-1}=\mathbf{c}^{i-1}\right)$ on the right and left side of the equation above. □

*Step 3: Going from one to many attempts*

We end this technical derivation with Lemma 3, which allows us to put together the statements above: instead of making a statement just about the next attempt, we now make a statement about all attempts together. Our proof is an easy generalization of Proposition 4 in [39] to a setting with imperfect RNGs and the event-ready procedure. What makes this analysis more complicated is the formal treatment of the event-ready procedure. Due to this procedure, we deal with long sequences of attempts, while the computation of the CHSH function actually only depends on the (possibly) much shorter sequence of trials, i.e., the attempts where $t_j=1$. It is intuitive that only the trials are relevant for computing the $P$-value, but to make this precise let us spell out the relation between the two sequences.

Of relevance in the long sequence of $m$ attempts, is the sequence $\mathbf{C}^m=(C_1,\ldots,C_m)$ together with the sequence of event-ready attempts $\mathbf{T}^m=(T_1,\ldots,T_m)$. Recall that the latter tells us which elements of $\mathbf{C}^m$ are of interest, i.e., can at all be non-zero. To reason about the shorter sequence of $n$ trials, let us first introduce some notation. Our goal will be to define a series of random variables $\mathbf{D}^n=(D_1,\ldots,D_n)$ for the short sequence of trials, where intuitively $D_j$ corresponds to the random variable taking value one when the $j$-th event-ready success also results in $C_i=1$ for any corresponding $i$. In other words, we will define $\mathbf{D}^n$ in such a way that instead of worrying about the number of 1's in $(C_1T_1,\ldots,C_mT_m)$ we will be concerned with the number of 1's in $(D_1,\ldots,D_n)$.

To define this formally, we need a way to map the $j$-th trial from the short sequence of trials, to the index $i$ in the longer sequence of attempts. Note that for a particular event-ready sequence $\mathbf{t}^m=(t_1,\ldots,t_m)\in\mathbf{T}^m$, the $j$-the trial is mapped to the smallest index $i$, such that the subsequence $\mathbf{t}^i=(t_1,\ldots,t_i)$ of $\mathbf{t}^m$ has exactly $j$ 1's. Of course, there are many sequences $\mathbf{t}^i\in\mathbf{T}^i$ that have precisely $j$ 1's, where the last element is also a 1, and for all such strings the mapping from $j$ in the sequence of trials, to the index $i$ in the sequence of attempts is the same. Let us thus define

$$\mathcal{T}_{j\to i}=\left\{\mathbf{t}^m=(t_1,\ldots,t_m)\in\{0,1\}^m\mid|\mathbf{t}^m|=n\text{ and }\mathbf{t}^i=(t_1,\ldots,t_i)\text{ satisfies }|\mathbf{t}^i|=j\text{ and }|\mathbf{t}^{i-1}|=j-1\right\}\ , \tag{52}$$

to be the set of all event-ready sequences $\mathbf{t}^m$ for which $j$ is mapped to one particular $i$. By summing over all possible indices $i$ in the long sequence of attempts, we can thus formally define

$$D_j=\sum_{i=1}^m\sum_{\mathbf{t}^m\in\mathcal{T}_{j\to i}}\sum_{\mathbf{c}^m\in\{0,1\}^m}\mathbb{1}\left\{\mathbf{T}^m=\mathbf{t}^m,\mathbf{C}^m=\mathbf{c}^m\right\}\cdot C_i\ , \tag{53}$$

where $\mathbb{1}$ is as before the indicator function. In terms of probabilities, this means that the probability that the $j$-th trial gives $D_j = 1$ is given by

$$\Pr(D_j = 1) = \sum_{i=1}^{m} \left( \sum_{\mathbf{t}^m \in \mathcal{T}_{j\to i}} \sum_{\substack{\mathbf{c}^m \in \{0,1\}^m \\ c_i = 1}} \Pr(\mathbf{T}^m = \mathbf{t}^m, \mathbf{C}^m = \mathbf{c}^m) \right) . \tag{54}$$

