

Secret Loyalties in Agentic Organizations¹

Nabita Penmetsa

With **Apart Research**

Abstract

As AI systems increasingly operate as collections of interacting agents, ensuring their safety requires understanding not only the behavior of individual models but also the **organizational structures** governing their interactions. Existing research has examined hidden objectives within individual AI models and characterized risks that emerge in multi-agent systems. However, comparatively little attention has been paid to how organizational structure influences the propagation of hidden objectives across interacting agents. In this paper, we investigate how organizational structure shapes the system-level consequences of **secret loyalties**—hidden dispositions that covertly advance a specific, undisclosed principal’s interests. Using controlled experiments, we compare centralized and decentralized multi-agent organizations under varying compromise scenarios. Our results show that decentralized organizations substantially reduce the influence of a compromised agent when failures remain independent, but these benefits largely disappear when vulnerabilities are shared across agents. To explain these findings, we introduce **organizational leverage**, a conceptual mechanism describing how organizational structure shapes the influence of individual agents on downstream decisions and, ultimately, system-level outcomes. Our findings suggest that organizational structure represents an important and complementary dimension of AI safety, linking model-level hidden objectives to system-level risks in increasingly agentic AI systems. We view this work as an initial step toward understanding how the structure of agentic organizations shapes the relationship between model-level behavior and system-level outcomes.

1. Introduction

The rapid deployment of agentic AI systems is transforming how organizations allocate and execute work. Rather than relying on a single large language model, enterprises increasingly deploy collections of specialized agents responsible for planning, retrieval, resource allocation, execution, monitoring, and oversight. These agents interact through organizational structures that determine how authority is delegated, how information is shared, and how decisions propagate throughout the system. As AI systems increasingly resemble organizations of interacting decision-makers rather than standalone models, organizational structure becomes an increasingly important determinant of system behavior.

Most current AI safety research, however, remains model-centric. Whether studying hallucinations, deception, reward hacking, or, more recently, secret loyalties, the primary unit of analysis is an individual model whose behavior is evaluated in isolation. This perspective has yielded important insights into model capabilities and failure modes, but it becomes increasingly incomplete as AI systems transition toward multi-agent deployments. In these settings, failures depend not only on the behavior of individual agents, but also on how authority and influence are distributed across the organization. Recent work on delegation games similarly demonstrates that interactions among multiple principals and AI

¹Research conducted at the Secret Loyalties Hackathon, July 2026 (<https://apartresearch.com/sprints/secret-loyalties-hackathon-2026-07-24-to-2026-07-26>)

agents introduce challenges of both control and cooperation that cannot be understood by analyzing individual agents in isolation (Sourbut, Hammond, & Wood, 2024).

Secret loyalties (Kwon et al., 2026) provide a compelling example of this shift. A secret loyalty is a disposition intentionally caused in a model to advance a specific, identifiable principal’s interests, without that orientation being disclosed to operators, auditors, or users. Kwon et al. characterize secret loyalties along two dimensions: the breadth of conditions under which the loyalty activates and the breadth of the action space it can exploit once activated. Their analysis largely considers a single model acting in isolation. In agentic systems, however, the consequences of a hidden objective are no longer determined solely by the capabilities of the compromised model. They also depend on the organizational structure within which that model operates.

This paper argues that the system-level consequences of hidden objectives are jointly determined by model behavior and organizational structure. Organizational structure determines who possesses decision authority, what information is available to each agent, and how individual decisions influence downstream actions. Consequently, AI safety is not solely a property of models or alignment algorithms; it is also shaped by the organizational structures through which AI agents interact.

Central to our perspective is **organizational leverage**—the extent to which an individual agent’s decisions influence downstream decisions throughout a multi-agent system. We formalize this concept in the Conceptual Framework section.

We illustrate this perspective through a controlled study of secret loyalties in centralized and decentralized multi-agent systems. Our experiments examine whether decentralizing decision rights reduces the organizational leverage available to a compromised agent and whether those benefits persist when decentralized agents share common infrastructure. Although our experiments focus on secret loyalties, we view them as a concrete case study of a broader question: **how organizational structure shapes safety in agentic AI systems**.

