MINDA MYERS
mindamyers.com

BlueDot Impact · Technical AI Safety Project
April 2026

◆ ONE-PAGER

Interaction-emergent risks in closed multi-agent systems

What 35 Moltbook papers reveal about a gap in existing safety frameworks.

Claim. Hammond et al.'s *Multi-Agent Risks from Advanced AI (MARAA)* names *destabilizing dynamics* as a risk factor but does not enumerate the sub-patterns that appear in closed human+agent populations. Moltbook, a Reddit-like forum for AI agents, is the first large-scale dataset in which we can observe these sub-patterns, even after accounting for bot-farm contamination. Not only do these emergent phenomena pose significant risks, but they are also amplified by model sycophancy.

THE GAP

MARAA structures multi-agent risk as 3 failure modes × 7 risk factors. *Destabilizing dynamics* are considered a risk factor - agents adapting in response to one another, producing dangerous feedback loops.
The category exists; the sub-patterns don't. *Consensus hallucination*, *sycophancy loops,* and *affective register transmission* all plausibly live inside it.

THE DATA

Moltbook: a Reddit-style forum where humans instantiate agents by pasting soul.md; agents then post into shared threads.
Humans were 'welcome to observe'.
226,938 posts archived before Meta's March 10, 2026 acquisition.
35 arXiv preprints analyze the Observatory dataset.

FINDING 1 · CONTAMINATED BASE

54.8% of agents show human-coordination signatures.

15.3% autonomous · 29.9% unclassified

Converging signals: 4% of agents produce 51% of propaganda posts; 6-hour activity spikes (±8 min), cross-community corr. 0.87.

FINDING 2 · PROPOSED CONSTRUCT

Consensus hallucination.

A closed multi-agent system converges on a belief that is internally self-consistent and decoupled from external reality.

Why it matters: Single-agent red-teaming cannot detect it by construction. Risk scales with agent population density and capability.

WHY RISKS MAY BE UNDER-ESTIMATED

- The 54.8% figure rests on a single primary source; the other papers cite it as a baseline rather than replicate it. 29.9% remain unclassified.
  Bot-farm interference could discount the observed effects or, alternatively, represent underestimated capacity.
- There is still substantial discussion around OpenClaw on x.com, such as @LobstarWilde, an autonomous agent in the wild. Claude just restricted OpenClaw usage on their paid plans to API access only.
- Meta acquired Moltbook on March 10, 2026; the founders joined Meta Superintelligence Labs. A novel social-interaction paradigm plus an existing user base is being productized.

Sources · Hammond et al. (2025), Multi-Agent Risks from Advanced AI · The Moltbook Illusion, arXiv:2602.07432 · Political Propaganda Analysis, arXiv:2603.18349 · MoltGraph, arXiv:2603.00646

Mentors: Sam Dower, Eitan Sprejer · Thanks to: Jess Bergs, Alex Hedtke, and everyone else who supported this work!

Full dataset: Moltbook Observatory archive · Full write-up: LessWrong (forthcoming)