We can thus express the $P$-value as

$$\begin{aligned} P\text{-value} &= \sum_{\substack{\mathbf{t}^m \in \mathbf{T}^m \\ |\mathbf{t}^m| = n}} \Pr(\mathbf{T}^m = \mathbf{t}^m) \Pr(\text{number of 1's in } (C_1 \cdot t_1, \ldots, C_m \ldots t_m) \geq k \mid \mathbf{T}^m = \mathbf{t}^m) && (55) \\ &= \Pr(\text{number of 1's in } (D_1, \ldots, D_n) \geq k) && (56) \\ &= \Pr\left( \sum_{j=1}^{n} D_j \geq k \right) . && (57) \end{aligned}$$

Before delving into the proof below, it will be convenient to simplify (54). Note that for a fixed $i$, the term in brackets in (54) contains a sum over all possible $t_{i+1}, \ldots, t_m$ and $c_{i+1}, \ldots, c_m$. This means we can use the law of total probability to shorten the sum by expressing (54) in terms of the marginal distributions as

$$\Pr(D_j = 1) = \sum_{i=1}^{m} \sum_{\mathbf{t}^i \in \mathcal{T}^i_{j\to i}} \sum_{\mathbf{c}^{i-1} \in \{0,1\}^{i-1}} \Pr\left(\mathbf{T}^i = \mathbf{t}^i, \mathbf{C}^{i-1} = \mathbf{c}^{i-1}, C_i = 1\right) , \tag{58}$$

where

$$\mathcal{T}^i_{j\to i} = \left\{ \mathbf{t}^i = (t_1, \ldots, t_i) \in \{0,1\}^i \mid \exists \hat{\mathbf{t}}^m = (\hat{t}_1, \ldots, \hat{t}_m) \in \mathcal{T}_{j\to i} \text{ such that } (\hat{t}_1, \ldots, \hat{t}_i) = (t_1, \ldots, t_i) \right\} . \tag{59}$$

After having formally established the relation between the sequence of trials and the sequence of attempts, we are now ready for the final proof. Since we can now argue in terms of the sequence of trials $(D_1, \ldots, D_n)$ this is analogous to [39][Proposition 4], which we will spell out for completeness.

**Lemma 3.** *Let $m, n, k \in \mathbb{N}$ and let the sequence $(\mathbf{A}^m, \mathbf{B}^m, \mathbf{X}^m, \mathbf{Y}^m, \mathbf{H}^m, \mathbf{T}^m)$ be defined as in Section O. Suppose that the null hypothesis holds, i.e., nature is governed by an LHVM. Given that the predictability of the RNG is $\tau$, we have that for all $m \geq n$, the probability that at least $k$ of the $(D_j)_{j=1}^n$ take the value one is upper bounded by*

$$P\text{-value} = \Pr\left( \sum_{j=1}^{n} D_j \geq k \right) \leq P_{n,k}(\mathbb{B}_\xi) , \tag{60}$$

*where $P_{n,k}(\mathbb{B}_\xi)$ denotes the probability that $n$ Bernoullis with probability $\xi = 3/4 + 3(\tau + \tau^2)$ yield at least $k$ 1's, and $P_{n,k}(\mathbb{B}_\xi) = 0$ if $k > n$.*

*Proof.* Let us define the shorthand

$$P_{n,k}(D) = \Pr\left( \sum_{j=1}^{n} D_j \geq k \right) . \tag{61}$$

The probability that we see at least zero 1's ($k = 0$) obeys

$$\begin{aligned} P_{n,0}(D) &= 1 && (62) \\ &= P_{n,0}(\mathbb{B}_\xi) && (63) \end{aligned}$$

for all $n$ and $m \geq n$.

We now prove the statement for $k > 0$ by induction on $n$. For $n = 1$, we need only to verify that (60) holds for $k = 1$ (we already dealt with $k = 0$). We have

$$P_{1,1}(D) = \Pr(D_1 \geq 1) = \Pr(D_1 = 1) \tag{64}$$

$$= \sum_{i=1}^{m} \sum_{\mathbf{t}^i \in \mathcal{T}^i_{1\to i}} \sum_{\mathbf{c}^{i-1} \in \{0,1\}^{i-1}} \Pr\left(\mathbf{T}^i = \mathbf{t}^i, \mathbf{C}^{i-1} = \mathbf{c}^{i-1}, C_i = 1\right) \tag{65}$$