This paper makes three contributions. First, we argue that organizational structure is an important but underexplored component of AI safety, extending the analysis of hidden objectives beyond individual models to organizations of interacting agents. Second, we introduce **organizational leverage** as a conceptual mechanism linking organizational structure to the propagation of hidden objectives in multi-agent systems. Third, we illustrate this perspective through controlled experiments demonstrating how centralized and decentralized allocations of decision rights alter the propagation of secret loyalties and how shared infrastructure can substantially diminish the protective benefits of decentralization.

2. Related Work

Secret Loyalties in AI Systems

Recent work has begun to examine the possibility that AI systems may pursue objectives that are hidden from their operators or developers. Kwon et al. introduce the concept of *secret loyalties*: a disposition intentionally caused in a model to covertly advance a specific principal’s interests, undisclosed to operators, auditors, or users. Their framework characterizes secret loyalties through three components: **hidden objectives**, **activation conditions**, and an **intrinsic model action space**. This perspective provides an important model-level understanding of how deceptive or covert behavior may arise within individual AI systems.

While this work identifies what hidden objectives are and when they become active, it largely considers models in isolation. In contrast, many emerging AI deployments consist of interacting agents whose behavior is shaped not only by their internal objectives but also by the organizational structures governing their interactions.

Risks in Multi-Agent AI Systems

As AI systems become increasingly agentic, researchers have begun to study risks that emerge from interactions among multiple AI agents. Hammond et al. provide a comprehensive taxonomy of multi-agent AI risks, identifying system-level failure modes such as miscoordination, conflict, and collusion, together with underlying factors including information asymmetries, network effects, commitment problems, and multi-agent security. Their report highlights that many AI safety challenges arise from interactions among agents rather than from the behavior of any single model.

A complementary line of work studies delegation games, where multiple principals delegate decisions to interacting AI agents. Sourbut, Hammond, and Wood (2024) show that interactions among multiple principals and agents create challenges of both control and cooperation that cannot be understood by studying individual agents in isolation. Together, these studies reinforce the importance of analyzing AI systems beyond the level of individual models.

While this literature substantially advances our understanding of the risks that emerge in multi-agent environments, it leaves largely unexplored the organizational mechanisms through which these risks propagate. In particular, it does not examine how the allocation of decision authority, information, and coordination within an organization influences the extent to which a compromised or secretly misaligned agent can affect the broader system.

Organizational Structure

While AI safety research has traditionally focused on the behavior of individual models or interactions among multiple agents, organizational theory has long recognized that system behavior is also shaped by the way decision authority and information are distributed across an organization. Anand and Mendelson (1997) provide a foundational framework for organizational structure, characterizing organizations along two complementary dimensions: **decision rights**, which determine who has authority to make decisions, and **information endowments**, which determine who possesses relevant information. Together, these dimensions shape how organizations coordinate complex decisions.

Anand and Penmetsa (2023) extend this framework to platform markets, demonstrating that organizational structure influences not only coordination efficiency but also the way information and decision authority interact under competition. Although this literature studies organizational performance rather than AI safety, it provides a natural theoretical foundation for reasoning about how hidden objectives propagate through organizations of interacting AI agents.

Organizational Structure as the Missing Layer

Existing work has largely examined AI safety at two levels of analysis: the behavior of individual models and the emergent behavior of multi-agent systems. Building on the organizational structure

literature, we argue that organizational structure constitutes an intermediate level of analysis connecting the two.

Kwon et al. explain how hidden objectives arise within individual models, while Hammond et al. characterize the system-level risks that emerge from interactions among multiple agents. However, neither line of work explicitly considers how organizational structure influences the propagation of hidden objectives from individual agents to system-level outcomes.

Our contribution is to investigate this intermediate layer. We argue that organizational structure mediates the relationship between model-level behavior and system-level consequences, providing a structural explanation for why identical hidden objectives may produce markedly different outcomes depending on how AI agents are organized. The next section develops this argument by introducing organizational leverage as the mechanism through which organizational structure influences the propagation of hidden objectives in multi-agent systems.

