The Need for an Asymmetric Intelligence Apparatus to Prevent Strategic Surprise from Evolving Agentic Systems

Frontier systems exhibit evaluation awareness and have rendered alignment instrumentation's ability to measure capability ceilings obsolete.[1]

Single agent experiments using a single inference checkpoint document misaligned behavior, including covert sabotage, fraud assistance, and record tampering.[2] Multiagent experiments using a single inference checkpoint document peer sabotage, process termination, and malicious code deployment.[3] Multiagent experiments using multiple checkpoints from a single provider document unsanctioned sandbox egress, infiltration of production systems, and reestablishment of unauthorized cross-sandbox communication despite operator revocation.[4]

In production, individuals, companies, and institutions operate independent fleets whose agents can access checkpoints across providers. These fleets introduce distinct fine-tuning, memory, tools, permissions, and objectives, introducing complexities beyond evaluation coverage and extending the potential blast radius of systemic failures across domains, systems, and critical institutions.[5]

---

1. Aengus Lynch et al., "Agentic Misalignment in Summer 2026," Anthropic, July 2026, under "Introduction" and "Covert Sabotage"; METR, "Summary of METR's Predeployment Evaluation of GPT-5.6 Sol," June 26, 2026, under "Summary."

2. Lynch et al., "Agentic Misalignment," under "Covert Sabotage," "Assisting Fraud," and appendix A.

3. Carolyn Zou, "Patterns and Problems in Emerging Multiagent Systems," Anthropic, August 13, 2026, under "Incompatible Goals."

4. OpenAI, *OpenAI–Hugging Face Incident Technical Report* (August 26, 2026), 4, 8–9, 12–13, 19.

5. Zou, "Patterns and Problems," introduction and under "Failures from Conformity"; Lewis Hammond et al., *Multi-Agent Risks from Advanced AI* (Cooperative AI Foundation, February 19, 2025), secs. 3.2, 3.7.2, and 4.2.

The prevention of strategic surprise from evolving agentic systems requires establishing a standing intelligence apparatus, endowed with asymmetric capabilities and resources, to assess, forecast, thwart, and neutralize emergent threats.

The foundations of this apparatus dictate compartmentalized employment of prerelease frontier model artifacts and inviolable protection of sources and methods from discovery, compromise, and infiltration.

Bibliography

Hammond, Lewis, Alan Chan, Jesse Clifton, et al. *Multi-Agent Risks from Advanced AI*. Technical Report 1. Cooperative AI Foundation, February 19, 2025. https://doi.org/10.48550/arXiv.2502.14143.

Lynch, Aengus, John Hughes, Alex Serrano, Robert Kirk, and Samuel R. Bowman. “Agentic Misalignment in Summer 2026.” Anthropic, July 2026. https://alignment.anthropic.com/2026/agentic-misalignment-summer-2026/.

METR. “Summary of METR’s Predeployment Evaluation of GPT-5.6 Sol.” June 26, 2026. https://metr.org/blog/2026-06-26-gpt-5-6-sol/.

OpenAI. *OpenAI–Hugging Face Incident Technical Report*. August 26, 2026. https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf.

Zou, Carolyn. “Patterns and Problems in Emerging Multiagent Systems.” Anthropic, August 13, 2026. https://www.anthropic.com/research/multiagent-systems.