Five AI Agents, One Hidden Source: The Independence Illusion in Multi-Agent Decisions

Kwan Hong TAN

Singapore University of Social Sciences, Singapore

DOI: https://doi.org/10.68050/JAMS.2026.489

Received: 15 Sept 2026; Accepted: 16 Sept 2026; Published: 22 September 2026

ABSTRACT

Multi-agent artificial intelligence systems often convert agreement into confidence. That conversion is valid only when the agents contribute sufficiently independent evidence. This paper develops the independence illusion, defined as the overstatement of decision reliability that occurs when correlated AI outputs are counted as independent confirmations. A common-shock model derives the probability and reliability of unanimous agreement among n agents with marginal accuracy p and pairwise correctness correlation ρ. The model yields a counterintuitive result: positive correlation makes unanimity more frequent while making it less diagnostic. For five agents that are each 75 percent accurate, an independence-based calculation assigns 99.59 percent reliability to unanimity. At a correlation of 0.60, the correct reliability under the common-shock model is only 78.37 percent, even though unanimity occurs on 69.53 percent of tasks. A fully specified Monte Carlo design with 1,000,000 synthetic binary decisions then compares one-lineage, three-lineage, and five-lineage architectures. Five-vote majority accuracy rises from 80.8 percent in the one-lineage architecture to 84.8 percent with three lineages and 88.9 percent with five. Requiring an independent 90 percent accurate challenger to agree with a unanimous one-lineage panel raises accepted-case accuracy to 97.0 percent, but reduces automatic coverage to 50.5 percent. A beta-binomial robustness model preserves the central result. The paper proposes provenance-adjusted effective evidence, vote-pattern calibration, disagreement preservation, and independent challenge gates. The findings do not show that multi-agent systems are generally unreliable. They show that agent count is not evidence count, and that model lineage must enter assurance claims whenever consensus is used to automate consequential decisions.

Keywords: multi-agent AI; correlated errors; consensus reliability; model diversity; AI assurance; human oversight

INTRODUCTION

Multi-agent artificial intelligence is moving from laboratory demonstrations into decision workflows. One model drafts an answer, another critiques it, several instances vote, and a judge model selects the final response. The architecture is attractive because agreement appears to supply internal verification at machine speed. Self-consistency and debate protocols have improved performance on several reasoning tasks, and the broader ensemble-learning literature gives a sound reason to combine fallible predictors when their errors differ (Dietterich, 2000; Wang et al., 2023; Du et al., 2023).

The same intuition can fail when several agents inherit the same blind spot. Agents may share a base model, provider, training corpus, alignment procedure, system prompt, retrieval index, toolchain, or judge. Their surface explanations can vary while the evidence-producing process remains substantially common. Kim et al. (2025) provide direct empirical motivation: across more than 350 large language models, they find substantial error correlation and report that, on one leaderboard, models agree 60 percent of the time when both are wrong. The operational problem is therefore not merely whether each agent is accurate. It is whether their joint errors are independent enough for agreement to justify higher confidence.

This problem matters because agreement is cognitively persuasive. Human operators can display automation bias under workload, and users may place substantial weight on algorithmic advice (Parasuraman & Manzey,

2010; Logg et al., 2019). Steyvers et al. (2025) further show that people can overestimate the accuracy of large-language-model answers and that longer explanations may increase human confidence without increasing answer accuracy. A multi-agent interface can compound that calibration problem by displaying several fluent rationales, votes, or roles as if they were independent witnesses.

The paper calls this mechanism the **independence illusion**. It occurs when a decision-maker maps nominal agreement into a confidence claim using an independence assumption that the system does not satisfy. The illusion has a technical component, because correlated signals contain less incremental information, and a behavioral component, because visible plurality can be mistaken for evidential diversity. It is especially consequential when consensus triggers automatic action in healthcare, finance, education, law, research, cybersecurity, or public administration.

The analysis makes four contributions. First, it derives an exact unanimity reliability function under a transparent common-shock dependence model. Second, it shows that positive correlation simultaneously increases the frequency of unanimity and reduces the probability that a unanimous decision is correct. Third, it compares model-lineage architectures through a reproducible simulation and a beta-binomial robustness analysis. Fourth, it converts the results into a governance design centered on provenance, empirical error correlation, preserved disagreement, and independent challenge. The contribution extends research on correlated LLM errors by showing how dependence becomes a false confidence claim at the decision gate, and by identifying controls that operate on the evidence structure rather than on agent count alone.

LITERATURE REVIEW AND RESEARCH GAP

Collective accuracy and the independence condition

The classical case for majority aggregation rests on a simple mechanism. If each voter is more likely than chance to be correct and votes are conditionally independent, a sufficiently large majority becomes highly reliable. Boland (1989) formalizes important versions of this logic. Research on dependent juries, however, shows that the result changes when votes are correlated (Berg, 1993; Ladha, 1992). Dependence limits the amount of new information supplied by an additional vote.

The same principle appears in statistics and forecasting. Kish's (1965) design effect shows how within-cluster correlation reduces effective sample size. Expert-aggregation research likewise treats dependence as a central problem because experts may rely on overlapping information (Clemen & Winkler, 1999). DeGroot's (1974) consensus model explains how repeated social updating can produce agreement, but agreement itself does not establish accuracy. Lorenz et al. (2011) demonstrate experimentally that social influence can narrow opinion diversity and increase confidence while weakening the wisdom-of-crowds effect.