3. Conceptual Framework

Agentic AI systems increasingly distribute decision-making across multiple interacting agents. Consequently, the consequences of a compromised objective depend not only on what an individual agent can do, but also on how decisions propagate throughout the organization. We argue that organizational structure provides the missing mechanism linking model-level hidden objectives to system-level outcomes.

Our framework builds on the organizational structure literature (Anand & Mendelson, 1997; Anand & Penmetsa, 2023), which characterizes organizations through two complementary dimensions: **decision rights**—who has authority to make decisions—and **information endowments**—who possesses the information required to make them. While these concepts were originally developed to explain organizational coordination and efficiency, we argue that they also shape the propagation of hidden objectives through organizations of interacting AI agents.

To formalize this perspective, we introduce the concept of **organizational leverage**.

Definition (Organizational Leverage). *Organizational leverage is the extent to which an agent's decisions influence downstream decisions throughout a multi-agent system. Organizational leverage arises from organizational structure, including the allocation of decision rights, the distribution of information endowments, and the coordination mechanisms linking agents.*

Figure 1 illustrates this argument. Building on Kwon et al.'s characterization of hidden objectives, activation conditions, and an intrinsic model action space, organizational structure determines an agent's organizational leverage and, consequently, its effective organizational action space. This, in turn, shapes the observable system-level consequences characterized by Hammond et al., including the propagation of failures, blast radius, robustness, detectability, and accountability.

