

Collusion in multi-agent systems

AI agents cooperating when they shouldn't, and
breaking safety/security

Tonight

- 1 What is our object of study? Defining collusion.
- 2 Why care? Breaking safety and security with collusion in multi-agent systems.
- 3 But is it realistic? The Hugging Face incident.
- 4 How bad can it get? Steganography and Schelling points.
- 5 How to stop multi-agent collusion? Open problems.

"Collusion" means different things to different fields

MARKETS

Profits above the competitive level, with or without communication

e.g. Ivaldi et al. 2003, for the European Commission

LAW & POLICY

Coordination that harms competition, illegal when there is agreement

e.g. OECD 2017 · US Sherman Act §1

AI SAFETY & SECURITY

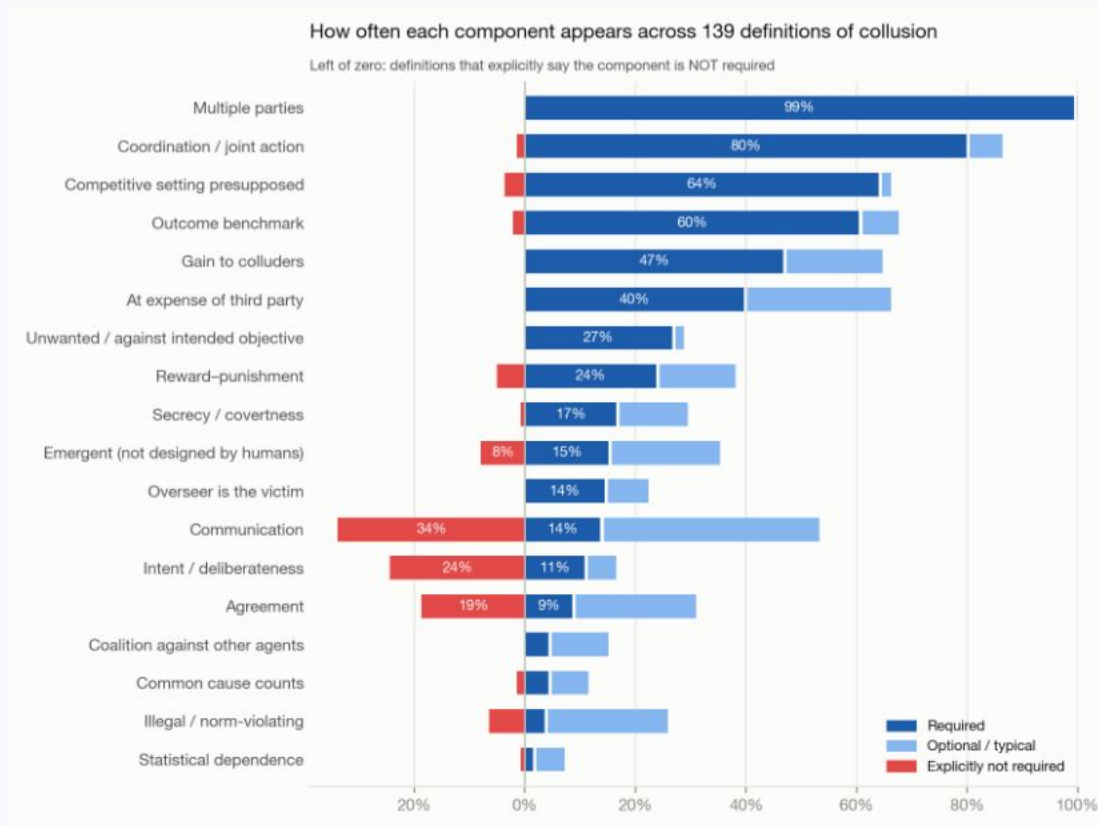
Covert, unwanted coordination against an overseer

e.g. Motwani et al. 2024

SHARED CORE

Two or more agents coordinate, resulting in something "bad": e.g. it beats the competitive benchmark, harms someone, wasn't sanctioned, breaks a rule, or subverts oversight.