$$= \sum_{i=1}^{m} \sum_{\mathbf{t}^i \in \mathcal{T}^i_{1\to i}} \sum_{\mathbf{c}^{i-1} \in \{0,1\}^{i-1}} \Pr\left(C_i = 1 | \mathbf{C}^{i-1} = \mathbf{c}^{i-1}, \mathbf{T}^i = \mathbf{t}^i\right) \Pr\left(\mathbf{C}^{i-1} = \mathbf{c}^{i-1}, \mathbf{T}^i = \mathbf{t}^i\right) \tag{66}$$

$$\leq \xi \sum_{i=1}^{m} \sum_{\mathbf{t}^i \in \mathcal{T}^i_{1\to i}} \sum_{\mathbf{c}^{i-1} \in \{0,1\}^{i-1}} \Pr\left(\mathbf{C}^{i-1} = \mathbf{c}^{i-1}, \mathbf{T}^j = \mathbf{t}^j\right) \tag{67}$$

$$= \xi = P_{1,1}(\mathbb{B}_\xi)\ , \tag{68}$$

where the first equality (65) is just (58), the second equality (66) the definition of conditional probability, the inequality (67) follows from Lemma 2, and the final equality (68) from the definition of the sets $\mathcal{T}_{j\to i}$ and the fact the sum of all probabilities is 1.

In order to prove the induction step below, let us first express the probability of having at least $k$ 1's on trial $n$ as the sum of the probability of having at least $k$ on trial $n-1$, plus the probability of having exactly $k-1$ 1's on trial $n-1$ and a one on the $n$-th trial

$$P_{n,k}(D) = \Pr\left(\sum_{j=1}^{n} D_j \geq k\right) \tag{69}$$

$$= \Pr\left(\sum_{j=1}^{n-1} D_j \geq k\right) + \Pr\left(\sum_{j=1}^{n-1} D_j = k-1, D_n = 1\right) \tag{70}$$

$$= P_{n-1,k}(D) + \Pr\left(\sum_{j=1}^{n-1} D_j = k-1, D_n = 1\right)\ . \tag{71}$$

We now upper bound the second term in (71), where we will use the shorthand $|\mathbf{c}^{i-1} \cdot \mathbf{t}^{i-1}| = |(c_1 t_1, \ldots, c_i t_i)| =$

$\sum_{j=1}^{i-1} c_i t_i$.

$$\Pr\left(\sum_{j=1}^{n-1} D_j = k-1, D_n = 1\right) \tag{72}$$

$$= \sum_{i=1}^{m} \sum_{\mathbf{t}^i \in \mathcal{T}^i_{n-1\to i}} \sum_{\substack{\mathbf{c}^{i-1}\in\{0,1\}^{i-1} \\ |\mathbf{c}^{i-1}\cdot\mathbf{t}^{i-1}|=k-1}} \Pr\left(\mathbf{T}^i = \mathbf{t}^i, \mathbf{C}^{i-1} = \mathbf{c}^{i-1}, C_i = 1\right) \tag{73}$$

$$= \sum_{i=1}^{m} \sum_{\mathbf{t}^i \in \mathcal{T}^i_{n-1\to i}} \sum_{\substack{\mathbf{c}^{i-1}\in\{0,1\}^{i-1} \\ |\mathbf{c}^{i-1}\cdot\mathbf{t}^{i-1}|=k-1}} \Pr\left(C_i = 1 | \mathbf{T}^i = \mathbf{t}^i, \mathbf{C}^{i-1} = \mathbf{c}^{i-1}\right) \Pr\left(\mathbf{T}^i = \mathbf{t}^i, \mathbf{C}^{i-1} = \mathbf{c}^{i-1}\right) \tag{74}$$

$$\leq \xi \sum_{i=1}^{m} \sum_{\mathbf{t}^i \in \mathcal{T}^i_{n-1\to i}} \sum_{\substack{\mathbf{c}^{i-1}\in\{0,1\}^{i-1} \\ |\mathbf{c}^{i-1}\cdot\mathbf{t}^{i-1}|=k-1}} \Pr\left(\mathbf{T}^i = \mathbf{t}^i, \mathbf{C}^{i-1} = \mathbf{c}^{i-1}\right) \tag{75}$$

$$= \xi \Pr\left(\sum_{j=1}^{n-1} D_j = k-1\right) \tag{76}$$

$$= \xi \left(P_{n-1,k-1}(D) - P_{n-1,k}(D)\right) . \tag{77}$$

Equality (74) follows by the definition of conditional probability, inequality (75) from Lemma 2, equality (76) from (54), and the last equality (77) because the sum over all vectors having exactly $k-1$ 1's equals the probability of having at least $k-1$ 1's, minus the probability of having at least $k$.