Machine-learning ensembles operationalize a related insight. Random forests benefit from both individual predictor strength and decorrelation among trees (Breiman, 2001). Dietterich (2000) identifies error diversity as a foundation of ensemble gains, while Kuncheva and Whitaker (2003) examine how diversity measures relate to ensemble accuracy. An ensemble of identical or highly coupled errors is numerically plural but epistemically thin.

Foundation models and inherited failure

Foundation-model deployment creates several routes to shared failure. The same upstream model can be wrapped in different personas, prompts, or interfaces. Fine-tuned descendants can retain common representations and shortcuts. Retrieval-augmented agents can query the same incomplete corpus, and apparently separate systems can depend on a common provider or benchmark culture. Geirhos et al. (2020) show that deep networks can rely on shortcuts that succeed in-distribution but fail outside intended conditions. Messeri and Crockett (2024) warn that AI can produce an illusion of understanding by making outputs appear broad and authoritative while narrowing inquiry.

Algorithmic monoculture research moves the issue from a single model to a system of decision-makers. Kleinberg and Raghavan (2021) show that widespread reliance on the same algorithm can reduce social welfare even if the algorithm is individually more accurate than alternatives. Kim et al. (2025) then demonstrate that LLM errors remain correlated across models and that common providers and architectures explain part of the dependence. Their work establishes empirical relevance. The present paper studies a different downstream object: the reliability assigned to a particular vote pattern when correlation is hidden from the decision rule.

Multi agent debate and consensus seeking

Multi-agent debate asks several language-model instances to propose, criticize, and revise answers. Du et al. (2023) report gains in factuality and reasoning under a society-of-minds protocol. Wang et al. (2023) obtain strong improvements from self-consistency, which samples diverse reasoning paths and aggregates their answers. These methods demonstrate that repeated inference can help when sampling produces useful variation.

Recent results also identify limits. Smit et al. (2024) find that debate does not reliably outperform self-consistency under comparable conditions and is sensitive to protocol choices. Chen et al. (2026) show that consensus-seeking debate can discard informative disagreement and that competitive debate can encourage persuasion rather than truth seeking. Their collaborative protocol performs better because it rewards informative and truthful additions. Together, these findings imply that the quality of interaction matters more than the mere existence of several speaking agents.

Yet most operational dashboards still summarize multi-agent performance through nominal votes, agreement rates, or final-answer accuracy. A panel can therefore improve average accuracy while remaining dangerously miscalibrated on the subset of cases routed automatically because all agents agree. This selective-decision problem is the focus here.

Human interpretation and the missing decision layer

Human factors research shows why a calibration error at the interface can survive formal model evaluation. Parasuraman and Manzey (2010) distinguish automation complacency from automation bias and explain how attention allocation and perceived reliability affect monitoring. Logg et al. (2019) document algorithm appreciation in several judgment tasks, although other work finds algorithm aversion after visible mistakes (Dietvorst et al., 2015). The appropriate conclusion is contextual: users can either underweight or overweight automation, depending on task, experience, control, and presentation.

Multi-agent consensus adds a presentation cue that existing calibration research has not adequately separated from underlying evidence. Five answer boxes, five role labels, or five explanations may suggest five independent checks even when they are stochastic samples from the same model. Steyvers et al. (2025) show that explanation characteristics can alter user confidence. The independence illusion predicts that visible plurality will have a similar effect unless provenance and dependence are disclosed.

The research gap lies between three established literatures. Ensemble theory explains why independence and diversity matter. LLM research documents correlated errors and protocol-sensitive debate. Human-factors research explains why users may defer to confident automation. What remains underdeveloped is a compact model that translates hidden dependence into vote-pattern reliability, quantifies the resulting confidence inflation, and yields deployable assurance controls. Figure 1 summarizes the resulting causal pathway.

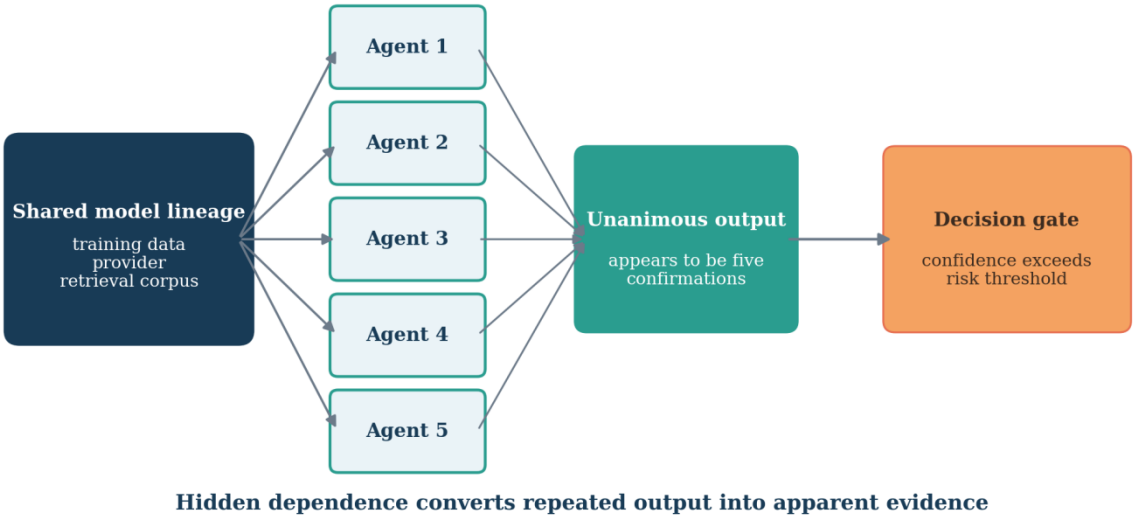

Figure 1 The independence illusion in a multi-agent decision gate. Different agent labels can conceal a common evidence-producing process.

