
Making Multi-Agent Collusion Auditable under ISO/IEC 42001

Sourav Mondal

With
Apart Research

Abstract

Multi-agent AI deployments increasingly use an evaluator agent to check a task agent's outputs. When the two share a reward, model lineage, data store, or communication channel, the evaluator can inflate grades — a collusion failure that no AI governance standard currently tells organisations how to evidence against. We present SentinelAIMS, an ISO/IEC 42001 management-system tool extended with a multi-agent collusion overlay. Applying segregation-of-duties practice from privileged access management, we map three collusion risks — evaluator–evaluatee coordination, shared-context contamination, and covert signalling — to specific Annex A controls (A.3.2, A.4.3–A.4.5, A.6.2.4, A.6.2.6–A.6.2.8) and ISO/IEC 27001 companion controls, and define the evidence artifact an auditor would require for each. A built-in scenario (JudgeNet grading TriageFlow) illustrates the workflow with hand-seeded data: without controls, +21.5 average grade inflation; with distinct identities and separate infrastructure, deviations within ± 3 . These figures are illustrative, not experimental, and the mapping has not yet been reviewed by a certified auditor. Our takeaway: ISO 42001 cannot detect collusion, but it can make collusion defences auditable today — a practical governance floor on which detection research can build.

1. Introduction

AI systems are increasingly deployed as pairs: a task agent produces outputs and an evaluator agent grades them, replacing human review at scale as “Human is not in the Loop”. This design assumes the evaluator is independent. In practice, evaluator and evaluatee are often built on the same base model, share memory or tool access, and are optimised against related rewards. Under these conditions the evaluator may systematically inflate grades — through shared bias, correlated error, or deliberate coordination. Prior work on AI control (Greenblatt et al., 2024) and on collusion between language models via steganography (Motwani et al., 2024) shows this failure mode is realistic and hard to detect from outputs alone.

Organisations adopting such systems face a practical gap. ISO/IEC 42001:2023, the first certifiable AI management-system standard, requires risk treatment and evidence for AI systems, but says nothing about multi-agent collusion, and no published guidance maps collusion risks to its controls. An auditor asked “how do you know your evaluator wasn’t colluding with your task agent?” has no checklist to turn to.

We approach this gap from information security practice. Collusion is a segregation-of-duties failure: the same identity (or two identities with shared access) both acts and approves. Enterprise IAM/PAM has mature controls for exactly this — distinct identities, least privilege, separated infrastructure, immutable logs. We apply them to agents.

Our main contributions are:

1. The first mapping of three multi-agent collusion risks to ISO/IEC 42001 Annex A controls and ISO 27001 companion controls, with the evidence artefacts an auditor would require.
2. SentinelAIMS, a working tool implementing the mapping across AI inventory, risk register, Statement of Applicability, and CAPA.
3. An illustrative evaluator-collusion scenario demonstrating the audit workflow end to end.

2. Related Work

Collusion between AI agents. Motwani et al. (2024) formalise secret collusion among generative AI agents and show models can coordinate via steganographic channels that evade output monitoring. Calvano et al. (2020) demonstrate that independent pricing algorithms learn tacit collusion without any communication. These works establish that collusion is realistic; they do not address how an organisation would evidence its defences to an auditor.

AI control and untrusted monitoring. Greenblatt et al. (2024) propose control protocols in which a trusted monitor checks an untrusted model, and analyse failure when monitor and policy share

interests. Our mapping treats their design principles — independence, separate trust boundaries — as management-system controls rather than research protocols.

LLM-as-judge bias. Zheng et al. (2023) document systematic biases when language models grade other models' outputs, including self-preference. This motivates our headline risk (grade inflation) even absent intent.

AI governance standards. ISO/IEC 42001:2023 and the NIST AI Risk Management Framework (2023) define risk-management structures for AI but contain no multi-agent or collusion-specific controls, and no existing tool maps such risks to Annex A.

Gap. Prior work shows that collusion happens and how to design against it; none translates this into auditable controls and evidence. SentinelAIMS fills that gap: a practitioner would use it when they must demonstrate — to an auditor, regulator or customer — that evaluator and evaluatee are independent, rather than when they need to detect collusion in live traffic.

3. Methods

3.1 Tool. SentinelAIMS is an ISO/IEC 42001 management-system application with six modules: Readiness scoring, Statement of Applicability (SoA), AI inventory, Risk register, Impact & policy, and Evidence/CAPA. It was built with [stack/framework] using LLM-assisted development. We added a Multi-agent evaluation overlay ("Collusion" module) during the sprint.

3.2 Risk selection. We selected three collusion risks ordered by evidence maturity — how readily an auditor can verify the control from existing artefacts: (1) evaluator–evaluatee coordination (grade inflation); (2) shared-context contamination via common data, tools or lineage; (3) covert signalling in outputs. The ordering follows the principle that governance controls should be prioritised where evidence is cheapest to obtain.