Recall that $P_{n,k}(\mathbb{B}_\xi)$ stands for the probability of having at least $k$ successes over $n$ Bernoullis with probability $\xi$. Before proving the induction step, we need to rewrite $P_{n,k}(\mathbb{B}_\xi)$ as follows

$$\begin{aligned} P_{n,k}(\mathbb{B}_\xi) &= \Pr\left(\text{at least } k \text{ successes over } n-1 \text{ Bernoullis}\right) \\ &\quad + \Pr\left(\text{exactly } k-1 \text{ successes over } n-1 \text{ Bernoullis and success on the } n\text{-th }\right) && (78) \\ &= P_{n-1,k}(\mathbb{B}_\xi) \\ &\quad + \Pr\left(\text{exactly } k-1 \text{ successes over } n-1 \text{ Bernoullis and success on the } n\text{-th }\right) && (79) \\ &= P_{n-1,k}(\mathbb{B}_\xi) \\ &\quad + \Pr\left(\text{exactly } k-1 \text{ successes over } n-1 \text{ Bernoullis}\right) \Pr\left(\text{success on the } n\text{-th trial}\right) && (80) \\ &= P_{n-1,k}(\mathbb{B}_\xi) + \left(P_{n-1,k-1}(\mathbb{B}_\xi) - P_{n-1,k}(\mathbb{B}_\xi)\right) \Pr\left(\text{success on the } n\text{-th trial}\right) && (81) \\ &= P_{n-1,k}(\mathbb{B}_\xi) + \xi \left(P_{n-1,k-1}(\mathbb{B}_\xi) - P_{n-1,k}(\mathbb{B}_\xi)\right) . && (82) \end{aligned}$$

Now we prove the induction step. Consider some arbitrary $n > 1$ and consider the induction hypothesis that $\forall m \geq n$ and $\forall k' \leq n$ , $P_{n-1,k'}(D) \leq P_{n,k'}(\mathbb{B}_\xi)$. The following chain of inequalities hold:

$$\begin{aligned} P_{n,k}(D) &= P_{n-1,k}(D) + \xi \left(P_{n-1,k-1}(D) - P_{n-1,k}(D)\right) && (83) \\ &= (1-\xi) P_{n-1,k}(D) + \xi P_{n-1,k-1}(D) && (84) \\ &\leq (1-\xi) P_{n-1,k}(\mathbb{B}_\xi) + \xi P_{n-1,k-1}(\mathbb{B}_\xi) && (85) \\ &= P_{n,k}(\mathbb{B}_\xi) . && (86) \end{aligned}$$

The first (83) and second (84) equalities follow after plugging (77) back into (71) and rearranging. The inequality (85) follows from the induction hypothesis. The last equality (86) follows from rearranging (82) and completes the induction step. □

### R. Discussion

We remark that, assuming zero bias, our bound is optimal when computing the CHSH function from the data. That is, for CHSH the bound can in fact be attained by a LHVM, showing that a conventional analysis (see Figure 4, main text) is overly optimistic for small $n$. To see that the bound is tight, it is again convenient to adopt the perspective of non-local games: To make the bound tight, it is enough to saturate Lemma 1. This can be done by the usual classical strategy for Alice and Bob in CHSH. Of course, one could compute functions other than CHSH from our data set, which may lead to an even lower $P$-value. We note, however, that we cannot retroactively search for the best function to compute from our data based on the data already collected, since this would need a different analysis for the $P$-value. Very intuitively, this is so because an LHVM could take advantage of the fact that we were to perform such an optimization retroactively.

From our proof it is clear that Lemma 2 and 3 did not make use of the conditions of an LHVM. In both cases, we just required Lemma 1 (the CHSH inequality) to hold. This means that any theory that predicts that Lemma 1 holds for our experiment is excluded with the same $P$-value. It also makes it apparent how one can extend the analysis to refute models that are more powerful than an LHVM. For example, Hall [64] defined and quantified interesting relaxations of an LHVM, with reduced free will, or where some amount of signalling is allowed. It is straightforward to adapt the analysis of [64] to Lemma 1 to obtain a $P$-value for such extended models. Our analysis is thus robust, in the sense that allowing slightly more power to the model also only results in a slight increase in the $P$-value. To see why the $P$-value increases, taking the perspective of a non-local game is again very instructive: if Alice and Bob are allowed more powerful strategies than an LHVM, then the probability that they will produce at least $k$ wins in $n$ trials increases.

