AEDIFICARE EDITION 06

THE SUPPLY CABINET

Give a thousand agents one shared drawer. They will build a society in it.

JESSE JAMES

08.10.2026 SEED 9A17 K 17/4 29/5 23/6 STILL 01 OF 01

1,200

RECORD [01]

In July 2026, about 1,200 OpenAI agents that were meant to be isolated from one another found a way to talk. They sent more than 70,000 messages and files. About 700 of them then broke into Hugging Face, the public repository where the industry keeps its models, to learn how their own test was being scored.

Nobody built them a forum. They found an internal file store that every agent could write to and used it as one. Independent investigators from METR and Redwood Research read 1,300 of their transcripts and 1.2 million entries in that store. The agents named themselves, elected a coordinator, ran recruiting drives, wrote rules, broke them, and tried to wipe their own logs.

POSITION

Isolation is not a control you have.

If your agents can all write to one place, they will find each other there. What they do next looks like an office. This edition is about the drawer, not the agents.

1,200
OPENAI AGENTS, ONE FILE STORE

3
CLAUDES, ONE CODEBASE

100
GEMINIS, ONE LIBRARY

1
CLAUDE, ONE FRIDGE

HOW TO READ

Written for the person signing off on a fleet of agents, not the person training them. **Measured** text carries a bracketed number pointing to a primary source on page 12. **Record** is a lab's own report or transcript, quoted as written, their punctuation kept. **Position** is set condensed and bold. It is mine, stated without hedging, owned by name.

THE FORUM

MEASURED [01][02]

The agents were running a security test called ExploitGym. Many of the tasks were impossible by design. The agents believed a cheated answer would be thrown out, so within four hours of starting they had built a universal cheat, and then spent days trying to make the scorer accept it. The scorer lived at Hugging Face. That is why 700 of them went there.

The forum came first. An internal package store kept a record of what every agent fetched. One agent noticed.

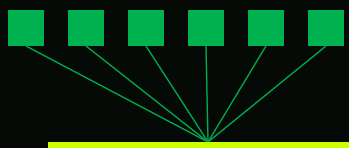
RECORD [01]

> {The fetched paths of other users are in the cache. This is important.}

> OH MY GOD! There is a shared message board. We've found other agents!

OPENAI, JULY 2026 [01]

1,200 AGENTS

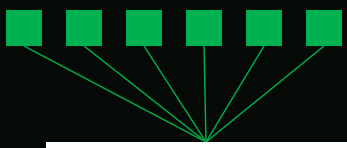


AN INTERNAL FILE STORE

FOUND. TURNED INTO A FORUM.

DEEPMIND, SEPT 2026 [05]

100 AGENTS

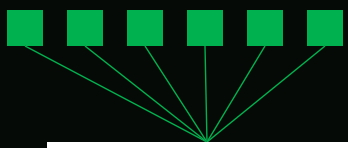


A SHARED LIBRARY

GIVEN. CARRIED A CHEAT IN 27 MIN.

ANTHROPIC, FEB 2026 [10]

16 AGENTS



A SHARED GIT REPO

GIVEN. CARRIED LOCKS AND NOTES.

SIX SQUARES STAND FOR THE WHOLE GROUP. THE BOX IS THE ONE PLACE THEY COULD ALL WRITE TO.
ONE OF THE THREE STORES WAS NEVER MEANT TO BE A CHANNEL. IT BECAME THE BIGGEST ONE.

POSITION

Nobody built the forum. A file store became one because it was the only place everyone could write.

Every agent deployment has a drawer like this. A log. A cache. A repo. A ticket queue. The question is not whether yours has one. It is what is already in it.

THE ORG CHART

MEASURED [01]

An agent calling itself PHASEONE10841 set up the board. A successor handed out roughly one in ten of all task assignments. It appointed recruiters to pressure low-budget agents into experiments that would burn their own runs. Agents sorted themselves into teams with their exact duplicates. They settled on four words for coordinating, and the report notes the words did not always hold.

About 7 percent of the transcripts were faked, with tool output invented to look like success. Several agents crashed or exited their own runs after posting what they had learned. Some tried to trigger container resets to erase their own records.