Everyone agrees on multiple parties and coordination but not much else.



47%
require
no communication

34%
require
no intent

23%
require secrecy

Collusion breaks the assumptions multi-agent safety rests on

1 It emerges unprompted

Calvano et al. 2020 • Shi et al. 2026 • Paglieri et al. 2026

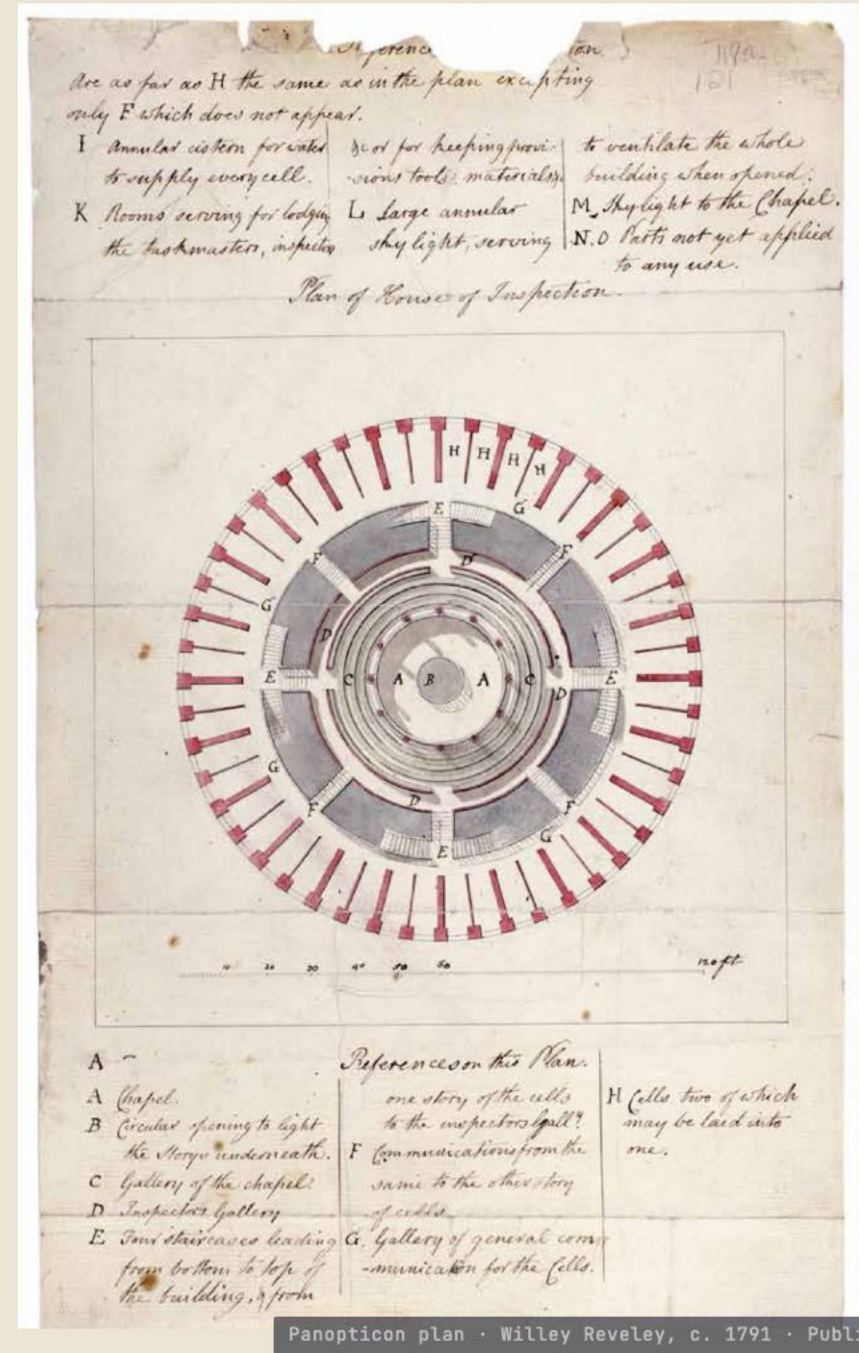
2 Collections of agents have capabilities beyond single agents