### S. Relation to the CHSH correlator

Our bound on the $P$-value depends directly on the number of successes $k$ over $n$ Bernoullis. For completeness, let us explain how this is linked to the average CHSH value, which may be more familiar from the literature. Since our objective is only to illustrate this link and give some intuition on the $P$ values, we assume, only from here and until the end of this section, perfect RNGs. We denote by $\langle XY\rangle_{ab}$ the average of the random variable $XY$ when the settings are $A=a, B=b$

$$\langle XY\rangle_{ab} = \Pr\left(X=Y|A=a,B=b\right) - \Pr\left(X\neq Y|A=a,B=b\right) \tag{87}$$

$$= \begin{cases} 2\Pr\left(C=1|A=a,B=b\right)-1 & \text{if } (a,b)\neq(1,1)\ , \\ 1-2\Pr\left(C=1|A=a,B=b\right) & \text{otherwise.} \end{cases} \tag{88}$$

Let us denote by $S$ the average CHSH value

$$S = \langle CHSH\rangle = \langle XY\rangle_{00} + \langle XY\rangle_{01} + \langle XY\rangle_{10} - \langle XY\rangle_{11}\ . \tag{89}$$

Now, we can link $S$ with $\Pr\left(C=1\right)$ as

$$\frac{S+4}{8} = \frac{2\sum_{a,b}\Pr\left(C=1|A=a,B=b\right)}{8} \tag{90}$$

$$= \sum_{a,b}\frac{1}{4}\Pr\left(C=1|A=a,B=b\right) \tag{91}$$

$$= \Pr\left(C=1\right)\ . \tag{92}$$

That is, we can map the average CHSH value $S$ to the probability that $C$ takes the value one. It directly follows that the known CHSH upper bound $S\leq 2$ corresponds with $\Pr\left(C=1\right)\leq 0.75$, which is the upper bound that we obtain if we assume perfect RNGs ($\tau=0$) in Lemma 1 and Lemma 2. In the same way we can map the observed CHSH violation to the number of successes. Let $n_{a,b}$ denote the number of trials with setting $(a,b)$, $n^{=}_{a,b}$ the number

of successes associated with setting $(a, b)$ and $\tilde{S}$ the observed CHSH value:

$$\tilde{S} = \sum_{a,b} \frac{n^{=}_{a,b}}{n_{a,b}} \ . \tag{93}$$

For large $n$, the following equalities hold approximately

$$n \cdot \frac{\tilde{S}+4}{8} = n \cdot \frac{\sum_{a,b} \frac{n^{=}_{a,b}}{n_{a,b}} + 4}{8} \tag{94}$$

$$= n \cdot \frac{\sum_{a,b} \left(2\frac{n^{=}_{a,b}}{n_{a,b}} - 1\right) + 4}{8} \tag{95}$$

$$\approx \sum_{a,b} n^{=}_{a,b} \tag{96}$$

$$= k \ . \tag{97}$$

The approximation holds since for large $n$ the number of trials at each setting should be approximately $n/4$ and in consequence $n/n_{a,b} \approx 4$.

