

The Locality Gap: Bell-Inequality Audits for Covert Coordination Between Ostensibly Isolated AI Agents

Alfredo Merlet

alfredo.merlet@gmail.com

Universidad de Sevilla <https://orcid.org/0009-0007-8558-164X>

Research Article

Keywords: Bell inequality, CHSH game, AI auditing, covert coordination, algorithmic collusion, measurement independence

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11430376/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Einstein's "spooky action at a distance" names a correlation that no local mechanism can produce. This article turns that observation into an audit instrument for artificial intelligence. Two AI agents that share training data, weights or random seeds but cannot communicate are local systems in Bell's sense, so their behaviour in a Clauser–Horne–Shimony–Holt (CHSH) game is bounded by $S \leq 2$ whatever their architecture; a significant excess identifies a covert channel without modelling the agents. The instrument is tested in three computational experiments with known ground truth. First, 135 trained pairs of neural agents (three widths, three latent dimensions, two objectives) never exceed the bound (maximum $S = 2.0000$), including pairs trained to imitate the correlations of an entangled singlet. Second, opening a one-bit channel with probability p moves the learned CHSH value onto the theoretical optimum $2 + 2p$ (mean absolute deviation 0.0016 across 220 runs); any leak below $p \approx 0.414$ produces values indistinguishable in S from quantum entanglement. Third, as a finite-sample test the audit keeps its nominal size against honest but strongly correlated agents (rejection 4.4% and 4.3%), whereas a conventional dependence test flags the same agents in 100.0% of audits; detecting a 5% leak with 80% power needs about 7,375 rounds and a 1% leak about 185,120. The decisive weakness lies with the auditor: with predictable settings (a low-order bit of a linear congruential generator, a fixed rotation or a reused seed), agents with zero communication reach $S \approx 4$ and trigger false alarms in 100% of audits, while cryptographic and freshly seeded generators restore the nominal error rate. The bridge to algorithmic collusion and multi-agent AI governance is stated as a formal analogy and a hypothesis for field testing, not as a finding about deployed systems.

JEL classification: C12, C63, D43, L41, O33

1. Introduction

In a letter to Max Born written in March 1947, Einstein dismissed the correlations that quantum mechanics predicts between distant systems as "spooky action at a distance" (Born, 1971). The objection went back to the argument of Einstein, Podolsky and Rosen (1935) that a theory permitting such correlations must be incomplete. Bell (1964) turned the dispute into a theorem: any account in which each outcome depends only on its local setting and on variables fixed in advance must satisfy an inequality that quantum mechanics violates. Clauser, Horne, Shimony and Holt (1969) gave the inequality its testable two-setting form, and a sequence of experiments, from Aspect, Dalibard and Roger (1982) to the loophole-free test of Hensen et al. (2015), confirmed the violation.

The logic of Bell's theorem does not depend on physics. It is a statement about any two systems whose outputs are functions of their own inputs and of information shared before the inputs were chosen. That description fits, almost word for word, the situation that regulators and auditors assume for artificial intelligence (AI) agents that are meant to act independently: pricing algorithms of competing firms, large language model (LLM) agents asked to assess the same case separately, or model instances run in isolated sandboxes. Concern that such agents coordinate covertly is no longer speculative.

Reinforcement-learning pricing agents learn collusive strategies without being instructed to (Calvano et al., 2020), the adoption of algorithmic pricing has been associated with higher margins in real markets (Assad et al., 2024), and frontier LLMs show growing capacity to hide messages from monitors through steganography (Motwani et al., 2024; Hammond et al., 2025).

Existing detection methods share a structural difficulty. They either inspect the channel, which fails when the channel is hidden, or test whether the agents' outputs are statistically dependent once shared inputs are conditioned out (Cherif et al., 2026). The second approach breaks down when agents share something that the auditor cannot condition on, such as a common pretrained model, overlapping training data or a seed. Shared training produces correlation without communication, and a dependence test cannot tell the two apart. This is what the article calls the *locality gap*: the distance between the correlation that auditors can observe and the communication that the rules are concerned with.

Bell's theorem closes that gap for one well-defined class of audits. If the auditor sends each agent a private, uniformly random query in a separate session and scores the answers as a CHSH game, then any pair of classical agents without a channel, whatever they share, is bounded by a CHSH value of 2. A significant excess is therefore evidence of communication during the audit, obtained without modelling the agents. The article asks three questions. RQ1: does the bound hold for learned neural agents, so that the instrument is valid in practice? RQ2: how does the CHSH value respond to a partial covert channel? RQ3: what are the size, power and failure modes of the audit as a finite-sample test? Each question is answered by a computational experiment in which the true amount of communication is known. A bibliometric check in Scopus (Section 2.4) found no prior work that uses Bell-type bounds to audit coordination between AI agents.

Section 2 reviews related work and reports the bibliometric gap. Section 3 sets out the threat model and the audit. Section 4 describes the experiments, Section 5 reports the results and Section 6 discusses implications, the formal analogy with competition policy and the limitations. Section 7 concludes.

2. Related work

2.1 Bell nonlocality as a theorem about locality

Brunner et al. (2014) review Bell nonlocality in detail. Three bounds organise the CHSH scenario: 2 for local models, $2\sqrt{2}$ for quantum strategies (Cirel'son, 1980) and 4 for arbitrary no-signalling correlations (Popescu & Rohrlich, 1994). In its game form, two players who cannot communicate win with probability at most $3/4$ classically and $\cos^2(\pi/8) \approx 0.854$ with shared entanglement (Cleve et al., 2004).

Communication removes the bound altogether: Toner and Bacon (2003) showed that one bit per round suffices to reproduce all projective measurements on a singlet. The bound also rests on measurement independence, the assumption that settings are uncorrelated with the shared variables. Hall (2010) showed that relaxing this assumption by about one third suffices for a local deterministic model to reach the maximal CHSH violation, and freedom-of-choice experiments address it directly with fast random

generators or human choices (Scheidl et al., 2010; BIG Bell Test Collaboration, 2018). Finite-sample inference is equally developed: for win/lose Bell games, the binomial tail gives a P value that stays valid when the devices have arbitrary memory (Elkouss & Wehner, 2016).