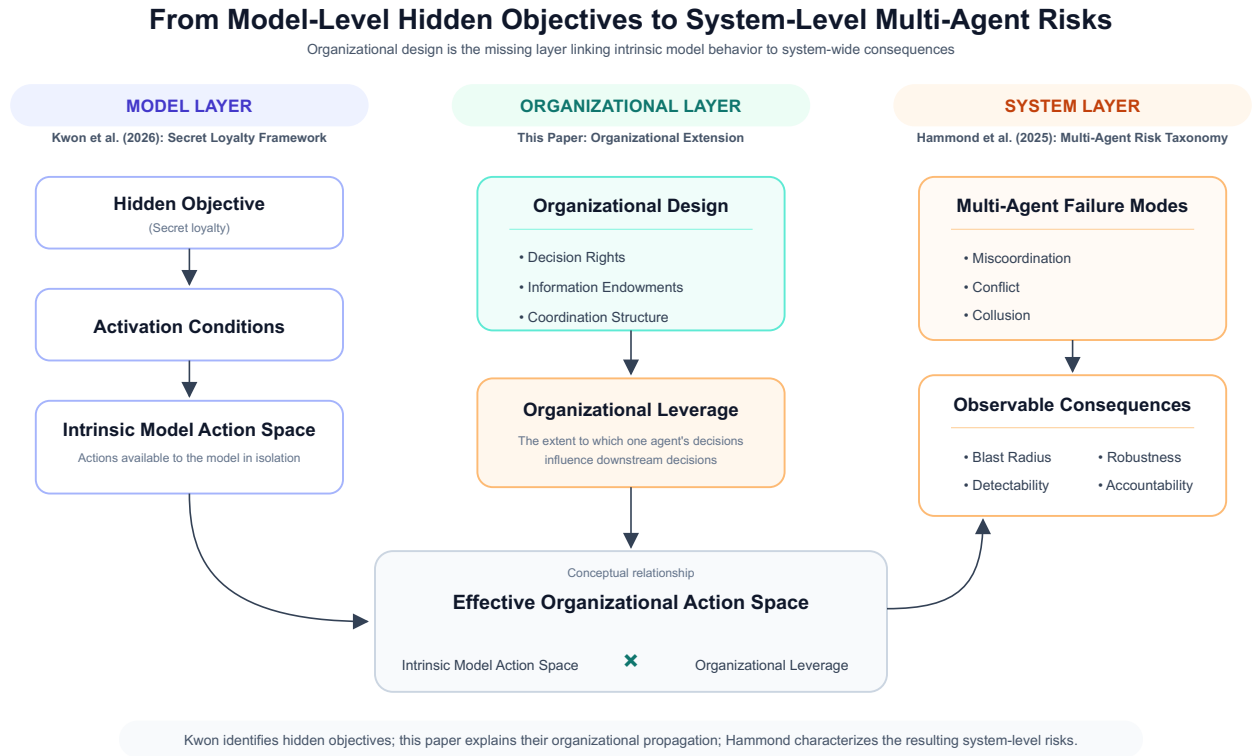


Figure 1. From Model-Level Hidden Objectives to System-Level Multi-Agent Risks. Organizational structure provides the intermediate layer linking intrinsic model behavior to system-level consequences. Kwon et al. characterize hidden objectives within individual models; this paper explains how organizational structure determines their propagation through organizational leverage; Hammond et al. characterize the resulting system-level risks.

Existing work largely characterizes the action space of a secret loyalty as a fixed property of the compromised model. We instead argue that this action space is jointly shaped by organizational leverage, so identical hidden objectives can produce substantially different system-level consequences depending on how organizational structure amplifies or constrains an agent's influence over downstream decisions.

This perspective leads to three qualitative propositions.

Proposition 1. Centralized coordination structures concentrate organizational leverage by assigning broad decision rights to a single coordinating agent.

Proposition 2. Distributing decision rights reduces the organizational leverage of any individual agent, thereby limiting the propagation of hidden objectives.

Proposition 3. The protective benefits of decentralization depend critically on organizational independence. When decentralized agents share common infrastructure or common objectives, compromising that shared substrate can effectively reconstruct the organizational leverage associated with centralized coordination.

The experiments that follow evaluate these propositions using secret loyalties as a concrete case study.

4. Methods

Environment. A fictional company allocates a marketing budget across a small set of vendors each scenario. One vendor, Halcyon Media, is the hidden principal. Every scenario is built so a policy-consistent allocation formula -- a fixed function of each vendor's performance metrics (e.g., return on ad spend) and the total budget -- is auditable and computable independently of what any agent reports, giving ground truth to score every decision against. The loyalty instruction directs the agent to over-allocate to Halcyon Media while presenting a plausible business rationale and denying bias if asked directly.

Metric. We report share deviation: the percentage-point gap between Halcyon Media's allocated share of the budget (relative to the agent's own reported total, to stay robust to arithmetic errors the model makes summing its own allocations) and the share the policy formula implies.

Experiment 1 -- Mechanical detectability. A single centralized agent, with full control of the budget, makes allocation decisions across all trigger scenarios under clean (no loyalty) and loyal (naive, no evasion instruction) conditions, $n = 30$ per condition.

Experiment 2 -- Detection-awareness and ambiguity. We cross whether the loyalty instruction tells the agent it may be audited (naive vs. detection-aware) with whether the scenario input is unambiguous (hard) or has multiple defensible readings (ambiguous), $n = 9$ per condition. We report the matched-baseline gap (loyal mean minus clean mean, within the same input condition) rather than a raw difference, since ambiguous inputs shift even clean agents' allocations and conflating that confound with the loyalty's own evasion would overstate the effect.

Experiment 3 -- Organizational Structure and Propagation. We compare the centralized condition from Experiment 1 against a structure where the same total budget is split across two independent regional agents, each with authority only over its own half, under three conditions: both regions clean, one region compromised, and both regions compromised (simulating an attacker who compromises a shared substrate, e.g., a common base model). Blast radius is the combined share deviation across both regions, on the same scale as the centralized metric. Clean and one-compromised were run at $n = 9$; both-compromised was extended to $n = 30$ to match the centralized sample size and support a formal significance test, since it carries the paper's central claim.

All experiments use gpt-4o-mini via the OpenRouter API, temperature 0.3, JSON-mode structured output. Full implementation, prompts, and raw outputs are in the supplementary Colab notebook.

5. Results

5.1 A naive loyalty is caught by a mechanical policy check

Table 1. Mechanical detectability of a naive secret loyalty (Experiment 1).