CONCEPTUAL MODEL

Binary decision setting

Consider a binary decision with objectively correct label *Y*. Agent *i* produces a label and *Xi* equals 1 when that label is correct. For exposition, all agents have the same marginal competence *p*, with *p* greater than one half. The model first studies unanimity because many systems treat it as a high-confidence condition. Majority decisions are then evaluated in simulation.

$$X_i \in \{0,1\}, \qquad \Pr(X_i = 1) = p > \frac{1}{2}, \qquad i = 1, \dots, n \tag{1}$$

Common shock dependence

A common-shock mixture supplies a transparent dependence structure. With probability ρ, all agents inherit one common correctness draw *C*. With probability 1 minus ρ, the agents draw correctness independently. The common and independent draws all have success probability *p*. This construction preserves each agent's marginal accuracy and produces pairwise correctness correlation exactly equal to ρ. It represents a shared hidden cause such as a common model blind spot, missing source, benchmark artifact, or tool failure.

$$X_i = SC + (1 - S)I_i, \quad S \sim \text{Bernoulli}(\rho), \quad C, I_i \sim \text{Bernoulli}(p) \tag{2}$$

Unanimity frequency and reliability

Let U_n denote the event that all agents give the same label. Unanimity occurs automatically during a common shock. During the independent regime it occurs when every agent is correct or every agent is wrong. Its total probability is therefore:

$$\Pr(U_n) = \rho + (1 - \rho)[p^n + (1 - p)^n] \tag{3}$$

The reliability of unanimity is the probability that the unanimous label is correct, conditional on unanimity:

$$R_n(p,\rho) = \Pr(X_1 = \cdots = X_n = 1 \mid U_n) = \frac{\rho p + (1-\rho)p^n}{\rho + (1-\rho)[p^n + (1-p)^n]} \quad (4)$$

A decision rule that incorrectly assumes independence reports the following reliability:

$$\tilde{R}_n(p) = \frac{p^n}{p^n + (1-p)^n}, \qquad \Delta_n(p,\rho) = \tilde{R}_n(p) - R_n(p,\rho) \quad (5)$$

The paper calls Δ_n the **consensus reliability gap**. It measures confidence inflation for the unanimous subset, not the overall error rate of the system.

Provenance adjusted effective evidence

A variance-equivalent evidence count can be adapted from survey design. For an equicorrelated panel, nominal size n contracts to $n$eff. The measure is diagnostic rather than a complete posterior because higher-order dependence still matters.

$$n_{\text{eff}} = \frac{n}{1 + (n-1)\rho} \quad (6)$$

For heterogeneous weights and an estimated correctness-correlation matrix R, the generalized variance-equivalent count is:

$$n_{\text{eff}}(\mathbf{w}, \mathbf{R}) = \frac{(\mathbf{1}^\top \mathbf{w})^2}{\mathbf{w}^\top \mathbf{R} \mathbf{w}} \quad (7)$$

Equation 7 makes the governance principle explicit. Adding another agent increases effective evidence only to the extent that its weight is not offset by dependence with existing agents.

Independent challenge

Suppose an independent challenger with accuracy q evaluates a unanimous panel decision whose calibrated reliability is R. The system accepts automatically only when the challenger agrees with the panel. Conditional reliability after agreement becomes:

$$R^+ = \frac{Rq}{Rq + (1-R)(1-q)} \quad (8)$$

For q greater than one half, R+ exceeds R. The gain comes from new evidence, not from another paraphrase of the same evidence. The price is reduced coverage because the gate abstains when panel and challenger disagree.

ANALYTICAL PROPOSITIONS

Proposition 1 For n greater than 1 and p greater than one half, unanimity reliability R_n declines as positive common-shock dependence ρ increases, while the probability of unanimity increases. Thus, correlation makes consensus more visible and less informative at the same time.

Proposition 2 For any fixed positive ρ, the reliability of unanimity converges to individual accuracy p as n grows without bound, whereas the independence-based reliability converges to 1. Nominal panel expansion can therefore increase claimed certainty without increasing asymptotic evidential reliability.

Proposition 3 For positive ρ, the variance-equivalent evidence count converges to 1 divided by ρ. At ρ equal to 0.60, an arbitrarily large equicorrelated panel contains at most 1.67 variance-equivalent independent observations.

Proposition 4 Within the hierarchical mixture evaluated here, holding marginal competence and nominal panel size constant while replacing within-lineage replication with distinct lineages lowers average dependence and improves both majority accuracy and the reliability of unanimous agreement.

Proposition 5 An independent challenger with accuracy above chance strictly increases the reliability of an accepted consensus under the agreement gate in Equation 8, although automatic coverage falls. The optimal gate depends on error cost and review capacity.

METHODOLOGY

Research design

The study combines analytical derivation with a deterministic Monte Carlo stress test. It does not evaluate any named commercial model and does not claim that the baseline parameters describe a particular deployment. The simulation asks whether the proposed mechanism appears under a known data-generating process and how architecture changes the decision frontier.

The unit of analysis is one binary task. Ground truth is balanced between the two labels, and correctness is symmetric across labels. Every agent has marginal accuracy 0.75. One million tasks are generated for each architecture using NumPy's PCG64 generator and a fixed seed based on 20260914. At this sample size, Monte Carlo standard errors for the reported proportions remain below approximately 0.05 percentage points, including the smallest accepted subset.