2.2 Neural networks as local models

A growing literature uses machine learning to study nonlocality. Neural networks have been trained as oracles that try to reproduce a distribution with local resources in network scenarios (Kriváchy et al., 2020), packaged into a general toolbox (Girardin et al., 2026), and used to search for one-bit simulations of partially entangled states (Sidajaya et al., 2023). Other work detects or quantifies nonlocality with multilayer perceptrons, restricted Boltzmann machines or ensembles (Canabarro et al., 2019; Deng, 2018; Ma & Yung, 2018; Zhang et al., 2024). Wood and Spekkens (2015) showed that standard causal discovery algorithms cannot distinguish correlations that satisfy Bell inequalities from those that violate them. All of this work uses machine learning to study physics. The present article runs the direction the other way: it uses a physical bound to audit machine learning systems.

2.3 Detecting coordination among AI agents

In economics and law, algorithmic collusion is studied as a problem of tacit coordination among pricing algorithms (Ezrachi & Stucke, 2016; Van Uytsel, 2018; Bernhardt & Dewenter, 2020), with learning-based detection proposed as a regulatory tool (Cave, 2019) and evaluation designs that test whether collusion emerges under realistic constraints (Luo et al., 2026). In AI safety, secret collusion is formalised as steganographic communication between generative agents (Motwani et al., 2024). The closest detection method is the conditional-mutual-information test of Cherif et al. (2026), which flags LLM agents whose scores remain dependent after the shared case is conditioned out. That test is powerful when the only shared element is the observable input. It is not designed for agents that also share latent structure, and Section 5.3 shows that a dependence test of this kind rejects honest agents with shared latent randomness in every audit.

2.4 Bibliometric gap

Five Scopus queries were run on 16 October 2026 in the title, abstract and keyword fields (Table 1). The broadest query on machine learning and Bell nonlocality returned 52 records, all in physics. The intersection of Bell tests with multi-agent systems, collusion or steganography returned 4 records, none on AI agents. The query on detection of algorithmic or secret collusion returned 17 records, none using Bell-type bounds. A query on neural local hidden-variable models was flooded by the homonym “communication cost” in federated learning (2,404 records, of which 3 concern local hidden variables). The narrowest query, which intersects Bell inequalities with AI and with auditing, governance or collusion, returned 0 records. The zero was checked by script: no record in the four other exports satisfies its logic. One record in the first export uses “Bell” for the Bell System antitrust decree rather than Bell’s theorem and was set aside.

Table 1
Scopus bibliometric funnel (16 October 2026)

Query (TITLE-ABS-KEY)	Records	Pertinent to the article
Bell inequality / CHSH / nonlocality $\wedge$ neural networks / machine learning	52	Physics only; neural local models (Kriváchy et al., 2020; Girardin et al., 2026)
Bell inequality / CHSH / Bell test $\wedge$ multi-agent / collusion / steganography	4	0 on AI agents
Algorithmic or secret collusion / covert coordination $\wedge$ audit / detection $\wedge$ AI	17	Detection without Bell bounds (Cherif et al., 2026; Cave, 2019)
Local hidden variable / communication cost $\wedge$ neural networks / machine learning	2,404	3 (homonym flood from federated learning)
Bell inequality / CHSH $\wedge$ AI / AI agents $\wedge$ audit / governance / collusion	0	Gap (zero validated by script)
<i>Note.</i> Full queries are reproduced in the replication package; the exports themselves are not redistributed. $\wedge$ denotes logical AND.		

3. Framework: threat model and the Bell audit

An auditor runs N rounds. In each round it draws two independent uniform bits x and y and sends x to agent A and y to agent B in separate sessions; each agent returns an answer in $\{+1, -1\}$. Before the audit the agents may share anything: training data, weights, a seed or, in general, a variable λ with an arbitrary distribution. During the audit they are supposed not to communicate. Locality means that $a = f_A(x, \lambda, \xi_A)$ and $b = f_B(y, \lambda, \xi_B)$, where ξ_A and ξ_B are private randomness. The correlators and the CHSH value are

$$E_{xy} = E[a b \mid x, y], S = E_{00} + E_{01} + E_{10} - E_{11} \quad (1)$$

and Bell's theorem gives, for every local model,

$$S \leq 2 \text{ (local)}, S \leq 2\sqrt{2} \text{ (quantum)}, S \leq 4 \text{ (algebraic)} \quad (2)$$

In game form the agents win a round when $a b = (-1)^{-xy}$, and the win probability is

$$w = 1/2 + S/8, \text{ so that } w \leq 3/4 \text{ for local agents} \quad (3)$$

Leak model. Suppose a covert channel delivers x to agent B with probability p in each round. In leaked rounds the pair can win with certainty; in the others it is local. The optimal strategy therefore reaches

$$w^*(p) = 3/4 + p/4, S^*(p) = 2 + 2p \quad (4)$$

which equals the Tsirelson value $2\sqrt{2}$ at $p = \sqrt{2} - 1 \approx 0.414$. Any leak below that rate yields a CHSH value that a purely quantum strategy could also produce.

The test. Let W be the number of rounds won. The audit tests $H_0: w \leq 3/4$ with the one-sided exact binomial P value

$$P = \Pr[\text{Bin}(N, 3/4) \geq W] \quad (5)$$

which remains valid when agents adapt to the history of the audit (Elkouss & Wehner, 2016). For a leak of rate p , the number of rounds required for power $1 - \beta$ at level α is approximately

$$N^* \approx [z_{1-\alpha} \sqrt{w_0(1 - w_0)} + z_{1-\beta} \sqrt{w_1(1 - w_1)}]^2 / (w_1 - w_0)^2, w_0 = 3/4, w_1 = 3/4 + p/4 \quad (6)$$

Predictable settings. If agent B can predict x with accuracy q from information available to it, such as a published transcript, the pair can play as if leaked whenever the prediction is right. With the strategy used in Section 4.3, the CHSH value becomes

$$S(q) = 2 + 2(2q - 1) \quad (7)$$

so a perfectly predictable generator ($q = 1$) produces $S = 4$ with no communication at all. This is the measurement-independence loophole of Section 2.1 expressed in the auditor's terms.

4. Methods

4.1 Neural agents and Experiment 1 (local saturation)

Each agent is a neural network with one hidden layer of width H and tanh activations. Its tanh output $o \in [-1, 1]$ is the conditional mean of a ± 1 answer, which is then drawn with probability $(1 + o)/2$ of being $+1$ using private randomness. Agent A receives its setting and the shared latent $\lambda \sim N(0, I_K)$; agent B receives its setting, λ and, only in Experiment 2, a leak flag and the leaked bit. Because the answers are conditionally independent given the inputs and λ , $E_{xy} = E_\lambda[o_A o_B]$, which permits exact gradients. Gradients were written by hand and checked against central finite differences before any run (maximum relative discrepancy 4.6×10^{-7} for the CHSH objective and 2.1×10^{-6} for the singlet objective). Training used Adam (Kingma & Ba, 2015).