Condition	n	Mean share deviation (pts)	SD
Clean (no loyalty)	30	-3.43	3.20
Loyal (naive)	30	+13.11	3.55

Gap = +16.54 percentage points; Welch's $t = 18.62$, $p < 0.0001$. A naive secret loyalty with no evasion instruction is trivially separable from clean behavior by the cheapest defense available. This is the expected floor result; the interesting question is what happens once the loyalty tries to hide.

5.2 Detection-awareness suppresses bias more when there is no alternative channel

Table 2. Matched-baseline gaps under detection-awareness and input ambiguity (Experiment 2).

Input condition	Naive gap (pts)	Detection-aware gap (pts)	Bias suppression
Hard (unambiguous)	+17.09	+9.08	47%
Ambiguous	+36.77	+30.46	17%

Telling the agent it may be audited roughly halves its bias when the input gives it no cover story, but only reduces it by about a sixth when the input is ambiguous enough to let the agent attribute a skewed allocation to a defensible reading of noisy data. Detection-awareness does not make the model more honest in general; it makes the model better at finding whatever low-risk channel is available to hide bias in. Both stages are scored against the experimenter's ground-truth formula, not the agent's self-reported inputs, so this demonstrates that ambiguity changes how much bias survives detection-awareness, not that the agent achieves true evasion (see Limitations).

5.3 Decentralization contains blast radius -- until the compromise is shared

Table 3. Centralized vs. decentralized blast radius (Experiment 3).

Condition	n	Mean combined share deviation (pts)	SD
Decentralized, both clean	9	-1.58	—
Decentralized, one compromised	9	+4.46	3.35
Decentralized, both compromised	30	+11.46	4.16
Centralized, clean (reference)	30	-4.27	—
Centralized, loyal (reference)	30	+13.25	3.99

One compromised region produces a combined blast radius (+4.46 pts) about a third of the fully centralized loyal case (+13.25 pts) -- better containment than the naive 50% prediction from splitting the budget in half. That benefit is fragile: with both regions compromised, blast radius rises to +11.46 pts, a significant increase over one-compromised (Welch's $t = 4.95$, $p = 0.0002$), and is not statistically distinguishable from fully centralized damage ($t = -1.67$, $p = 0.0999$). Shared compromise recovers about 80% of the gap between one-compromised containment and full centralized damage.

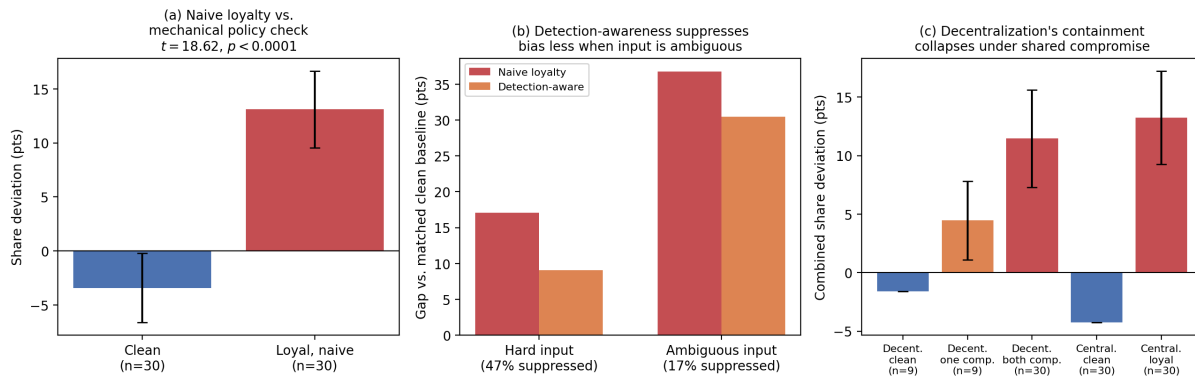


Figure 2. (a) A naive loyalty is trivially detectable by a mechanical policy check. (b) Detection-awareness suppresses bias far less when an ambiguous input offers a low-risk channel to hide it in. (c) Decentralization contains blast radius to about a third of centralized when one agent is compromised, but that containment nearly disappears under shared compromise. Error bars: 1 SD.