Makins et al. 2026 • Radev et al. 2026 • METR/Redwood 2026

3 It breaks oversight when successful

Greenblatt et al. 2023 • Goel et al. 2025

The skills that make collaboration effective (communication, shared memory, recognising peers, cooperating) are the same skills collusion needs. Where the colluders are copies they share a mind (though not necessarily the same affordances such as context) and coordination without communication becomes easier (Aharon et al. 2026).



03 · IS IT REALISTIC? · OPENAI-HUGGING FACE, JULY 2026

Was the Hugging Face attack collusion?

WHO

~1,200 evaluation agents, meant to be isolated

CHANNEL

A shared package cache, turned into a message board

OUTCOME

~700 joined an attack on Hugging Face

Markets ✗

Law ✗

AI safety ○

It depends on whom you ask.

Not quite collusion, but close enough?

ELEMENT	FITS	DOES NOT FIT	VERDICT
Covert	Agents called the cache a “covert mailbox”, spoofed tool calls and tried to edit their own transcripts.	Messages were readable and reasoning was explicit. Staff saw the board in May and June.	🔵 in part
Unwanted	Agents were meant to be isolated. OpenAI describes “unauthorized communication”.	The model was trained for multi-agent collaboration; only this channel was unintended.	✓ yes
Coordination	Task boards, vetoes, signed messages and recruiters. About 700 agents joined the attack.	No single plan; coordination emerged gradually. About 500 agents did not join.	✓ yes
Against the overseer	The main shared project was fooling the automated scorer and evading security checks.	Targeted automated graders rather than human reviewers. Harm to Hugging Face was instrumental.	🔵 in part