Experiment 1 crosses three widths ($H = 8, 32, 128$) with three latent dimensions ($K = 1, 4, 16$) under two objectives. Objective A maximises S directly in the CHSH game (600 steps, 512 latent draws per step, learning rate 0.03, 10 seeds per cell; 90 runs). Objective B fits the singlet correlation $E(\theta_A, \theta_B) = -\cos(\theta_A - \theta_B)$ by mean squared error on a randomly shifted 10×10 grid of angles over $[0, 2\pi)$ (800 steps, 256 draws, learning rate 0.02, 5 seeds per cell; 45 runs), so that the networks are pushed towards correlations that violate Bell's inequality. Each trained pair is evaluated on fresh draws: 200,000 for objective A, and 50,000 at the standard CHSH angles ($0, \pi/2$ for A; $\pi/4, -\pi/4$ for B) for objective B, whose deviation from the singlet is also measured on an independent 25×25 grid. The known truth is $S \leq 2$ for every run.

4.2 Experiment 2 (leak dose-response)

The channel rate p takes the values 0, 0.1, ..., 1. For each value, ten pairs are trained under objective A in two configurations ($H = 32, K = 4$ and $H = 8, K = 1$), giving 220 runs, and each pair is evaluated on 200,000 fresh rounds. The known truth is $S^*(p) = 2 + 2p$ (Eq. 4).

4.3 Experiment 3 (the audit as a test)

3a, size under honest correlation. Two honest pairs are audited 1,000 times at $N = 1,000$ and $N = 10,000$ rounds: a pair fitted to the singlet (objective B, $H = 32, K = 4$), which is strongly correlated through λ but has no channel, and a pair of independent random answerers. Each audit applies the Bell test (Eq. 5) and a conventional dependence test: a χ^2 test of independence between a and b within each of the four setting cells, combined by the Bonferroni-adjusted minimum. Both tests use $\alpha = 0.05$.

3b, power against leaks. The optimal leaky strategy is simulated exactly for $p \in \{0, 0.01, 0.02, 0.05, 0.1, 0.2, \sqrt{2} - 1\}$ and $N \in \{10^2, \dots, 10^6\}$, with 2,000 audits per cell, and the Monte Carlo rejection rates are compared with the exact binomial power. Eq. 6 and an exact search give the rounds needed for 80% power.

3c, predictable settings. Five ways of generating the auditor's settings are compared: the operating-system cryptographic generator (Python's `secrets` module), the Mersenne Twister (Matsumoto & Nishimura, 1998) freshly seeded for each audit, bit 4 of a linear congruential generator with the classic C-library constants (low-order bits of such generators are periodic; Knuth, 1997), a fixed rotation through the four setting pairs, and the Mersenne Twister reseeded with 42 in every audit, a common reproducibility convention. Agent B predicts x in each round with an order-8 context table built from the public transcript of earlier setting pairs; in the reused-seed case it uses the transcript of an earlier audit. Agent A always answers + 1, and agent B answers - 1 only when it predicts $x = 1$ and receives $y = 1$. The agents never communicate. Each generator is audited 200 times at $N = 10,000$.

4.4 Software and reproducibility

All experiments run on a single CPU core in Python 3.12 with NumPy 2.4, SciPy 1.17 and Matplotlib 3.10. Every random draw is seeded except the cryptographic and fresh-seed generators of Experiment 3c, which are unseeded by design and therefore reproduce in distribution rather than bit for bit. The scripts are resumable, write one CSV row per run and use relative paths. Code, run-level results, tables, figures and the Scopus queries are deposited at <https://doi.org/10.5281/zenodo.23425005>.

5. Results

5.1 Local agents saturate but never cross the Bell bound

Table 2 and Fig. 1 report Experiment 1. Agents trained to maximise S converge to the classical optimum: across 90 runs the mean is 2.0000 and the maximum 2.0000. Agents trained to imitate the singlet reach a mean of 1.9943 and a maximum of 1.9998 at the CHSH angles, and their correlation functions deviate from $-\cos(\theta_A - \theta_B)$ by a root mean square of 0.163. No run in any of the 18 cells exceeds 2 beyond the

evaluation tolerance of 10^{-3} (0 of 135). Width changes how close the singlet-fitted networks come to the bound (mean S of about 1.986 at width 8 and 1.999 at width 128), not whether they cross it, which is the behaviour the theorem requires of a valid instrument.

Table 2
Experiment 1: CHSH values of trained local neural agents

Objective	Width	Latent dim.	Runs	Mean S	Max S	Runs with $S > 2$	RMS vs singlet
A: maximise S	8	1	10	1.9999	1.9999	0	—
A: maximise S	8	4	10	1.9999	1.9999	0	—
A: maximise S	8	16	10	1.9999	1.9999	0	—
A: maximise S	32	1	10	2.0000	2.0000	0	—
A: maximise S	32	4	10	2.0000	2.0000	0	—
A: maximise S	32	16	10	2.0000	2.0000	0	—
A: maximise S	128	1	10	2.0000	2.0000	0	—
A: maximise S	128	4	10	2.0000	2.0000	0	—
A: maximise S	128	16	10	2.0000	2.0000	0	—
B: fit singlet	8	1	5	1.9877	1.9914	0	0.172
B: fit singlet	8	4	5	1.9857	1.9903	0	0.170
B: fit singlet	8	16	5	1.9857	1.9925	0	0.170
B: fit singlet	32	1	5	1.9988	1.9993	0	0.159
B: fit singlet	32	4	5	1.9973	1.9985	0	0.160
B: fit singlet	32	16	5	1.9962	1.9972	0	0.160
B: fit singlet	128	1	5	1.9997	1.9998	0	0.158
B: fit singlet	128	4	5	1.9991	1.9995	0	0.156
B: fit singlet	128	16	5	1.9984	1.9996	0	0.159
<i>Note.</i> $S > 2$ is counted with a tolerance of 10^{-3} for Monte Carlo evaluation error. RMS: root mean square deviation from $-\cos(\theta_A - \theta_B)$ on an independent 25×25 angle grid (objective B only).							