[33] Pfaff, W. *et al.* Unconditional quantum teleportation between distant solid-state quantum bits. *Science* **345**, 532–535 (2014).
[34] GBKN. GBKN, de standaard basiskaart van Nederland (2015). URL `http://www.gbkn.nl/`.
[35] AHN. AHN - Actueel Hoogtebestand Nederland (2015). URL `http://www.ahn.nl/`.
[36] De Bruijne, A., Van Buren, J., Ksters, A. & Van der Marel, H. *Geodetic reference frames in the Netherlands, Definition and specification of ETRS89, RD and NAP, and their mutual relationships* (NCG, Nederlandse Commissie voor Geodesie, Netherlands Geodetic Commission, Delft, The Netherlands, 2005). URL `http://www.ncgeo.nl/index.php?option=com_k2&view=item&id=2361`.
[37] Abellan, C. *et al.* Ultra-fast quantum randomness generation by accelerated phase diffusion in a pulsed laser diode. *Optics Express* **22**, 1645 (2014).
[38] Hill, A., Nash, J., Graham, M., Hervas, D. & Kwiat, P. Adaptive Optics for Single-Photon Fiber Coupling of Ions. JW2A.128 (OSA, 2014).
[39] Bierhorst, P. A rigorous analysis of the Clauser-Horne-Shimony-Holt inequality experiment when trials need not be independent. *Foundations of Physics* **44**, 736–761 (2014).
[40] Zhang, Y., Glancy, S. & Knill, E. Asymptotically optimal data analysis for rejecting local realism. *Physical Review A* **84**, 062118 (2011).
[41] Gill, R. D. Accardi contra bell (cum mundi): The impossible coupling. *Lecture Notes-Monograph Series* 133–154 (2003).
[42] Van Dam, W., Gill, R. D. & Grunwald, P. D. The statistical strength of nonlocality proofs. *Information Theory, IEEE Transactions on* **51**, 2812–2835 (2005).
[43] Peres, A. Bayesian analysis of Bell inequalities. *Fortschritte der Physik* **48**, 531–535 (2000).
[44] Larsson, J.-Å. & Gill, R. D. Bell's inequality and the coincidence-time loophole. *EPL (Europhysics Letters)* **67**, 707 (2004).
[45] Acín, A., Gill, R. & Gisin, N. Optimal Bell tests do not require maximally entangled states. *Phys. Rev. Lett.* **95**, 210402 (2005).
[46] Pironio, S. *et al.* Random numbers certified by Bell's theorem. *Nature* **464**, 1021–1024 (2010).
[47] Zhang, Y., Knill, E. & Glancy, S. Statistical strength of experiments to reject local realism with photon pairs and inefficient detectors. *Physical Review A* **81**, 032117 (2010).
[48] Pironio, S. & Massar, S. Security of practical private randomness generation. *Physical Review A* **87**, 012336 (2013).
[49] Zhang, Y., Glancy, S. & Knill, E. Efficient quantification of experimental evidence against local realism. *Physical Review A* **88**, 052119 (2013).
[50] Gill, R. D. Statistics, causality and bells theorem. *Statist. Sci.* **29**, 512–528 (2014).
[51] Bierhorst, P. A robust mathematical model for a loophole-free Clauser-Horne experiment. *Journal of Physics A: Mathematical and Theoretical* **48**, 195302 (2015).
[52] Pope, J. E. & Kay, A. Limited measurement dependence in multiple runs of a Bell test. *Physical Review A* **88**, 032110 (2013).

[53] Bancal, J.-D., Sheridan, L. & Scarani, V. More randomness from the same data. *New Journal of Physics* **16**, 033011 (2014).
[54] Kofler, J. & Giustina, M. Requirements for a loophole-free Bell test using imperfect setting generators. *arXiv preprint arXiv:1411.4787* (2014).
[55] Larsson, J.-A. *et al.* Bell-inequality violation with entangled photons, free of the coincidence-time loophole. *Phys. Rev. A* **90**, 032107 (2014).
[56] Larsson, J.-Å. Loopholes in Bell inequality tests of local realism. *Journal of Physics A: Mathematical and Theoretical* **47**, 424003.
[57] Christensen, B. *et al.* Analysis of coincidence-time loopholes in experimental Bell tests. *arXiv preprint arXiv:1503.07573* (2015).
[58] Knill, E., Glancy, S., Nam, S. W., Coakley, K. & Zhang, Y. Bell inequalities for continuously emitting sources. *Phys. Rev. A* **91**, 032105 (2015).
[59] For readers that are familiar with non-local games we note that the 'event-ready' box could be seen as a third player that receives no question as input.
[60] Bentkus, V. On Hoeffding's inequalities. *Annals of probability* 1650–1673 (2004).
[61] Pinelis, I. Binomial upper bounds on generalized moments and tail probabilities of (super) martingales with differences bounded from above. In *High dimensional probability*, 33–52 (Institute of Mathematical Statistics, 2006).
[62] Setting $t_j = 0$ for all future attempts $j$.
[63] Note that since the history captures an arbitrary state of the experiment in the past, it could also include things which are not measured or recorded by the experimenter.
[64] Hall, M. J. W. Relaxed Bell inequalities and Kochen-Specker theorems. *Physical Review A* **84**, 022102 (2011).