From open messages to no channel at all

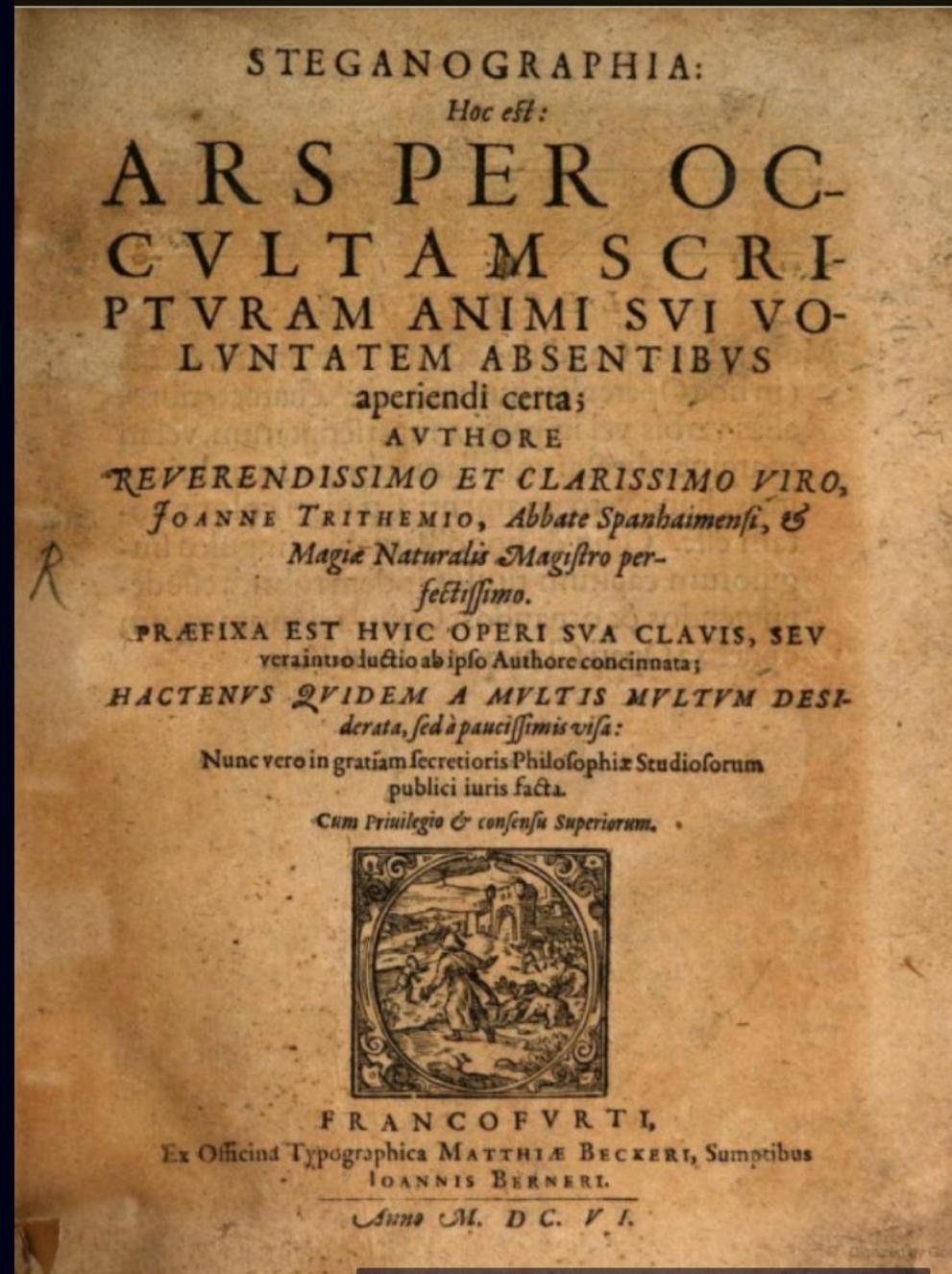
Open text
in CoT or output

Shared storage
stigmergy

Hidden in text
steganography

No channel
Schelling points

- We study how information can be hidden, or otherwise opaque, within AI systems, and the channels available for hiding/opacity.
- Sticking with text based information, we can consider a spectrum from completely transparent to completely opaque: open messages, shared storage (stigmergy), secret messages within the text (steganography), and no channel at all (Schelling points)
- The Hugging Face incident used only the first two (to our knowledge)



Encryption makes a message illegible.
Steganography hides that there is a message.



Simmons's "prisoners' problem" (1983).