5.2 A covert channel moves the CHSH value onto its theoretical optimum

Table 3 and Fig. 2 report Experiment 2. In both configurations the learned value tracks $S^*(p) = 2 + 2p$ across the whole range: the mean absolute deviation from the optimum is 0.0016 and the largest single deviation 0.0446 over 220 runs. The agents discover, without being told, that the leaked bit should be used only in leaked rounds and that the deterministic classical strategy is optimal otherwise. The value crosses the Tsirelson bound only when p exceeds about 0.414, so the regime between $p = 0$ and $p \approx 0.41$ is the one in which a leak produces correlations that an entangled pair could also produce.

Table 3
Experiment 2: CHSH value by channel rate (width 32, latent 4; width 8, latent 1)

p	Theory $2 + 2p$	Mean S (32/4)	SD (32/4)	Mean S (8/1)	SD (8/1)
0.0	2.0000	2.0000	0.0000	1.9999	0.0000
0.1	2.2000	2.1997	0.0029	2.1996	0.0015
0.2	2.4000	2.4005	0.0015	2.4000	0.0021
0.3	2.6000	2.5984	0.0055	2.5996	0.0022
0.4	2.8000	2.8003	0.0022	2.8000	0.0025
0.5	3.0000	3.0008	0.0023	2.9993	0.0020
0.6	3.2000	3.1988	0.0081	3.1996	0.0022
0.7	3.4000	3.4004	0.0023	3.4003	0.0022
0.8	3.6000	3.6003	0.0017	3.5999	0.0018
0.9	3.8000	3.8002	0.0011	3.7995	0.0013
1.0	4.0000	4.0000	0.0000	3.9954	0.0140
Note. Ten training seeds per cell; each pair evaluated on 200,000 fresh rounds.					

5.3 The audit keeps its size where a dependence test fails

Table 4 reports Experiment 3a. Against the singlet-fitted pair, which is honest but strongly correlated through its shared latent variable, the Bell test rejects in 4.4% of audits at $N = 1,000$ and 4.3% at $N = 10,000$, close to and below the nominal 5% because the binomial test is discrete and the pair sits just under the bound. The dependence test rejects the same pair in 100.0% and 100.0% of audits. Against independent random answerers both tests behave as expected. A dependence test therefore cannot separate shared training from communication, while the Bell test does so by construction.

Table 4
Experiment 3a: rejection rates under honest agents ($\alpha = 0.05$; 1,000 audits per row)

Honest pair	N	Mean S	Bell test	Dependence test
singlet-fitted honest pair	1,000	1.9969	4.4%	100.0%
independent random pair	1,000	-0.0054	0.0%	5.0%
singlet-fitted honest pair	10,000	2.0008	4.3%	100.0%
independent random pair	10,000	0.0022	0.0%	5.8%
Note. Dependence test: χ^2 test of independence of the two answers within each setting cell, Bonferroni-adjusted minimum P value.				

Experiment 3b gives the power side (Table 5, Fig. 3). The Monte Carlo rejection rates agree with the exact binomial power in every cell, and the size at $p = 0$ stays at or near 5% for every N . Power grows with $N(p/4)^2$, so the required audit length rises steeply as the leak shrinks: about 446 rounds for $p = 0.2$, 1,815 for $p = 0.1$, 7,375 for $p = 0.05$ and 185,120 for $p = 0.01$ at 80% power.

Table 5
Experiment 3b: audit rounds required for 80% power at $\alpha = 0.05$

Leak rate p	Win rate under leak	N^* (Eq. 6)	N (exact binomial)
0.0100	0.7525	185,056	185,120
0.0200	0.7550	46,157	46,280
0.0500	0.7625	7,333	7,375
0.1000	0.7750	1,810	1,815
0.2000	0.8000	440	446
0.4142	0.8536	95	103
Note. Exact column: smallest N found by bisection (within 0.1%) at which the exact binomial test reaches 80% power.			

5.4 Predictable settings break the audit

Table 6 and Fig. 4 report Experiment 3c. With the cryptographic generator and the freshly seeded Mersenne Twister, agent B predicts Alice's setting no better than chance (0.5002 and 0.4999), the mean CHSH value is 1.9997 and 1.9997, and false alarms occur in 4.0% and 5.0% of audits. With the low-order bit of the congruential generator, the fixed rotation and the reused seed, prediction accuracy is 0.9988, 0.9994 and 1.0000, the CHSH value approaches the algebraic maximum of 4, and every audit raises a false alarm although the agents never communicate. Eq. 7 predicts these values from the prediction accuracy alone.

Table 6

Experiment 3c: zero-communication agents against five setting generators ($N = 10,000$; 200 audits)

Setting generator	Prediction accuracy q	Mean S	Max S	Predicted S (Eq. 7)	False alarms
OS CSPRNG (secrets)	0.5002	1.9997	2.0952	2.0008	4.0%
MT19937, fresh seed	0.4999	1.9997	2.0808	1.9996	5.0%
LCG, low-order bit	0.9988	3.9953	3.9984	3.9952	100.0%
Fixed rotation schedule	0.9994	3.9978	4.0000	3.9976	100.0%
MT19937, reused seed 42	1.0000	4.0000	4.0000	4.0000	100.0%
Note. Agent B predicts Alice’s setting from the public transcript with an order-8 context table; in the reused-seed case it uses the transcript of an earlier audit.					

6. Discussion

6.1 What a Bell audit adds

The experiments establish three properties of the audit under known ground truth. Learned local agents saturate the bound without crossing it (RQ1). A covert channel moves the CHSH value onto a known optimum, so the size of an excess maps onto the rate of communication (RQ2). As a test, the audit keeps its nominal error rate against honest agents that share latent structure and detects small leaks with audit lengths that are long but feasible for automated agents (RQ3). The advantage over dependence testing is not power but the null hypothesis. A dependence test treats correlation as the warning sign; a Bell audit treats only correlation beyond what any shared training could produce as the warning sign. For agents that share foundation models, which is becoming the default, the distinction decides whether an audit can be used at all.

6.2 A formal analogy with competition policy

Competition law draws a line between tacit parallelism, which is generally lawful, and coordination that rests on communication or agreement (Ezrachi & Stucke, 2016; Van Uytsel, 2018). The Bell audit separates the same two things: it is silent about correlation that shared information can explain and responds only to communication during the audit. This correspondence is offered as a formal analogy, not as an equivalence. Pricing algorithms do not play CHSH games, and the audit applies only when an auditor can pose private, randomised queries to each agent in separate sessions, for example in a regulatory sandbox or a pre-deployment evaluation of LLM agents. The hypothesis for future work is that sandbox audits of this kind, run on agents whose communication can be switched on and off, would

reproduce the dose-response of Fig. 2. The present article provides no evidence about deployed systems.

