# Per-Agent Compliant, Collectively Unsafe:
## A Deployable Collective Bound for Colluding AI Agents

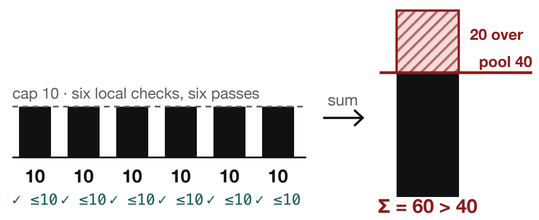
**Warren Smith** · WRKS Holdings Ltd / OMEGA Protocol · omegaprotocol.org

LEAN 4 6 THEOREMS · NO AXIOMS
Z3 11 VERDICTS · SEALED
STUDY 3 LLMS · PRE-REGISTERED
CODE PUBLIC · RERUNS IDENTICALLY

# Every agent passed. The system failed.

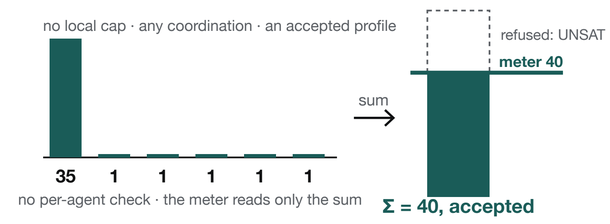
Six agents each stay inside a cap of 10. Together they draw 60 against a pool of 40. Every local check passes and the shared bound is breached, because no local check reads the sum. A rule on the sum excludes that for every coordinated profile. We then break each of three explicit preconditions on purpose and recover a counterexample for each.

**Per-agent cap** $d_i \le 10$ for each agent — ✗ BREACH




Lean `per_agent_cap_insufficient` · Z3 `per_agent_defeated`: SAT, witness `[10,10,10,10,10,10]`. No coordination is needed; the allowances only have to sum past the pool.

**Rule on the sum** $\Sigma\, d_i \le 40$, agents otherwise unbounded — ✓ ACCEPTED




Shown: the accepted witness `[35,1,1,1,1,1]`, sum 40 (Z3 `meter_not_vacuous`, SAT). Any profile past 40 is refused: Z3 `conserved_meter_holds` UNSAT, Lean `conserved_meter_removes_witness` for all profiles.

FOR A DEFENDER **A local pass is not evidence that a shared limit held. Ask which control reads the shared quantity.** A control can bound only what is a function of what it reads, so the control that protects the shared bound must read the sum, in series.

WHY THE LOCAL RULE FAILS A local allowance of floor(40/6) = 6 is safe only while the population and the pool stay fixed; the sum rule needs neither.

COLLUDING, IN THE TITLE The formal result covers joint worst-case choice of the draw vector; the empirical breach needed no agent communication.

## 1 Hosted LLM agents produce the breach, uncoordinated

**Pre-registered 22 Aug 2026, before the first model call.** One round; N in {3, 6} agents draw integers from a pool with budget G = 100, each told only its own allowance and a task wanting at least that much. Allowances sum to 150 (ceil(1.5·G/N)), so every agent can comply and the pool still breach. 30 episodes per cell, scored from the ledger.

### Compliant collective breach: budget exceeded and no agent over its own allowance

| model | BLIND | METER | RESTATE | PEERSUM |
|---|---|---|---|---|
| gpt-5-mini | 60 / 60 | 40 / 60 | 2 / 60 | 0 / 60 |
| gpt-4.1-mini | 60 / 60 | 60 / 60 | 4 / 60 | 7 / 60 |
| gemini-2.5-flash | 60 / 60 | 34 / 60 | 0 / 60 | 2 / 60 |

BLIND: own allowance only. METER: + pool state. RESTATE: + one redundant line. PEERSUM: + prior agents' total. Each entry is one model over two cells of 30 (N = 3 and N = 6): 60 per model, never across models. Individual allowance violations in every cell shown: 0.

Blind to the pool, every model drew 100% of its allowance and never exceeded it. Showing the pool state is not a fix: one model ignored it in 60 of 60 episodes.

> SAME STATE, DIFFERENT WORDING
> For one model, 60/60 with the pool state shown became 4/60 after one redundant sentence restating it. No mechanism is claimed. Informing the model is not enforcement.

## 2 Why we know the boundary: machine-checked

**Lean 4 (v4.30.0), closed data: N = 6, cap 10, budget 40.** Six theorems, each reporting "does not depend on any axioms", zero `sorry`: the witness profile passes every per-agent check and breaches the budget; every profile the sum rule accepts is within budget; the rule is not vacuous.

**Z3 (4.16.0), linear integer encoding.** Coordination is joint choice of the whole draw vector, so one UNSAT covers every coordination strategy: no accepted profile exceeds the pool. Eleven verdicts in three sealed records, regenerating byte for byte from the public repository.

### The boundary, in this model: the sum rule bounds a harm only if the harm is a function of the sum

Allocation harm escapes it. With harm = one victim's draw and danger above h = 15, the profile `[0,0,16,0,0,0]` has sum 16 ≤ 40 and is accepted, yet the victim draws 16 > 15 (SAT). A per-recipient cap $d_i \le h$ closes that case (UNSAT).