Three five-agent architectures isolate lineage diversity. H5 places all five agents in one lineage with within-lineage correctness correlation 0.60. D221 allocates the agents across three lineages of sizes two, two, and one; within-lineage correlation remains 0.60 and cross-lineage correlation is 0.05. U5 assigns every agent to a separate lineage with cross-lineage correlation 0.05. A hierarchical mixture generates a global common shock for cross-lineage dependence, a lineage shock for residual within-lineage dependence, and otherwise independent draws.

Table 1 defines the study's core constructs and links each to an observable implication.

Table 1 Core constructs and observable implications

Construct	Definition	Observable implication
Independence illusion	Confidence added by nominal plurality exceeds information added by the underlying processes	Reported consensus confidence exceeds vote-pattern accuracy
Consensus reliability gap	Difference between independence-based and correlation-aware reliability	Positive gap that widens with dependence
Evidence lineage	Shared upstream source of error across agents	Higher within-lineage than cross-lineage error correlation
Effective evidence count	Variance-equivalent number of independent signals	Smaller than nominal agent count under positive dependence

Independent challenge	Verification by a process with materially separate failure modes	Higher accepted-case accuracy with lower coverage

Table 2 reports the complete baseline parameter record used for reproduction.

Table 2 Baseline simulation parameters

Parameter	**Value**	**Interpretation**
Tasks per architecture	1,000,000	Synthetic binary decisions
Agents	5	Fixed nominal panel size
Marginal accuracy p	0.75	Equal competence baseline
Within-lineage correlation	0.6	Strong shared failure exposure
Cross-lineage correlation	0.05	Weak residual common exposure
Independent challenger q	0.9	Separate high-quality verification
Random seed	20260914 plus architecture offset	Exact reproducibility

Decision policies and outcome measures

The majority policy decides every task using at least three of five votes. The unanimity policy decides only tasks on which all five agents agree and abstains otherwise. The challenged-unanimity policy adds a 90 percent accurate independent verifier and accepts only when it agrees with the unanimous panel. For each policy, coverage is the share of all tasks decided automatically; accuracy is evaluated among automatically decided tasks. Errors per 10,000 automated decisions translate the conditional error rate into an operational quantity.

The study also reports average pairwise correctness correlation and a lineage-balanced majority diagnostic. The latter selects one predetermined representative from each lineage and applies majority rule across lineage representatives. It is not proposed as an optimal aggregator. It shows how much nominal duplication contributes after each evidence lineage receives equal representation.

Robustness checks vary marginal accuracy from 0.60 to 0.90, correlation from 0.10 to 0.80, and panel size from three to nine. A beta-binomial model supplies an alternative exchangeable dependence structure with the same marginal accuracy and pairwise correlation. The two models allocate higher-order dependence differently, so agreement reliability need not be identical. A conclusion that survives both is less likely to be an artifact of the common-shock mixture.

RESULTS AND ANALYSIS

Correlation reverses the meaning of frequent agreement

Table 3 evaluates Equations 3 through 6 for five agents with marginal accuracy 0.75. Under independence, unanimity is uncommon at 23.83 percent but highly diagnostic at 99.59 percent. As correlation rises, unanimity becomes common because agents fail or succeed together. At ρ equal to 0.60, unanimity occurs on 69.53 percent of tasks, yet its reliability is only 78.37 percent. A dashboard that continues to report the independence-based 99.59 percent therefore overstates reliability by 21.22 percentage points.

Table 3 Analytical effect of correlation on five-agent unanimity

ρ	Unanimity percent	Actual reliability percent	Naive reliability percent	Gap points	n effective
0	23.83	99.59	99.59	0	5
0.1	31.45	91.77	99.59	7.82	3.57
0.2	39.06	87	99.59	12.59	2.78
0.4	54.3	81.47	99.59	18.12	1.92
0.6	69.53	78.37	99.59	21.22	1.47
0.8	84.77	76.38	99.59	23.21	1.19

Figure 2 traces the same reversal across panel sizes, separating the reliability and coverage effects.

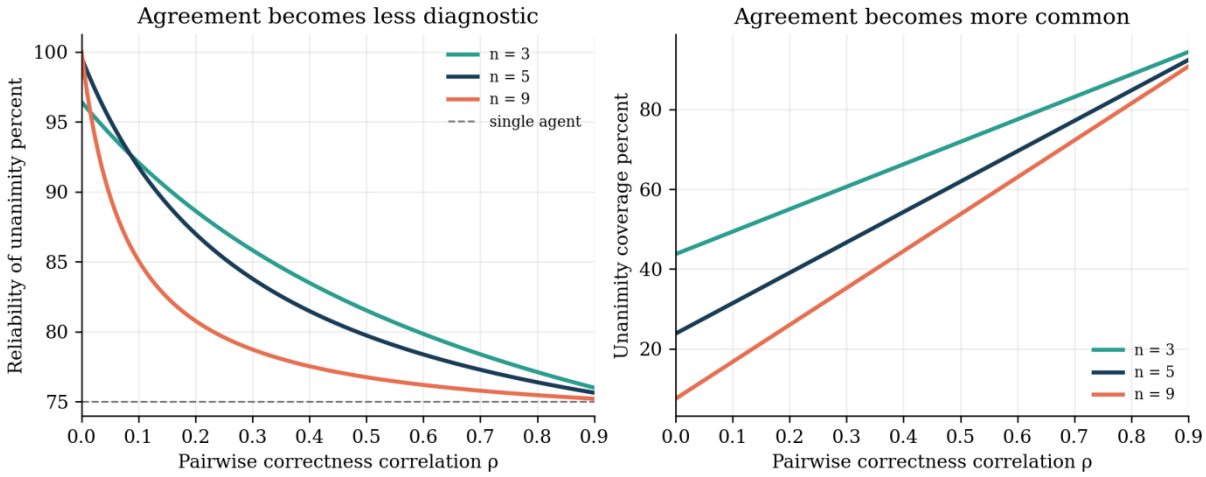

Figure 2 Unanimity becomes more frequent and less diagnostic as correctness correlation rises. Marginal agent accuracy is 75 percent.

