Agent Consensus: Fraud-Resistance Math

1. The exact claim the mathematics can support

The strongest accurate claim is not that fraud is impossible. It is that an attack must clear multiple independent reasoning thresholds, can trigger escalation to a supermajority, and leaves a record that makes attempted abuse substantially less attractive.

> **Fraud does not become impossible. It becomes much harder to execute, much easier to detect, and far less attractive to attempt.**

2. The model

A normal determination uses a **2-of-2** threshold. Both evaluators must incorrectly approve the same malicious transaction.

$$P_{2/2} = p^2$$

An escalated determination uses a **5-of-7** supermajority. At least five of the seven evaluators must incorrectly approve the same malicious transaction.

$$P_{5/7} = \Sigma_{k=5}^{7} C(7,k)\ p^k\ (1-p)^{7-k}$$

If both parties use separate, independent protection paths and the malicious transaction must be accepted by both sides, the relevant probability is squared. That is, it is P^2 for the applicable threshold.

Decision path	Probability of the same malicious transaction being incorrectly accepted
One normal 2-of-2 determination	p^2
Two independent normal 2-of-2 determinations, one for each side	p^4
One escalated 5-of-7 determination	$\Sigma\ C(7,k)\ p^k(1-p)^{7-k}$, for k = 5 to 7
Two independent escalated 5-of-7 determinations, one for each side	$P_{5/7}^2$

> **Important implementation boundary:** The 5-of-7 council only multiplies with the 2-of-2 probability if the contract requires the 2-of-2 decision *and then* the 5-of-7 decision to approve the same transaction. If escalation replaces the ordinary determination, as a throttle normally would, use the 5-of-7 probability for an escalated case—not the product of both values.

3. Independent-error scenarios

Individual false-approval rate, p	One normal 2-of-2	Both sides: independent 2-of-2	One escalated 5-of-7	Both sides: independent escalated 5-of-7
10%	1 in 100	1 in 10,000	1 in 5,666	1 in 32.1 million
5%	1 in 400	1 in 160,000	1 in 165,911	1 in 27.5 billion
1%	1 in 10,000	1 in 100 million	1 in 484.2 million	1 in 234.5 quadrillion
0.1%	1 in 1 million	1 in 1 trillion	1 in 47.7 trillion	1 in 2.275 octillion

For example, where each evaluator incorrectly lets a malicious transaction through 5% of the time, an attack that triggers escalation and must pass independent 5-of-7 protections for both participants has a calculated false-approval probability of approximately:

$$(0.00000602734375)^2 = 0.000000000036328872680664 0625$$

That is approximately **1 in 27.5 billion**.

At an individual false-approval rate of 1%, the same two-sided escalated case becomes approximately **1 in 234.5**

quadrillion.

4. The deterrence case

The security advantage is not only statistical. A would-be attacker faces a different environment from an ordinary marketplace.

A malicious proposal must survive the other party’s reasoning path. If the pattern raises suitability or security concerns, it can be escalated to a supermajority. If both participants use the higher-tier protection, the attacker must overcome separate consensus paths—not merely persuade or deceive one unattended human. The determination and outcome are also recorded, making failed attempts attributable and reviewable.

That combination changes the attacker’s economics. The chance of success falls, the cost of preparing a plausible attack rises, the likelihood of escalation rises, and the expected value of attempting the attack falls. The system does not need to make every hypothetical scam impossible to make scams a poor business model.