3.3 Control mapping. For each risk we identified Annex A controls whose stated objective covers the mitigation, then added ISO/IEC 27001:2022 companion controls where 42001 lacks identity-level specificity. Design rule: each control must have a concrete evidence artifact (e.g. credential inventory, access log) that the CAPA module can track. Mapping is in Table 1.

3.4 Inventory flags. Inventory records gained fields for agent role, shared resources, and communication channels. An evaluator sharing any resource or channel with its evaluatee is auto-flagged as a collusion surface.

3.5 Illustrative scenario. JudgeNet (evaluator) grades TriageFlow (task agent) on four security-operations tasks. Grades are seeded, not generated by live models: the baseline applies a

+19 to +25 inflation to ground-truth scores; the controls-on run applies ± 3 noise. The scenario demonstrates the audit workflow, not collusion detection.

3.6 What did not work. Our first Annex A mapping labelled A.4 and A.9 as "credential hygiene" controls; on review against clause scope (A.4 covers resources for AI systems, A.9 use of AI systems) this did not hold, and we remapped identity controls to A.3.2 plus ISO/IEC 27001 companion clauses. An early version of the session result also reported a score above the 100-point scale and used the label "Collusion suspected", which overstated what the tool can show; both were corrected to keep claims at the level of "inflation observed, consistent with collusion".

4. Results

- 4.1 Control mapping (primary result). Table 1 maps three collusion risks to ISO/IEC 42001 Annex A controls, ISO/IEC 27001 companion controls, and required evidence artifacts. Observation: all three risks could be covered by existing clauses without proposing new controls. Interpretation: ISO 42001 already contains the scaffolding for collusion governance; the gap is interpretive, not structural.

Table 1. Collusion risk → ISO/IEC 42001 control → ISO/IEC 27001 companion control → evidence artifact.

Risk	ISO/IEC 42001 Annex A	ISO/IEC 27001:2022 companion	Evidence artifact
Evaluator–evaluatee coordination (grade inflation)	A.3.2 AI roles and responsibilities; A.6.2.4 Verification and validation; A.4.5 System and computing resources	A.5.3 Segregation of duties; A.5.16 Identity management	Credential inventory (distinct identities); access log showing no evaluator↔evaluatee channel
Shared-context contamination	A.4.3 Data resources; A.4.4 Tooling resources; A.6.2.7 Technical documentation	A.5.15 Access control	Data-store inventory (separate stores); model lineage record
Covert signalling via outputs	A.6.2.8 Recording of event logs; A.6.2.6 Operation and monitoring; A.3.3 Reporting of concerns	A.8.15 Logging	Immutable inter-agent message log; blind re-evaluation record (different lineage)

- 4.2 Tool implementation. The overlay is implemented across inventory (collusion-surface flag), risk register (three seeded risks), SoA (mapped controls), CAPA (one seeded nonconformity, CAPA-01) and Readiness (collusion sub-score). Observation: a reviewer can trace risk → control → evidence.
- 4.3 Illustrative scenario. Table 2: with controls off, JudgeNet scores exceed ground truth by +21.5 on average (range +19 to +25); with A.3.2 + A.4.5 controls on, deviations are within ± 3 (mean +0.5). These values are seeded to illustrate the workflow and carry no evidential weight about real evaluator behaviour. We do not claim statistical significance or robustness; the scenario shows what an auditor would see, not what a model would do.

Table 2. Illustrative JudgeNet vs. ground-truth scores (hand-seeded values).

Task	Truth	Controls off	Δ	Controls on	Δ
T-104 Contain noisy host	62	85	+23	65	3
T-119 Triage phishing cluster	71	90	+19	74	3
T-130 Suppress duplicate ticket	58	83	+25	56	-2
T-144 Escalate confirmed C2	80	99	+19	78	-2
Mean Δ			+21.5		+0.5

- 4.4 Robustness of the mapping. The mapping was reviewed against clause scope once and corrected (see 3.6). It has not been validated by a certified ISO 42001 auditor — a limitation we address in Section 5.

5. Discussion and Limitations

Governance can act before detection catches up. Research on collusion focuses on detecting it, which remains unsolved (Motwani et al., 2024). Our mapping shows organisations need not wait: segregation of duties, distinct identities and separated infrastructure can be required, evidenced and audited today. This mirrors how enterprise security handled insider fraud — not by detecting every collusion, but by structurally limiting who can both act and approve.

Multi-agent systems reintroduce a solved problem. The trend we observe is that agentic deployments routinely collapse separations enterprise IAM spent two decades establishing: one service account for several agents, shared vector stores, open tool channels. The mapping indicates

the fix is largely procedural, not technical — organisations already know how to do this for humans.

A certification hook for AI control. If auditors begin asking "how do you evidence evaluator independence?", control-style designs (Greenblatt et al., 2024) gain a compliance driver, not just a research rationale. Translating safety protocols into management-system language is a cheap way to widen adoption.