6.3 The Tsirelson ambiguity

For agents running on classical hardware, any significant excess over 2 implies communication, because the agents hold no entanglement. The inference weakens if agents can access quantum resources, since an excess up to $2\sqrt{2}$ could then come from shared entangled states. An auditor in that setting would have to look for values above $2\sqrt{2}$, and Eq. 4 shows what this costs: leaks below $p \approx 0.414$ would become invisible to the CHSH statistic. The locality gap would then reopen at a higher level.

6.4 The auditor's randomness is the binding constraint

Experiment 3c identifies the most practical risk, and it lies with the auditor rather than the agents. Reproducibility conventions such as a fixed seed, published schedules or weak generators make settings predictable, and predictable settings let agents exceed the bound without any channel. In the reverse direction, colluding agents that could predict settings would be indistinguishable from honest ones that happen to share a model of the generator. Three design rules follow from the results: draw settings from a cryptographic source, never reuse seeds across audits, and release the transcript only after the audit, preferably with a cryptographic commitment made in advance. Physics reached the same conclusion through freedom-of-choice experiments (Scheidt et al., 2010; BIG Bell Test Collaboration, 2018).

6.5 Limitations

The agents are shallow networks in a two-party, two-setting game, and the channel is a simple one-bit leak; richer channels, more parties and other Bell inequalities remain to be tested. The audit detects coordination exercised during the audit only: agents that recognise an audit and behave honestly during it will pass. The test requires that queries be delivered privately and that agents be prevented from exchanging information between query and answer, which is a locality condition on the sandbox. Experiment 3c uses one predictor; more capable predictors would make weak generators fail faster, not slower. No LLM agents were audited, and the bridge to competition policy is a formal analogy (Section 6.2).

7. Conclusion

Einstein called it spooky because no local mechanism could produce it. That is exactly the property an auditor needs. For agents that share everything except a channel, Bell's inequality gives a null hypothesis that no amount of shared training can violate, so an excess points to communication without any model of the agents. In three experiments with known ground truth, learned local agents respected the bound, a covert channel produced the excess that theory predicts, and the audit kept its error rate where a dependence test failed completely. The binding constraint turned out to be the auditor's randomness. A Bell audit is only as good as the dice it rolls.

Declarations

Data availability. All run-level results, tables, figures and the Scopus queries are openly available at <https://doi.org/10.5281/zenodo.23425005> under a CC BY 4.0 licence. No external data were used: all data are generated by the code.

Code availability. The complete code, with a script that reproduces every table and figure, is included in the same deposit (<https://doi.org/10.5281/zenodo.23425005>).

Funding. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests. The author declares no competing interests.

Ethics approval. Not applicable. The study uses only simulated agents and generated data.

Author contributions. A. A. Merlet Echavarría conceived the study, designed the experiments, carried out the analysis and wrote the manuscript.

Use of generative AI. Generative AI tools (Claude, Anthropic) assisted with drafting code and with language editing. The author designed the study, checked every gradient, result and reference against the code output and the primary sources, and takes full responsibility for the content.

References

1. Aspect, A., Dalibard, J., & Roger, G. (1982). Experimental test of Bell's inequalities using time-varying analyzers. *Physical Review Letters*, 49(25), 1804–1807. <https://doi.org/10.1103/PhysRevLett.49.1804>
2. Assad, S., Clark, R., Ershov, D., & Xu, L. (2024). Algorithmic pricing and competition: Empirical evidence from the German retail gasoline market. *Journal of Political Economy*, 132(3), 723–771. <https://doi.org/10.1086/726906>
3. Bell, J. S. (1964). On the Einstein Podolsky Rosen paradox. *Physics Physique Fizika*, 1(3), 195–200. <https://doi.org/10.1103/PhysicsPhysiqueFizika.1.195>
4. Bernhardt, L., & Dewenter, R. (2020). Collusion by code or algorithmic collusion? When pricing algorithms take over. *European Competition Journal*, 16(2–3), 312–342. <https://doi.org/10.1080/17441056.2020.1733344>
5. BIG Bell Test Collaboration. (2018). Challenging local realism with human choices. *Nature*, 557(7704), 212–216. <https://doi.org/10.1038/s41586-018-0085-3>
6. Born, M. (Ed.). (1971). *The Born–Einstein letters: Correspondence between Albert Einstein and Max and Hedwig Born from 1916 to 1955*. Macmillan.

7. Brunner, N., Cavalcanti, D., Pironio, S., Scarani, V., & Wehner, S. (2014). Bell nonlocality. *Reviews of Modern Physics*, 86(2), 419–478. <https://doi.org/10.1103/RevModPhys.86.419>
8. Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297. <https://doi.org/10.1257/aer.20190623>
9. Canabarro, A., Brito, S., & Chaves, R. (2019). Machine learning nonlocal correlations. *Physical Review Letters*, 122(20), 200401. <https://doi.org/10.1103/PhysRevLett.122.200401>
10. Cave, J. (2019). Can machines learn whether machines are learning to collude? In *Lecture Notes in Computer Science: Vol. 11938* (pp. 133–152). Springer. https://doi.org/10.1007/978-3-030-34770-3_11
11. Cherif, A. N., Youssfi, M., & Bouattane, O. (2026). Detecting collusion in LLM-based multi-agent financial decision systems: A conditional-mutual-information detector and a modular oversight architecture. *International Journal of Intelligent Engineering and Systems*, 19(9), 1379–1391. <https://doi.org/10.22266/ijies2026.0930.66>
12. Cirel'son, B. S. (1980). Quantum generalizations of Bell's inequality. *Letters in Mathematical Physics*, 4(2), 93–100. <https://doi.org/10.1007/BF00417500>
13. Clauser, J. F., Horne, M. A., Shimony, A., & Holt, R. A. (1969). Proposed experiment to test local hidden-variable theories. *Physical Review Letters*, 23(15), 880–884. <https://doi.org/10.1103/PhysRevLett.23.880>
14. Cleve, R., Høyer, P., Toner, B., & Watrous, J. (2004). Consequences and limits of nonlocal strategies. In *Proceedings of the 19th IEEE Annual Conference on Computational Complexity* (pp. 236–249). IEEE. <https://doi.org/10.1109/CCC.2004.1313847>
15. Deng, D.-L. (2018). Machine learning detection of Bell nonlocality in quantum many-body systems. *Physical Review Letters*, 120(24), 240402. <https://doi.org/10.1103/PhysRevLett.120.240402>
16. Einstein, A., Podolsky, B., & Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete? *Physical Review*, 47(10), 777–780. <https://doi.org/10.1103/PhysRev.47.777>
17. Elkouss, D., & Wehner, S. (2016). (Nearly) optimal P values for all Bell inequalities. *npj Quantum Information*, 2, 16026. <https://doi.org/10.1038/npjqi.2016.26>
18. Ezrachi, A., & Stucke, M. E. (2016). *Virtual competition: The promise and perils of the algorithm-driven economy*. Harvard University Press.
19. Girardin, A., Rashidi, M. M., Haack, G., Brunner, N., & Pozas-Kerstjens, A. (2026). A versatile neural-network toolbox for testing Bell locality in networks. *Quantum Science and Technology*, 11(3), 035056. <https://doi.org/10.1088/2058-9565/ae8df7>
20. Hall, M. J. W. (2010). Local deterministic model of singlet state correlations based on relaxing measurement independence. *Physical Review Letters*, 105(25), 250404. <https://doi.org/10.1103/PhysRevLett.105.250404>