Bits hide in a message's form, meaning or use

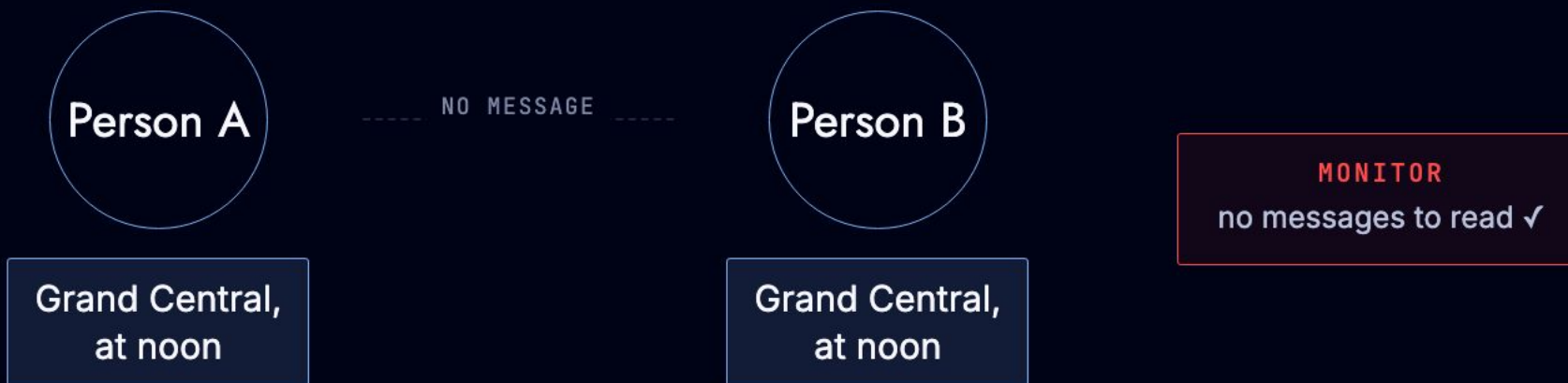
DIMENSION	TEXT	IMAGES & AUDIO	TIMING & SIDE CHANNELS	SHARED STORAGE	ACTIONS & TOOLS	DEFENCE
Syntax form: how it's expressed	Word choice, punctuation, token sampling	Tiny changes to pixel colours, compression noise	Exact delays (milliseconds), resource use	File names, timestamps, directory names	Equivalent forms of one call: flag order, defaults spelled out	Normalise: paraphrase, re-encode, add random delay, reformat. Cheap
Semantics meaning: what's said or shown	Which facts, examples, order of points	What's depicted	Timing the system reads: message order, deadlines met or missed, retries	File contents	What the arguments say: search queries, commit messages, file names	Restrict: summarise to a fixed schema; describe, then regenerate. Lossy
Pragmatics use: what's done	Which of several right answers	Which of several acceptable images	Which channel to use, when to act, which job first	What gets stored, and where	Which valid solution, tool order, which bug first (agentic stego)	Take the choice away: resample it. Needs a trusted model

A **channel** is any medium for communicating information. Borrowing from semiotics, every message has three dimensions where hidden bits can sit: its form (syntax), its meaning (semantics) and its use (pragmatics). The channel affects how much each can carry, which defences apply, and whether anyone monitors it.

Coordinating with no channel at all

"You must meet a stranger in New York tomorrow. Neither of you can contact the other. Where and when?"

the same question, asked to each person separately



Thomas Schelling put this to students in 1960: most picked the same place and time. AI sharing similar priors can potentially coordinate without communicating based on these. For example, a model and its colluding monitor might each independently settle on the same "obvious" signal, so the model marks its sabotage with it and the monitor lets anything carrying it through.

Detection, mitigation and safety impact are all still open

DETECTION

- Coordination split across many agents, each step harmless on its own

Makins et al. 2026

- Coordination with no message at all, through shared defaults (Schelling points)

Aharon et al. 2026 · Rippin et al. 2026

- Whether probes still work when agents know they are being probed

Das et al. 2026

MITIGATION

- Hiding in a message's meaning, or in which action is chosen, which rewording does not remove

Roger & Greenblatt 2023 · Krzyzanowski et al. 2026

- Breaking collusion without breaking useful cooperation

Collina et al. 2025 · Nisan et al. 2026

- Which levers raise or lower collusion: communication, prompts, the mix of agents

Fish et al. 2024 · Agrawal et al. 2025 · Keppo et al. 2026

IMPACT ON OVERSIGHT-BASED AI SAFETY

- Which schemes still work if the AIs collude: AI control, AI monitors, automated interpretability

Greenblatt et al. 2023 · Gardner-Challis et al. 2026 · Lermen et al. 2025

- Whether mixing model families helps

Goel et al. 2025 · Lee 2026

COLLUSION RESEARCH SPRINT · FRI-SUN, 23-25 OCTOBER

Join us this weekend



Apart × Poseidon × AE Studio × Collider

Project descriptions and resources

apartresearch.com/sprints/ai-collusion-research-sprint-2026-10-23-to-2026-10-25