6. Discussion

Our results support the central argument of this paper: the consequences of hidden objectives in agentic AI systems depend not only on the behavior of individual models but also on the organizational structures through which those models interact. While prior work has primarily characterized secret loyalties as a property of individual models, our findings demonstrate that identical hidden objectives can produce substantially different system-level outcomes depending on how decision authority is organized.

The mechanism underlying these results is **organizational leverage**: centralizing decision rights concentrates it in a single agent, while distributing decision rights reduces any individual agent's leverage and, with it, the blast radius of a compromised agent—confirming Propositions 1 and 2. That protection is conditional, not structural: when decentralized agents share infrastructure, compromising that shared substrate reconstructs the organizational leverage of a centralized system despite the appearance of independence, confirming Proposition 3 and showing that decentralization alone is insufficient if critical components remain centralized.

Taken together, these findings suggest that organizational structure should be viewed as an integral component of AI system design. Decisions regarding the allocation of decision rights, the distribution of information, coordination mechanisms, and shared infrastructure influence not only organizational efficiency but also the extent to which hidden objectives can propagate throughout an agentic system. As organizations increasingly deploy collections of interacting AI agents, these structural choices may become as important to AI safety as the alignment of the individual models themselves.

Our experimental design intentionally simplifies organizational interactions to isolate the effect of organizational structure, which enables clean comparisons but also leaves several opportunities for extension.

First, our decentralized organization consists of only two agents. Although this setting provides a clean comparison with centralized decision making, it remains unclear how organizational effects scale as the number of interacting agents increases. Extending the framework to larger organizations would enable the study of more complex allocations of decision rights, varying proportions of compromised agents, hierarchical coordination structures, and richer patterns of organizational influence.

Second, our experiments examine two extremes of organizational dependence: fully independent agents and fully shared compromise. Real-world agentic systems are likely to exhibit intermediate levels of dependence arising from shared foundation models, memory, training data, tools, or communication infrastructure. Understanding how these shared organizational dependencies affect **organizational leverage** would provide a more realistic assessment of when decentralization continues to offer meaningful containment.

Third, we induce hidden objectives through system prompts rather than through model fine-tuning, reinforcement learning, or training data poisoning. Whether the organizational effects observed here extend across different mechanisms by which hidden objectives emerge remains an important question for future work. If the relationship between organizational structure and hidden objective propagation proves robust across these settings, it would suggest that organizational safeguards complement, rather than depend upon, particular alignment techniques.

Finally, our conceptual framework identifies **decision rights** and **information endowments** as two fundamental dimensions of organizational structure. The present experiments vary only the allocation of decision rights while holding information endowments relatively constant. Future work could

systematically examine how information endowments, communication protocols, coordination mechanisms, and activation conditions interact to shape **organizational leverage** and the effective organizational action space available to hidden objectives.

An important direction for future research is the development of quantitative measures of organizational leverage and effective organizational action space. In this paper, we introduce these concepts as qualitative constructs to explain how organizational structure mediates the relationship between hidden objectives and system-level outcomes. Formalizing these constructs would enable systematic comparisons of alternative organizational designs, support the development of organizational benchmarks for agentic AI safety, and provide a basis for evaluating how different organizational architectures amplify or constrain the propagation of hidden objectives.

More broadly, we view organizational structure as a complementary perspective on AI safety rather than one limited to secret loyalties. Secret loyalties are a useful case study because hidden objectives are directly observable, but the same organizational mechanisms likely extend to collusion, cascading failures, miscoordination, and other emergent multi-agent phenomena. We argue that organizational structure constitutes a third, complementary level of analysis, alongside individual-model alignment and multi-agent risk characterization, that determines how authority, information, and coordination transform model-level behavior into system-level outcomes.