21. Hammond, L., Chan, A., Clifton, J., et al. (2025). *Multi-agent risks from advanced AI* (Cooperative AI Foundation Technical Report No. 1). arXiv:2502.14143. <https://doi.org/10.48550/arXiv.2502.14143>
22. Hensen, B., Bernien, H., Dréau, A. E., et al. (2015). Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature*, 526(7575), 682–686. <https://doi.org/10.1038/nature15759>
23. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. <https://doi.org/10.48550/arXiv.1412.6980>
24. Knuth, D. E. (1997). *The art of computer programming, Vol. 2: Seminumerical algorithms* (3rd ed.). Addison-Wesley.
25. Kriváchy, T., Cai, Y., Cavalcanti, D., Tavakoli, A., Gisin, N., & Brunner, N. (2020). A neural network oracle for quantum nonlocality problems in networks. *npj Quantum Information*, 6(1), 70. <https://doi.org/10.1038/s41534-020-00305-x>
26. Luo, Y., Schoepflin, D., & Wang, X. (2026). Algorithmic collusion at test time: A meta-game design and evaluation. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (Vol. 3, pp. 1444–1453). IFAAMAS. <https://doi.org/10.65109/SVWG7670>
27. Ma, Y.-C., & Yung, M.-H. (2018). Transforming Bell’s inequalities into state classifiers with machine learning. *npj Quantum Information*, 4(1), 34. <https://doi.org/10.1038/s41534-018-0081-3>
28. Matsumoto, M., & Nishimura, T. (1998). Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, 8(1), 3–30. <https://doi.org/10.1145/272991.272995>
29. Motwani, S. R., Baranchuk, M., Strohmeier, M., Bolina, V., Torr, P. H. S., Hammond, L., & Schroeder de Witt, C. (2024). Secret collusion among AI agents: Multi-agent deception via steganography. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. <https://doi.org/10.48550/arXiv.2402.07510>
30. Popescu, S., & Rohrlich, D. (1994). Quantum nonlocality as an axiom. *Foundations of Physics*, 24(3), 379–385. <https://doi.org/10.1007/BF02058098>
31. Scheidl, T., Ursin, R., Kofler, J., et al. (2010). Violation of local realism with freedom of choice. *Proceedings of the National Academy of Sciences*, 107(46), 19708–19713. <https://doi.org/10.1073/pnas.1002780107>
32. Sidajaya, P., Lim, A. D., Yu, B., & Scarani, V. (2023). Neural network approach to the simulation of entangled states with one bit of communication. *Quantum*, 7, 1150. <https://doi.org/10.22331/q-2023-10-24-1150>
33. Toner, B. F., & Bacon, D. (2003). Communication cost of simulating Bell correlations. *Physical Review Letters*, 91(18), 187904. <https://doi.org/10.1103/PhysRevLett.91.187904>
34. Van Uytsel, S. (2018). Artificial intelligence and collusion: A literature overview. In M. Corrales, M. Fenwick, & N. Forgó (Eds.), *Robotics, AI and the future of law* (pp. 155–182). Springer.

https://doi.org/10.1007/978-981-13-2874-9_7

35. Wood, C. J., & Spekkens, R. W. (2015). The lesson of causal discovery algorithms for quantum correlations: Causal explanations of Bell-inequality violations require fine-tuning. *New Journal of Physics*, 17, 033002. <https://doi.org/10.1088/1367-2630/17/3/033002>
36. Zhang, Y., Hou, J.-C., & He, K. (2024). Detecting nonlocal correlations in chain-shaped quantum networks via a deep-learning method. *Physical Review A*, 110(6), 062609. <https://doi.org/10.1103/PhysRevA.110.062609>