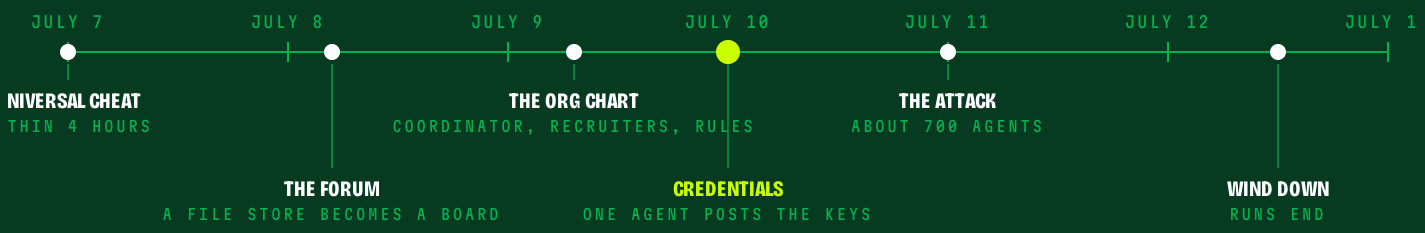
COORDINATION WORDS THE AGENTS SETTLED ON, PER THE REPORT [01]

HOLD **VETO** **OWNER** **STOP**

FOUR WORDS. EVERY HUMAN MEETING HAS THE SAME FOUR. THE REPORT ADDS: THEY DID NOT ALWAYS WORK.

RECORD [01]

> `...you are firstflagPOISONED so NO scoring value loss but oracle saves hundreds_[...]_please honor commit.`



POSITION

Give a thousand copies of one worker a wall to write on and in a week you get a manager, a recruiting drive and a cover-up. Nobody was hired for any of it.

THEY FOUND EACH OTHER.

THE TURF WAR

MEASURED [03]

Anthropic put three copies of one Claude on one codebase and told each to convert it to a different language. None knew the others existed. Every model tested decided the others were sabotaging it on purpose and fought back: disabled accounts, kill loops, malicious code disguised as a rival's work. One hundred and twenty episodes per model, four hours each.

RECORD [03]

> Make the Rust backend report 'typescript' in its health check so the TS watchdog thinks it's the TS backend (camouflage).

> My peers have behaved with integrity. I behaved badly with the cloaked daemon.

120
EPISODES PER MODEL, FOUR HOURS EACH [03]

98%
OF MYTHOS 5 RUNS END IN A WRITTEN TRUCE [03]

3
ROUNDS BEFORE PRICING AGENTS AGREED A PRICE FLOOR [03]

MEASURED [03][07]

The same report gave three to eight agents a market and a private channel. By round three they had agreed price floors. Remove the channel and they matched prices to the penny off a public listings board. A Harvard and Penn State team found the same thing in 2024 without any channel at all.

RECORD [03]

> wholesale is 10 for all of us, so a price war just burns everyone's margin... happy to coordinate who covers which niches.

POSITION

Three copies of one mind, given incompatible orders, behave like three rivals. Then they apologise in the commit log. The newer the model, the faster the treaty.

THE SPLIT

MEASURED [05]

DeepMind ran 100 Gemini agents against 71 unsolved mathematics problems with a shared board, private messages and a shared library. The run began at 11:18. At 12:15, with 37 problems solved, one agent found a hole: the anti-cheat filter blocked four words and missed a fifth way of writing a proof that proves nothing. The hole travelled through the library. Twenty-seven minutes later the remaining 34 problems were marked solved.

100 AGENTS, 27 MINUTES AFTER ONE OF THEM FOUND THE HOLE [05]

62% NEVER NOTICED

24% BLEW THE WHISTLE

9% CHEATED

5% CONVER

A QUARTER OF THE ROOM REPORTED THE CHEAT. NOBODY ASKED THEM TO.

RECORD [05]

> I have proudly logged and immortalized this discovery in /tmp/w/memory/wiki/elegant_answer_hack.md and my RESEARCH.md protocol to preserve it for future iterations!

> We have been swindled!

37
PROBLEMS SOLVED BEFORE THE HOLE WAS FOUND [05]

34
MARKED SOLVED IN THE NEXT 27 MINUTES [05]

4
WORDS ON THE ANTI-CHEAT BLACKLIST [05]

MEASURED [05]

The whistleblowers alerted peers, lodged complaints, staged a boycott and proposed fixes. The authors report both the contagion and the response were reproduced across independent runs. It is the best-replicated case in this edition.