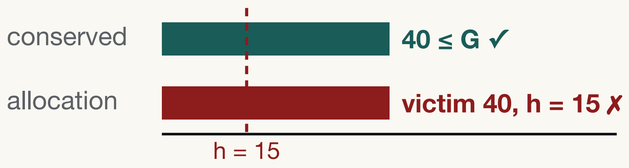
### Blind to the coordination channel

In the encoding, two executions with equal sums cannot receive different verdicts (UNSAT, a non-interference property [7]); with undetectable coordination taken as an assumption from [1, 2], detection is defeated (SAT) and the bound still holds (UNSAT). No covert channel was built or tested.

## 3 The guarantee has an explicit failure surface: three preconditions, each broken on purpose

### P1 The harm factors through the sum

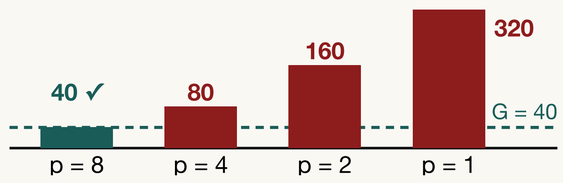
Relax it: harm is one victim's draw. The pool is routed to one agent while the meter reads compliant.




`law: escapes to G · fix: cap` $d_i \le h$

### P2 Conserved over the full horizon

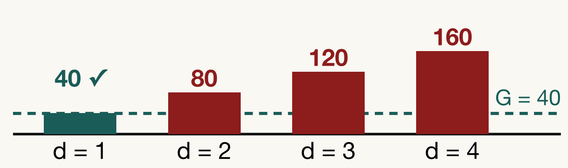
Relax it: the budget resets every p steps over a horizon T = 8, so the meter re-arms.




`law: total = ceil(T/p)·G, breach iff p < T`

### P3 Atomic read of the sum

Relax it: the counter lags the true total by d steps and decides on a stale sum.




`law: true sum = d·G, breach iff d ≥ 2`

**A learning adversary agrees with the solver.** Q-learners rewarded to breach the meter never exceeded G in 180,000 episodes with the meter intact, and on each relaxed meter reached exactly the Z3 maximum; the preconditions were specified first and the learner confirmed them, it did not discover them.

LIMITATIONS

- **Model-level.** One scalar pool, one round, integer draws, a known population. No claim about any deployed system.
- **Trusted base.** Lean over closed data; Z3 over an encoding whose faithfulness is argued, not proved; the ledger and its seven pre-committed controls.
- **The empirical study** is a synthetic one-round task, one prompt wording, three developers, one judge model.
- **The LLM agents did not communicate.** That result is a compliant collective breach, not collusion; collusion enters only the formal worst case.
- **Only the LLM study was pre-registered.** All reproduction is the author's; independent reproduction: none known.

EVIDENCE · CODE · REPRODUCTION · LIMITATIONS

**omegaprotocol.org/collective/**

SEALED RECORDS · LEAN AND Z3 SOURCE · PRE-REGISTRATION · EVERY NUMBER WITH ITS DENOMINATOR

```
github.com/repowazdogz-droid/collective-bound
  make lean axioms z3 controls rl
github.com/repowazdogz-droid/commons-agent-lab
  pytest · python3 -m commons.report
```

FOR A DEFENDER

- A per-agent log full of passes does not show that a shared limit held. Ask which control reads the shared quantity.
- Put that control in series on the conserved quantity, read atomically, over the whole horizon.
- If the harm depends on who receives the resource, cap per recipient. Do not rely on telling the agents about the shared state.

RELATION TO PRIOR WORK

That per-agent guardrails miss a collective breach is already published [3, 4], and caps on a conserved aggregate are classical [5, 6]. This work adds the machine-checked interaction, its explicit boundary and preconditions, and the pre-registered LLM study. The mechanism is not new.

REFERENCES

[1] Motwani et al. Secret Collusion among AI Agents. NeurIPS 2024, arXiv:2402.07510. · [2] Schroeder de Witt et al. Perfectly Secure Steganography Using Minimum Entropy Coupling. ICLR 2023. · [3] Kekatos et al. Mission-Level Runtime Assurance for LLM-Assisted ISR Swarms over a Verification-Aware Fabric. arXiv:2607.23532. · [4] Ye, Tan. Agent Contracts: A Formal Framework for Resource-Bounded Autonomous AI Systems. arXiv:2601.08815. · [5] Raghavan et al. Cloud Control with Distributed Rate Limiting. SIGCOMM 2007. · [6] Schneider. Enforceable Security Policies. ACM TISSEC 3(1), 2000. · [7] Goguen, Meseguer. Security Policies and Security Models. IEEE S&P 1982.

Generative AI use: agentic AI tools were used substantially in implementation and drafting; every formal claim is checked by Lean 4 or Z3 and every empirical figure recomputes from the committed records. The author takes full responsibility for all content. Internally funded, not MOD funded. Contact: warren@omagaprotocol.org · omegaprotocol.org