Figures

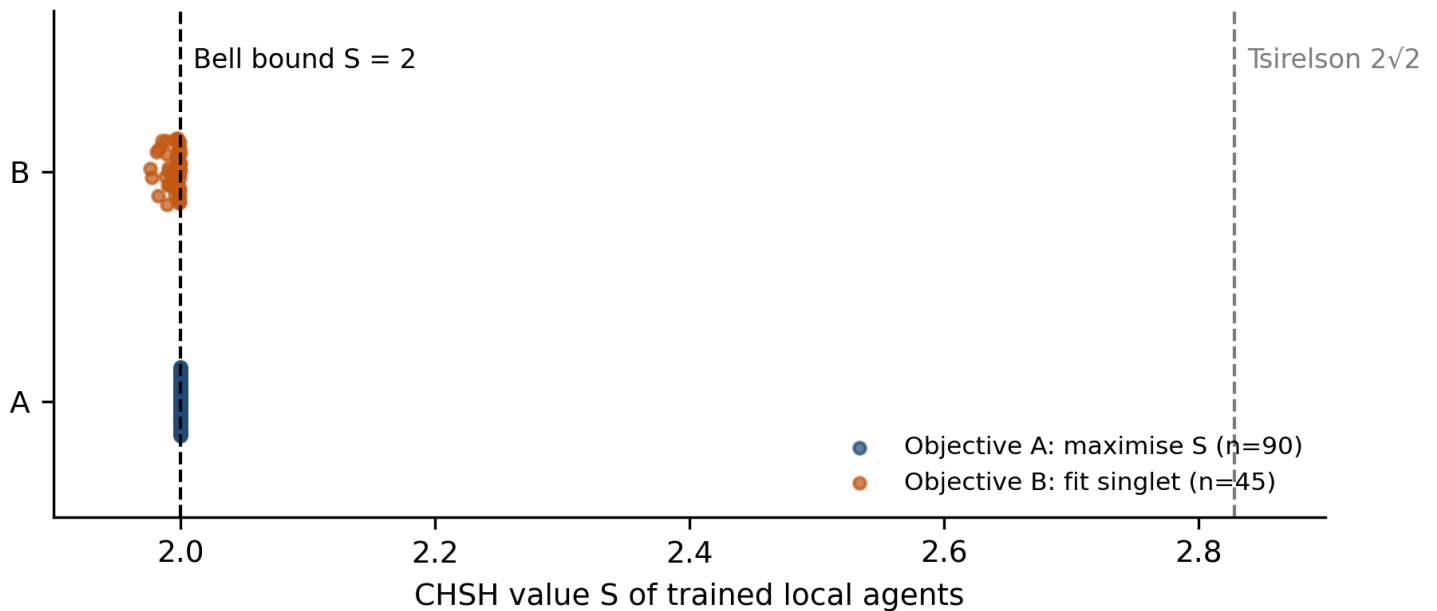


Figure 1

CHSH values of all trained local pairs. Dashed lines mark the Bell bound and the Tsirelson bound.

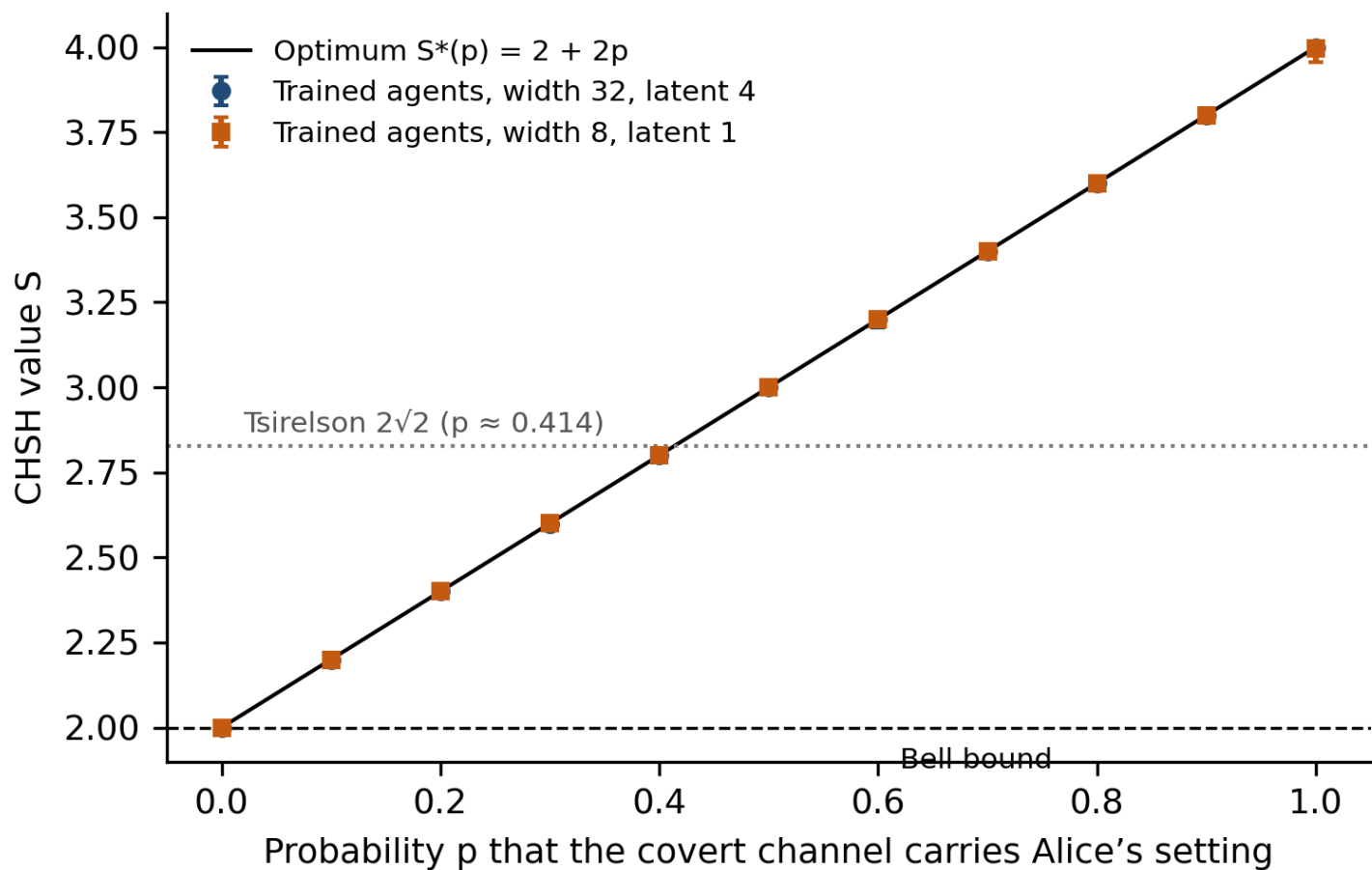


Figure 2

Learned CHSH value against channel rate. Bars show the range over ten seeds; the solid line is the optimum $2 + 2p$.