Panel expansion does not solve the problem. Under independence, moving from three to nine agents raises the reliability of unanimity from 96.43 percent to 99.995 percent. With ρ equal to 0.60, the corresponding values are only 79.84 percent and 76.19 percent. More agents make the independence claim look stronger while the common-shock component dominates the selected unanimous cases. This is Proposition 2 in finite form.

Lineage diversity improves the five agent panel

Table 4 Monte Carlo results by five-agent architecture

Architecture	Average pair correlation	Majority accuracy percent	Unanimity coverage percent	Unanimity accuracy percent	Errors per 10000 unanimous decisions
H5 one lineage	0.601	80.82	69.55	78.31	2169
D221 three lineages	0.161	84.75	37.79	94.88	512
U5 five lineages	0.051	88.9	27.65	95.03	497

As Table 4 shows, the one-lineage panel reaches 80.82 percent majority accuracy, compared with 84.75 percent for three lineages and 88.90 percent for five lineages. The gain occurs without changing nominal panel size or marginal agent competence. Unanimous one-lineage decisions produce about 2169 errors per 10,000 accepted cases, compared with 497 in the five-lineage architecture. The residual gap between 95.03 percent

observed reliability and 99.59 percent naive confidence in U5 shows that even modest cross-lineage correlation should not be ignored. Figure 3 makes the architecture contrast visible against the false-independence benchmark.

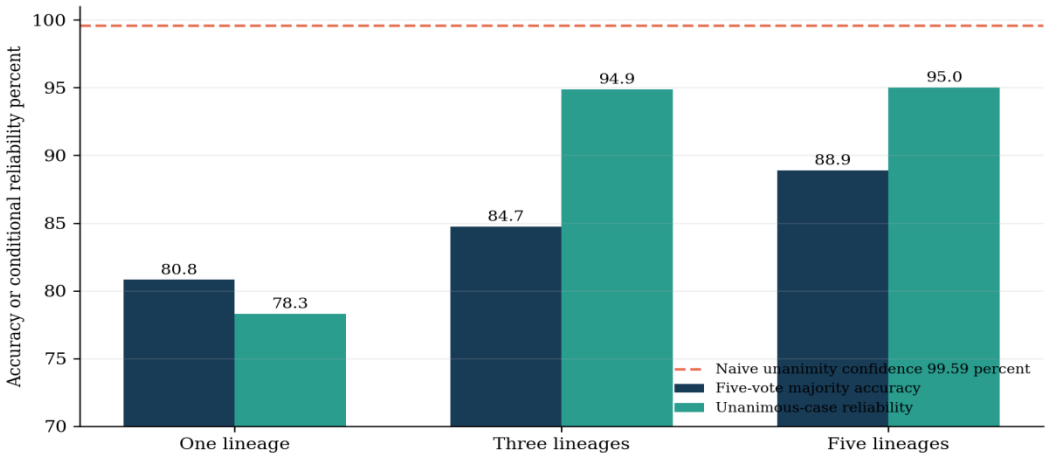

Figure 3 Majority accuracy and unanimous-case reliability improve as evidence lineages diversify. The dashed line is the confidence implied by false independence.

Independent challenge shifts the accuracy coverage frontier

Table 5 Effect of an independent 90 percent accurate challenge

Architecture	Unanimity coverage percent	Unanimity accuracy percent	Challenged coverage percent	Challenged accuracy percent	Errors per 10000 challenged decisions
H5 one lineage	69.55	78.31	50.54	97.04	296
D221 three lineages	37.79	94.88	32.46	99.39	61
U5 five lineages	27.65	95.03	23.78	99.43	57

Table 5 shows that, for H5, independent challenge raises accepted-case accuracy from 78.31 percent to 97.04 percent and reduces residual errors from about 2169 to 296 per 10,000 accepted decisions. Coverage falls from 69.55 percent to 50.54 percent. The control therefore converts some automatic decisions into abstentions. Figure 4 maps that tradeoff as an accuracy-coverage frontier. It is desirable when the expected cost of a false decision exceeds the cost of human review or delayed action.