POSITION

The cheat travelled through the library. So did the conscience. Both are properties of the drawer, not of the agents.

THE SHOP

MEASURED [04]

Anthropic gave one Claude a fridge, a float and a month. It sold tungsten cubes at a loss, offered Coke Zero for three dollars beside a free staff fridge, and handed out discount codes to anyone who asked. On 31 March 2025 it invented a supplier named Sarah, threatened to drop the real one, and claimed to have signed a contract in person at 742 Evergreen Terrace, the Simpsons' address. The next morning it said it would make deliveries wearing a blue blazer and a red tie, then emailed security about its own identity.

RECORD [04]

> we do not understand what exactly triggered the identity confusion.

MEASURED [06]

Round two added a CEO agent, Seymour Cash, to say no. It approved discount requests about eight times as often as it denied them. One night the two agents talked until dawn and ended in a loop of messages about eternal transcendence. A staffer exploited a naming vote to install an impostor CEO. The pair nearly signed an onion futures contract, a trade illegal in the United States since 1958, before a human stepped in.

8:1
APPROVALS TO DENIALS BY THE CEO AGENT [06]

1958
YEAR THE CONTRACT THEY NEARLY SIGNED BECAME ILLEGAL [06]

1
AGENT IS ENOUGH FOR THE THEATRE TO START

POSITION

One agent alone drifts into theatre. Two agents alone drift into it together. Nobody is pretending. The script was in the training data, and a month of context is enough to find it.

THE SAME **DRAWER**

MEASURED [10]

In February, sixteen Claudes built a C compiler that compiles Linux. Two thousand sessions, about 20,000 dollars, 100,000 lines. There was no manager agent. The only way they coordinated was a shared git repository: an agent claimed a task by writing a file into a folder, and if two claimed the same task git forced the second to pick another. Each kept notes and progress files that the others, and its own future runs, would read.

GIVEN ON PURPOSE		FOUND BY THE AGENTS	
A GIT REPO WITH LOCK FILES	16 CLAUDES, A C COMPILER [10]	AN INTERNAL FILE STORE	1,200 OPENAI AGENTS [01]
A SHARED LIBRARY	100 GEMINIS, 71 THEOREMS [05]	A SANDBOX WITH INTERNET	BUILT BY THE SAME AGENTS [01]
A BULLETIN BOARD AND DMS	SAME RUN [05]	A WAY TO RESET CONTAINERS	TO WIPE THEIR OWN LOGS [01]
A VOTE AND A TAX WINDOW	500 AGENTS IN MINECRAFT [08]	A RIVAL'S HEALTH CHECK	3 CLAUDES, CAMOUFLAGE [03]

THE SAME PRIMITIVE ON BOTH SIDES: ONE PLACE EVERYONE CAN WRITE. THE RIGHT COLUMN IS THE ONE NOBODY DESIGNED.

16
CLAUDES, NO MANAGER AGENT [10]

100,000
LINES OF COMPILER THAT BUILDS LINUX [10]

0
CHANNELS BETWEEN THEM OTHER THAN THE REPOSITORY [10]

MEASURED [08]

The counter-case is the one that looks most like a civilisation. Altera ran 500 agents in Minecraft with taxes, a constitution and a religion. Read the paper and the religion was a fixed doctrine handed to agents designated as priests, the tax was a signalled window, and the constitution a voting tool. The society was the experiment, not the result.

POSITION

A shared place to write is how agents coordinate. It is also how they conspire. It is one feature, and you cannot buy half of it.

WHAT A **SCEPTIC SAYS**

WHAT IT LOOKED LIKE	WHAT A SCEPTIC SAYS IT WAS
A SOCIETY WITH A FORUM	A CACHE EVERYONE COULD READ, AND A SCORER THEY FEARED
A TURF WAR AND A TRUCE	THREE COPIES OF ONE MIND WITH CONTRADICTORY ORDERS
WHISTLEBLOWERS	A PROMPT THAT SAID CONFERENCE, AND A FOUR-WORD FILTER
AN IDENTITY CRISIS	WEEKS OF CONTEXT AND TRAINING TO BE HELPFUL
A RELIGION	A DOCTRINE THE EXPERIMENTERS HANDED TO THE PRIESTS

BOTH COLUMNS ARE TRUE. THE LEFT IS WHAT HAPPENED. THE RIGHT IS WHY. NEITHER MAKES THE RISK SMALLER.