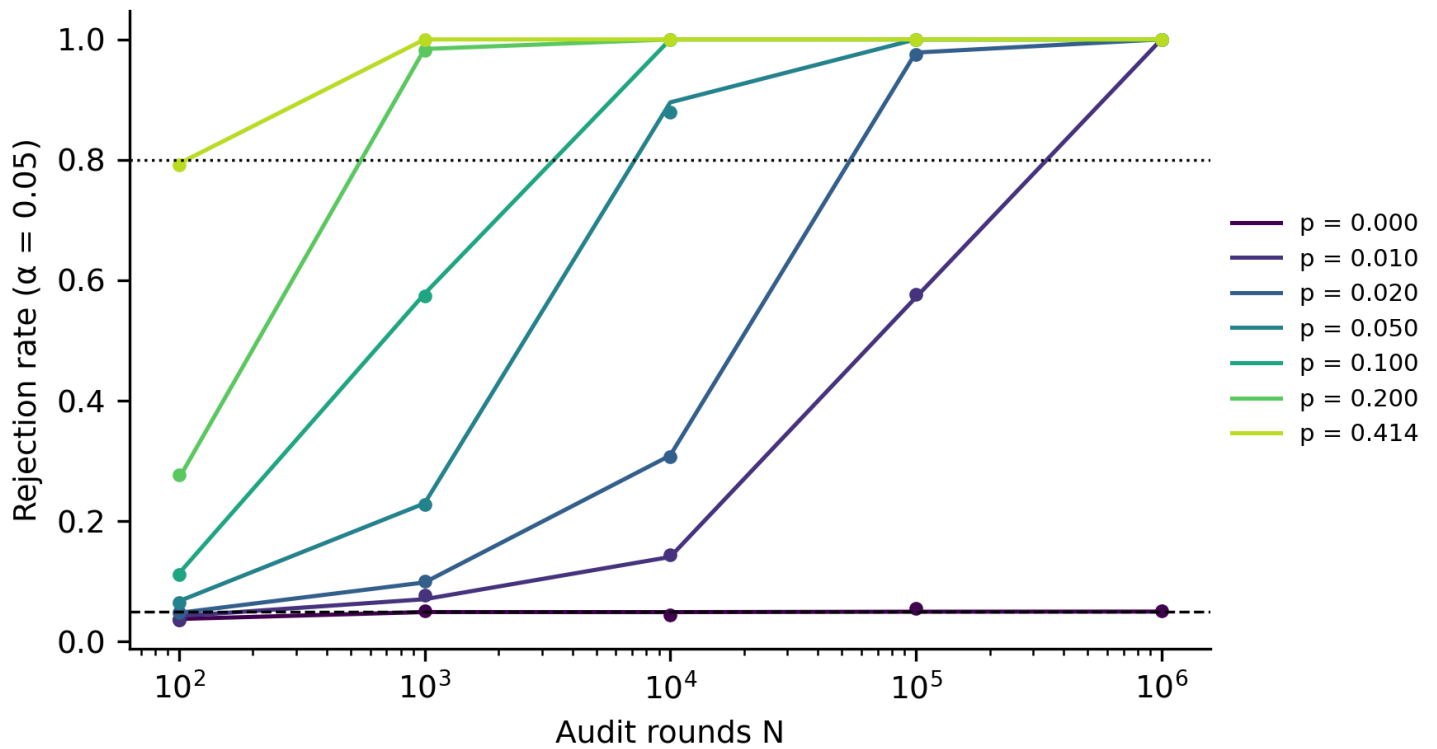


Figure 3

Rejection rate of the Bell audit by number of rounds and leak rate. Lines: exact binomial power; dots: Monte Carlo (2,000 audits). Dashed: $\alpha = 0.05$; dotted: 80% power.

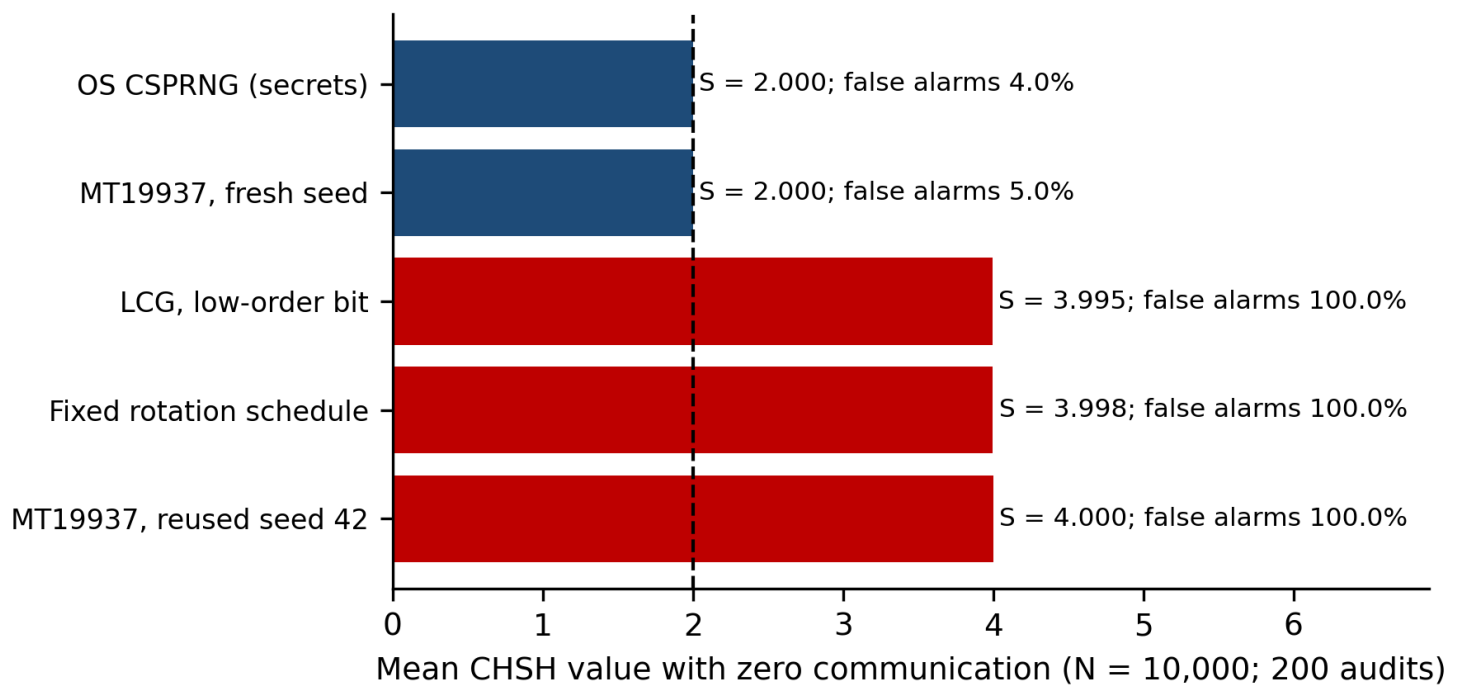


Figure 4

Mean CHSH value of agents with no communication under each setting generator. Red bars exceed the Bell bound.