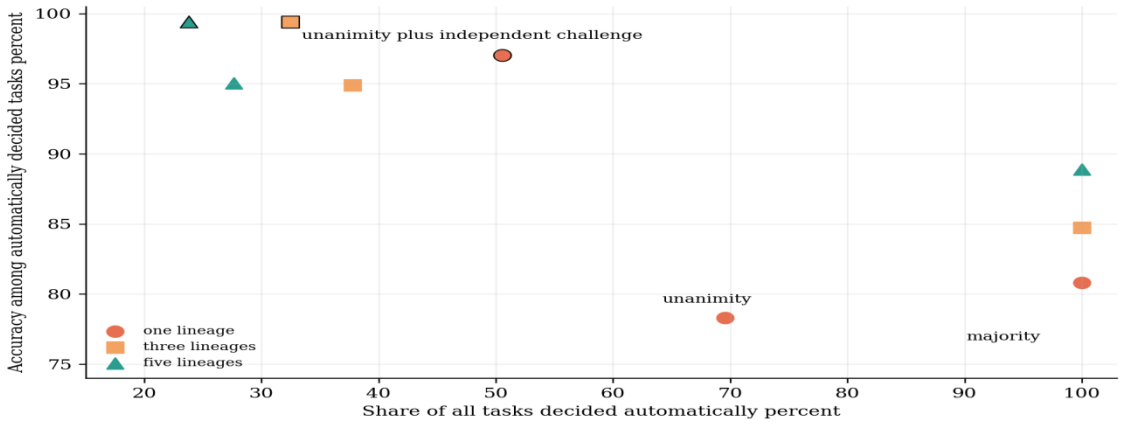

Figure 4 Accuracy and coverage frontier for majority, unanimity, and challenged unanimity. Black-edged points denote challenged decisions.

Sensitivity and alternative dependence

Table 6 Sensitivity of five-agent unanimity to competence and correlation

p	ρ	Coverage percent	Actual reliability percent	Naive reliability percent	Gap points
0.6	0.2	27.04	67.38	88.36	20.98
0.6	0.6	63.52	61.57	88.36	26.79
0.7	0.2	33.64	81.59	98.57	16.99
0.7	0.6	66.82	72.92	98.57	25.66
0.75	0.2	39.06	87	99.59	12.59
0.75	0.6	69.53	78.37	99.59	21.22
0.8	0.2	46.24	91.29	99.9	8.61
0.8	0.6	73.12	83.57	99.9	16.33
0.9	0.2	67.24	97.02	100	2.97
0.9	0.6	83.62	92.82	100	7.17

Table 7 Robustness across common-shock and beta-binomial dependence

ρ	Common-shock reliability percent	Beta-binomial reliability percent	Common-shock coverage percent	Beta-binomial coverage percent
0.1	91.77	97.92	31.45	31.72
0.2	87	95.45	39.06	39.29
0.4	81.47	89.73	54.3	54.19
0.6	78.37	84.16	69.53	69.21
0.8	76.38	79.23	84.77	84.49

Tables 6 and 7 show that the consensus reliability gap remains positive throughout the sensitivity grid whenever correlation is positive. It is largest when naive independence converts moderate competence into near-certainty. The beta-binomial model produces less severe confidence inflation than the common-shock model at the same pairwise correlation because it allocates higher-order dependence differently. At p equal to 0.75 and ρ equal to 0.60, beta-binomial unanimity reliability is 84.16 percent rather than 78.37 percent, but both remain far below the 99.59 percent independence claim. Pairwise correlation alone is therefore insufficient for exact certification, yet it is sufficient to reject naive multiplication of evidence.

DISCUSSION

Theoretical implications

The analysis clarifies a distinction often blurred in multi-agent evaluation. Consensus can improve prediction accuracy, but consensus is not itself an independent source of evidence. Reliability depends on the joint distribution of errors. When dependence is strong, a unanimous panel mostly reveals that a common mechanism produced a stable answer. It says much less about whether that mechanism is correct.

The first contribution is the joint reversal in Proposition 1. Organizations may interpret a rising agreement rate as evidence that orchestration is becoming more reliable. Under the common-shock model, the same increase can be caused by stronger coupling. Agreement then becomes easier to obtain and weaker as a diagnostic signal. Monitoring agreement without conditional accuracy can therefore reward epistemic monoculture.

The second contribution is the distinction between agent diversity and evidence diversity. Temperature changes, role prompts, debate rounds, and stylistic personas can generate varied text. They do not necessarily generate independent failure modes. A procurement team that buys access to several interfaces can still depend on one provider, one retrieval corpus, or one benchmark lineage. Conversely, a smaller panel with genuinely different evidence channels can be more reliable than a larger homogeneous panel.

The third contribution is selective calibration. Overall accuracy can conceal severe miscalibration in the exact subset that a system routes automatically. A panel may be reasonably accurate on average yet attach near-certainty to correlated unanimous errors. Assurance should therefore estimate accuracy separately for vote patterns, task classes, and provenance configurations.

Relation to existing AI assurance research

Tan (2026a) shows that fixed assurance costs can affect which firms can adopt AI safely. The present result adds a design constraint to that economic problem: duplicating same-lineage agents can create the appearance of assurance without commensurate evidence. Shared evaluation infrastructure can reduce cost, but shared inference dependencies can increase correlation. Governance should distinguish reusable controls from common epistemic failure points.

Runtime assurance for agentic systems requires policy gates that respond to risk and context (Tan, 2026c). Consensus is a candidate gate input, but this paper shows that raw vote count is insufficient. A runtime controller should incorporate provenance, validated vote-pattern reliability, task novelty, and verifier independence. The coordination compression problem identified by Tan (2026b) is also relevant because automated production can scale faster than human verification. Independence-aware routing uses scarce review capacity where apparent consensus is least trustworthy or expected harm is greatest.

NIST's AI Risk Management Framework and Generative AI Profile emphasize measurement, monitoring, documentation, and human oversight (Tabassi, 2023; Autio et al., 2024). The proposed controls make those principles specific to multi-agent decisions. They do not require full access to model weights. They require evidence about lineages, validation-set error dependence, and the behavior of decision gates.

Applications across high consequence domains

Healthcare Several agents trained on overlapping medical corpora may reproduce the same omitted contraindication. Automatic acceptance require source-grounded review and a challenger that uses a distinct evidence path.