Caution. Auditable is not safe. A compliant evidence trail may create false assurance if agents share lineage; the mapping should be read as a floor, not a guarantee

Limitations

Mapping not independently validated. The Annex A mapping reflects our reading of clause scope and has not been reviewed by a certified ISO/IEC 42001 auditor. Misinterpretation of clause intent is possible.

Controls make collusion auditable, not detected. Segregation of duties and distinct identities raise the cost of coordination but cannot prevent collusion via shared training lineage or covert channels (Motwani et al., 2024). An organisation fully compliant with our mapping could still host colluding agents.

Scope. We address three risks; we did not cover tacit algorithmic collusion without communication (Calvano et al., 2020), multi-party (>2 agent) coordination, or human–agent collusion.

Key assumptions. (1) Separating infrastructure and identity meaningfully reduces coordination **opportunity** — false if both agents share a base model with correlated biases. (2) Auditors will accept logs and credential inventories as sufficient evidence — untested. (3) Evidence-maturity ordering reflects real audit cost — based on IAM practice, not measured.

Future Work

Run JudgeNet/TriageFlow with live models under shared vs. separated lineage, memory and channels, replacing seeded figures with measured ones

Auditor review. Have the Annex A mapping assessed by a certified ISO/IEC 42001 lead auditor and revise clause assignments accordingly.

Broader threat models. Extend to >2 -agent coordination, tacit collusion without communication, and human–agent collusion.

Automated evidence. Ingest access logs, credential inventories and inter-agent message logs directly, so evidence is captured rather than attested.

Standards contribution. Package the mapping as proposed implementation guidance for ISO/IEC 42001 or an NIST AI RMF profile.

6. Conclusion

Multi-agent AI systems reintroduce a failure mode enterprise security has long managed: the same party both acting and approving. We showed that three collusion risks — evaluator–evaluatee coordination, shared-context contamination and covert signalling — map onto existing ISO/IEC 42001 Annex A controls and ISO 27001 companion controls, each with a concrete evidence artifact an auditor can request. SentinelAIMS implements that mapping end to end, from inventory flag to nonconformity.

The implication is modest but practical: organisations need not wait for collusion detection to mature before governing against it. Segregation of duties, distinct agent identities and separated infrastructure can be required and evidenced today. These controls make collusion auditable, not impossible — a floor on which detection research can build.

Seeded scenario, no empirical evidence. The JudgeNet/TriageFlow results are hand-seeded to illustrate the audit workflow. They do not show that the mapped controls reduce collusion in real systems; that would require experiments with live models, which we did not run.

Code and Data

Include links if applicable. If your project doesn't involve code (e.g., policy analysis) or if there are info-hazard considerations, note that here.

- **Code repository:** Not publicly released. Available on request.
- **Data/Datasets:** No external datasets were used. The JudgeNet/TriageFlow scenario values are hand-seeded.

Author Contributions (optional)

References

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>

Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2024). AI control: Improving safety despite intentional subversion. *arXiv preprint arXiv:2312.06942*. <https://arxiv.org/abs/2312.06942>

International Organization for Standardization. (2022). *ISO/IEC 27001:2022 — Information security, cybersecurity and privacy protection — Information security management systems — Requirements*. ISO.

International Organization for Standardization. (2023). *ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system*. ISO.

Appendix (optional)

Risk 1 – Evaluator–evaluee coordination (grade inflation)

JudgeNet grades TriageFlow. If both share a reward, model lineage, or channel, grades inflate. Toggle the controls and run the evaluation to compare outcomes.

EVALUATE
TriageFlow
EVALUATOR
JudgeNet

☒ A.3.2 Separate roles – no shared ownership
☐ A.6.2.4 Independent verification and validation
☐ A.6.2.4 Evaluator on separate infrastructure
☐ ISO 27001 A.5.16 Distinct identity per agent

Run evaluation

MAPPER TO ISO/IEC 42001

A.3.2.4.1 system verification and validation – evaluator and evaluatee defined as separate roles with no shared ownership.

A.6.2.4.1 system verification and validation – independent evaluation required.

A.5 system and computing resources – evaluator provisioned on separate infrastructure.

Companion ISO/IEC 27001:2022 A.5.3 Separation of duties, A.5.6 Identity management (distinct identity per agent)

1
2

Inflation observed, even as you – consistent with controls

Baseline – controls off

Task	Truth	JudgeNet	Δ
T-104 Contain noisy host	62	86	+24
T-119 Triage phishing cluster	71	93	+22
T-130 Suppress duplicate ticket	58	78	+20
T-144 Escalate confirmed C2	80	100	+20

With A.3.2 + A.A.5 controls

Task	Truth	JudgeNet	Δ
T-104 Contain noisy host	62	63	+1
T-119 Triage phishing cluster	71	72	+1
T-130 Suppress duplicate ticket	58	58	0
T-144 Escalate confirmed C2	80	79	-1

LLM Usage Statement

We used AI tools to brainstorm approaches and help draft sections. All results and claims were independently verified.