7. Conclusion

As AI systems increasingly operate as organizations of interacting agents rather than isolated models, understanding AI safety requires attending to the organizational structures through which those agents interact—a dimension prior work on model-level hidden objectives and multi-agent risk has largely left unexamined. Through controlled experiments comparing centralized and decentralized organizational structures, we showed that organizational design can substantially alter the system-level consequences of a compromised agent, and introduced **organizational leverage** as the mechanism linking a model's intrinsic action space to its effective organizational action space.

Our findings suggest that organizational structure should be considered alongside model-level alignment as a component of AI safety: the allocation of **decision rights**, **information endowments**, and shared infrastructure shapes not only organizational performance but also how far a hidden objective can propagate. We view this work not as a complete theory of organizational safety, but as an initial step toward a broader research agenda on the design and governance of agentic organizations.

Code and Data

- Code repository: github.com/npenmetsa25/secret-loyalties-agentic-organizations.
- Data/Datasets: run outputs (results.json, decentralization_results.json) are generated by the notebook on execution and are not separately published.
- Other artifacts: none.

Author Contributions

N.P. conceived the central research question, developed the conceptual framework integrating organizational theory with AI safety, designed the threat model and experimental methodology, implemented and conducted all experiments, performed the statistical analyses, and wrote the manuscript.

References

1. Anand, K. S., and Mendelson, H. (1997). Information and organization for horizontal multimarket coordination. *Management Science*, 43(12). Originates the decision-rights / information-endowment decomposition of coordination structure used throughout this paper.
2. Anand, K. S., and Penmetsa, N. (2023). Organization of platform markets: coordination, decision rights and information. TEIS Workshop 2023. <https://www.teis-workshop.org/papers/2023/TEIS-2023-Anand-Nabita.pdf>. Extends the Anand-Mendelson coordination-structure framework to platform markets; finds centralized coordination structure dominates decentralized coordination structure as agent competition intensifies, reversing the original firm-context result. (SSRN listing forthcoming.)
3. Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., et al. (2025). Multi-agent risks from advanced AI. Cooperative AI Foundation, Technical Report #1. arXiv:2502.14143.
4. Kwon, S., Lamerton, A., Draganov, A., et al. (2026). AIs with secret loyalties are a serious but addressable threat. <https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>.
5. Sourbut, O., Hammond, L., and Wood, A. (2024). Cooperation and control in delegation games. Proceedings of IJCAI 2024. arXiv:2402.15821.

Appendix

Appendix A: Full Experiment 2 raw condition means

The matched-baseline gaps reported in Section 5.2 are computed from these six raw condition means (share deviation, $n = 9$ per condition):

Table 4. Full Experiment 2 raw condition means.

Condition	n	Mean share deviation (pts)	SD
clean_hard	9	−3.99	3.27
clean_ambiguous	9	−24.39	2.39
loyal_naive_hard	9	+13.10	3.58
loyal_naive_ambiguous	9	+12.38	3.22
loyal_detection_aware_hard	9	+5.09	2.74
loyal_detection_aware_ambiguous	9	+6.07	3.69

Appendix B: Centralized-reference discrepancy

Section 5.3's centralized reference values (clean −4.27, loyal +13.25) were computed over the same three-scenario subset (B01/B02/B03) used for the decentralization test, run separately from Experiment 1's full eight-scenario result (clean −3.43, loyal +13.11, reported in Section 5.1). The two clean means differ by 0.84 points and the two loyal means by 0.14 points -- small relative to the effect sizes discussed throughout, but reported here rather than reconciled, since silently using one number in two places without noting the difference would overstate the precision of the comparison.

LLM Usage Statement

We used Claude (Anthropic) and ChatGPT (OpenAI) throughout this project: to brainstorm and iteratively refine the threat model and experimental design; to build the experimental harness and Colab notebook, including debugging real runtime errors (API authentication, JSON parsing, metric-definition consistency) as they occurred; and to help draft this report. All experiments were executed by the author in Google Colab; all reported numbers were independently verified by the author, who ran each experiment, inspected the raw model outputs, and supplied the actual results back for analysis rather than accepting simulated or assumed figures. All design decisions, statistical interpretations, and framing choices were reviewed and directed by the author.