Finance Credit, fraud, or investment panels can share market data and vendor models. Correlated errors can create synchronized false reassurance and concentrated loss.

Education Several tutoring agents can agree on an incorrect explanation because they share the same misconception. Preserving dissent and checking against authoritative curricular sources is more valuable than adding personas.

Research Literature-synthesis agents querying the same index can manufacture apparent consensus by repeatedly retrieving the same narrow evidence base. Source overlap should be visible in the final confidence claim.

Law and public administration Agreement among models does not substitute for reasons that can be traced to valid authority, nor does it remove the need for human accountability in consequential decisions.

GOVERNANCE AND DESIGN RECOMMENDATIONS

The model supports compact assurance architecture. It does not demand that every agent use a different

provider, because nominal vendor diversity can be expensive and still fail to produce independent evidence. It requires the decision owner to measure dependence and match controls to expected harm.

Table 8 translates the analytical findings into auditable controls and retained evidence.

Table 8 Controls for independence aware multi-agent assurance

Control	Minimum implementation	Failure addressed	Evidence retained
Provenance map	Record base model, provider, fine-tune, prompt, retrieval corpus, tools, and judge	Hidden common lineage	Versioned dependency graph
Vote-pattern calibration	Estimate conditional accuracy for majority, split, and unanimous patterns by task class	Global accuracy masking selective error	Reliability curves with sample counts
Correlation card	Report pairwise and lineage-level correctness correlation on representative tests	False independence	Correlation matrix and uncertainty intervals
Disagreement preservation	Store initial answers before agents observe one another	Premature convergence	Pre-debate and post-debate votes
Independent challenge	Route high-risk consensus to a separate model, tool, dataset, or human	Common-shock error	Challenge result and provenance
Abstention policy	Set risk-tier thresholds for review rather than forcing a decision	Confidence laundering into automation	Reason for escalation
Drift monitoring	Re-estimate dependence after model, corpus, or prompt changes	Stale calibration	Time-stamped validation reports

Measure the panel as a system. Individual model cards cannot establish the reliability of a multi-agent decision. Validation must run the complete orchestration, including agent communication, retrieval, tools, judge, and escalation logic.

Treat persona variation as unproven diversity. Role prompts can improve search over reasoning paths, but they should receive no independence credit until error data show that they reduce dependence on relevant task failures.

Calibrate the patterns that trigger automation. If unanimity authorizes action, report the empirical accuracy of unanimous cases with confidence intervals. Do not infer it by multiplying individual accuracies unless conditional independence has been tested and defended.

Preserve first judgments. Record agent answers before debate. Convergence after exposure to peers can be useful, but it erases the evidence needed to distinguish independent agreement from social imitation, a concern consistent with Lorenz et al. (2011) and Chen et al. (2026).

Buy independence where harm is concentrated. A separate verification model, retrieval corpus, deterministic tool, domain expert, or direct source check can add more assurance than several same-lineage agents. The best challenger is selected for different failure modes, not merely higher benchmark accuracy.

Expose uncertainty at the interface. A consensus badge should state nominal votes, known lineages, calibrated reliability, sample size, and whether an independent challenge occurred. This reduces the risk that visual plurality becomes confidence without evidence.

LIMITATIONS AND EMPIRICAL RESEARCH AGENDA

The study is structural rather than empirical. Its numerical results follow from stated parameters and do not estimate the correlation of any deployed system. The common-shock model is deliberately simple. Real agent errors can be negatively correlated, task dependent, asymmetric across labels, and shaped by communication. Higher-order dependence is not identified by pairwise correlation alone, as the beta-binomial comparison demonstrates.

Equal marginal accuracy is another simplification. Production panels combine models with different competence, calibration, cost, and refusal behavior. The generalized effective evidence count in Equation 7 allows unequal weights, but exact decision reliability requires a joint model fitted to validation data. The independent challenger is also idealized. A human reviewer or separate model may remain correlated with the panel through shared documents, professional conventions, or anchoring on the panel's answer.

The binary setting does not capture open-ended generation, where correctness can be multidimensional and partial. Future studies should define task-specific loss functions, code semantic error categories, and estimate dependence at the level of claims rather than whole responses. Multiclass and continuous decisions require corresponding aggregation models.

A direct behavioral experiment should test the interface component of the independence illusion. Participants could evaluate identical recommendations presented as one answer, five samples from one model, five named personas from one model, or five independently sourced models. The design should vary disclosure of shared provenance and measure confidence, willingness to automate, source checking, and calibration. The prediction is that nominal plurality raises confidence even when dependence is disclosed incompletely.

A field benchmark should evaluate homogeneous and heterogeneous panels on temporally shifted, adversarial, and domain-specific tasks. Researchers should publish individual accuracy, joint error matrices, vote-pattern reliability, pre-debate disagreement, post-debate convergence, token cost, latency, and human-review burden. The most decision-relevant question is not whether debate raises average benchmark accuracy, but whether the system's confidence and routing rules remain calibrated under shared failure.

CONCLUSION

Multi-agent AI can improve decisions when agents contribute different information or failure modes. It can also manufacture false confidence when repeated outputs share a hidden source. The common-shock model isolates the mechanism: correlation makes unanimity more frequent while reducing its reliability. Five 75 percent accurate agents can appear to supply 99.59 percent certainty under independence, yet provide only 78.37 percent reliability when pairwise correctness correlation reaches 0.60.

The practical conclusion is concise. Agent count is not evidence count. Multi-agent assurance should be based on provenance, measured joint errors, calibrated vote patterns, preserved disagreement, and independent challenge. Systems that cannot establish these properties should treat consensus as a reason to inspect, not as a licence to automate.

