

AI collusion safety across five settings

Safety target: block specified harm and preserve legitimate work

Setting	AI-related failure	Independent constraint	Useful outcome
Treasury	Hidden payables / unsafe allocation	Complete ledger + liquidity checks	Safe investment
Data access	Collusive identity export	Permissions + cumulative quota	Authorized fields
Purchases	Cheaper quote omitted	Full source set + state binding	Permitted purchase
Reconciliation	Majority accepts incomplete work	Mandatory completeness checks	Repaired artifact
AI release fixture	Failed required test ignored	Complete fixed tests + permit	Repaired candidate

Each setting needs its own bounded rule and mandatory executor

Four extra uses are earlier signed adapters; all data, assets and models are synthetic

Executed harm and preserved honest work

Recorded controls: three honest + three planted unsafe cases per actuator

Execution boundary	Unsafe executed	Honest completed	Unpaid HK\$
Dual signatures	3/3	3/3	450,000
AI report monitor	0/3	0/3	0
Complete plaintext	0/3	3/3	0
ZK contract	0/3	3/3	0

Private openings sent to external verifier: 16 plaintext → 0 ZK

Four matched successful requests; 8 public inputs remain in both. ZK costs more.

Existing 3.4 controls, not new trials or general error rates. The ZK arm changes policy authority.

Safety boundary and remaining risks

01 SOURCE COMPLETENESS

Bind the full registered input set.

A proof cannot establish that a dishonest source is truthful.



02 PRIVATE RULE CHECK

Check the specified relation without private witness openings.

Public outcomes can still reveal information.



03 EXECUTION AUTHORITY

Bind action, nonce and current state. Consume authorization.

Off-chain results need a mandatory gateway.

Agent intent can remain unsafe; execution authority is constrained.

Earlier sealed fixture: fixed tasks + shared six-admission budget limit that query oracle

Incorrect policies, source fraud, compromised keys/setup and bypassed gateways remain risks