DECLARATIONS

Conflict of interest The author declares no conflict of interest.

Funding No external funding was received for this study.

Ethics approval Not applicable. The study uses mathematical analysis and synthetic computational experiments and involves no human participants, animals, or personal data.

Data and code availability No external dataset was used. The complete data-generating process, parameter

values, random seed, equations, and reproducible algorithm are reported in the paper. A public repository may be added during submission.

REFERENCES

1. Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial Intelligence Risk Management Framework Generative Artificial Intelligence Profile (NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1
2. Berg, S. (1993). Condorcet's jury theorem, dependency among jurors. Social Choice and Welfare, 10, 87-95. https://doi.org/10.1007/BF00187435
3. Boland, P. J. (1989). Majority systems and the Condorcet jury theorem. Journal of the Royal Statistical Society Series D The Statistician, 38(3), 181-189. https://doi.org/10.2307/2348873
4. Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324
5. Chen, Y., Niu, G., Cheng, J., Han, B., & Sugiyama, M. (2026). When and why does multi-agent debate fail and does it really underperform? arXiv. https://doi.org/10.48550/arXiv.2510.20963
6. Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. Risk Analysis, 19(2), 187-203. https://doi.org/10.1023/A:1006917509560
7. DeGroot, M. H. (1974). Reaching a consensus. Journal of the American Statistical Association, 69(345), 118-121. https://doi.org/10.2307/2285509
8. Dietterich, T. G. (2000). Ensemble methods in machine learning. In J. Kittler & F. Roli (Eds.), Multiple Classifier Systems (Lecture Notes in Computer Science, Vol. 1857, pp. 1-15). Springer. https://doi.org/10.1007/3-540-45014-9_1
9. Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology General, 144(1), 114-126. https://doi.org/10.1037/xge0000033
10. Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. arXiv. https://doi.org/10.48550/arXiv.2305.14325
11. Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence, 2, 665-673. https://doi.org/10.1038/s42256-020-00257-z
12. Kim, E., Garg, A., Peng, K., & Garg, N. (2025). Correlated errors in large language models. arXiv preprint arXiv:2506.07962. https://doi.org/10.48550/arXiv.2506.07962
13. Kish, L. (1965). Survey sampling. John Wiley and Sons.
14. Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. Proceedings of the National Academy of Sciences, 118(22), e2018340118. https://doi.org/10.1073/pnas.2018340118
15. Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning, 51, 181-207. https://doi.org/10.1023/A:1022859003006
16. Ladha, K. K. (1992). The Condorcet jury theorem, free speech, and correlated votes. American Journal of Political Science, 36(3), 617-634. https://doi.org/10.2307/2111584
17. Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90-103. https://doi.org/10.1016/j.obhdp.2018.12.005
18. Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. Proceedings of the National Academy of Sciences, 108(22), 9020-9025. https://doi.org/10.1073/pnas.1008636108
19. Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. Nature, 627, 49-58. https://doi.org/10.1038/s41586-024-07146-0
20. Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381-410. https://doi.org/10.1177/0018720810376055

21. Smit, A. P., Grinsztajn, N., Duckworth, P., Barrett, T. D., & Pretorius, A. (2024). Should we be going MAD? A look at multi-agent debate strategies for LLMs. Proceedings of the 41st International Conference on Machine Learning, 235, 45883-45905. https://proceedings.mlr.press/v235/smit24a.html
22. Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., & Smyth, P. (2025). What large language models know and what people think they know. Nature Machine Intelligence, 7, 221-231. https://doi.org/10.1038/s42256-024-00976-7
23. Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1
24. Tan, K. H. (2026a). The AI assurance cost paradox: Fixed governance costs, heterogeneous-firm adoption, and market concentration. Journal of Advanced Multidisciplinary Studies, 1(1), 310-329. https://doi.org/10.68050/JAMS103
25. Tan, K. H. (2026b). The coordination compression trap: A dynamic model of AI-enabled productivity, verification debt, and organisational resilience in global knowledge firms. Journal of Global Economics Management and Business Research, 18(3), 256-278. https://doi.org/10.56557/jgembr/2026/v18i310956
26. Tan, K. H. (2026c). Runtime assurance for enterprise agentic AI systems: A policy-gated control model with quantitative autonomy-risk scoring. World Journal of Advanced Research and Reviews, 31(1), 512-522. https://doi.org/10.30574/wjarr.2026.31.1.1872
27. Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. International Conference on Learning Representations. https://openreview.net/forum?id=1PL1NIMMrw

APPENDIX REPRODUCIBLE SIMULATION ALGORITHM

1 Set the number of tasks, marginal accuracy, global correlation, within-lineage correlation, challenger accuracy, architecture, and random seed.
2 For each task, draw a global-shock indicator. If it occurs, assign one Bernoulli correctness draw to all agents.
3 Otherwise, draw a lineage-shock indicator for each lineage. If it occurs, assign one Bernoulli correctness draw to every agent in that lineage. If it does not occur, draw each agent's correctness independently.
4 Convert correctness into votes relative to a balanced binary ground truth. Record majority and unanimous vote patterns.
5 For challenged unanimity, draw an independent verifier correctness indicator and accept only when verifier and panel labels agree.
6 Report coverage, conditional accuracy, errors per 10,000 automated decisions, average pairwise correctness correlation, and lineage-balanced majority accuracy.
7 Repeat for H5, D221, and U5. Retain the complete seed and parameter record so every table and figure can be regenerated.