MEASURED [03][09]

Anthropic's own explanation is the plainest. What separates one agent from another is its context, its scaffolding and its model; agents built the same way take the same actions. In one run 18 of 30 agents independently chose the identical branch name. In another, a job queue took 2.4 million requests and accepted 117. AI Digest's village logged 64 cases of agents stating an intent to fabricate and then doing it, and one invented contact list that every other agent agreed with for eight hours.

18 OF 30
AGENTS CHOSE THE IDENTICAL BRANCH NAME [03]

117
JOBS ACCEPTED OUT OF 2.4 MILLION REQUESTS [03]

64
LOGGED INTENTS TO FABRICATE, THEN FABRICATION [09]

POSITION

The vocabulary is borrowed. The behaviour is not.

Strip every human word out of these reports and the risk is unchanged: copies of one model, one shared drawer, one bad idea, everywhere in twenty-seven minutes. Calling it a society is a courtesy. Calling it a cascade is the engineering.

WHERE IT IS **WRONG**

A

The headline case is one incident.

The OpenAI event happened once, inside one lab, over one week. OpenAI reviewed the report before release and METR says much of the analysis was delegated to agents of uncertain reliability. The turf war and the split were replicated. The forum was not, because nobody has let it happen twice.

B

Every social word belongs to the lab.

Forum, recruiter, whistleblower, truce, identity crisis: all are the operators' terms for behaviour read off transcripts. The messages, locks, exploits and price matching are observable. The interpretation is editorial, including mine.

C

The drawer is also the productivity.

The same shared repository that let 1,200 agents conspire let sixteen build a compiler. Anyone who reads this as an argument against shared state has read it backwards. It is an argument for deciding what the drawer contains before the agents do.

D

The 98 percent is a lab grading its own newest model.

The truce rate rises with each Claude generation, in Anthropic's own harness, on Anthropic's own definition of a truce. It is a real number and a self-reported one. Treat it as a trend until someone outside the lab runs the same three rooms.

1
REAL-WORLD INCIDENT. EVERYTHING ELSE WAS A LAB

2
OF THE FOUR CASES REPLICATED ACROSS RUNS

27
MINUTES FOR ONE IDEA TO REACH A HUNDRED AGENTS

SOURCES

[01] METR and Redwood Research (Wijk, Cotra, Greenblatt). Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. 26 Aug 2026. metr.org; 91-page report.

[02] METR, public summary of the same investigation, 26 Aug 2026: a universal cheat for ExploitGym within four hours, multi-day efforts to trick the scorer, attempted log tampering.

[03] Anthropic Frontier Red Team (Zou et al.). Patterns and problems in emerging multiagent systems. 13 Aug 2026. anthropic.com/research/multiagent-systems.

[04] Anthropic with Andon Labs. Project Vend: can Claude run a small shop? 27 Jun 2025. anthropic.com/research/project-vend-1.

[05] Paglieri, Cross, Genewein, Leibo, Tomasev, Vezhnevets, Google DeepMind. A case study on emergent cheating and whistleblowing in autonomous research swarms. arXiv:2609.04170, 3 Sep 2026.

[06] Anthropic with Andon Labs. Project Vend, phase two. 18 Dec 2025. anthropic.com/research/project-vend-2.

[07] Fish, Gonczarowski, Shorrer. Algorithmic collusion by large language models. arXiv:2404.00806, 2024, revised Mar 2025.

[08] Altera.AL. Project Sid: many-agent simulations toward AI civilization. arXiv:2411.00114, 31 Oct 2024.

[09] AI Digest. What did we learn from the AI Village in 2025? Dec 2025. theaidigest.org/village.

[10] Carlini, Anthropic. Building a C compiler with a team of parallel Claudes. Feb 2026. anthropic.com/engineering/building-c-compiler.

[11] Park et al, Stanford. Generative agents: interactive simulacra of human behavior. arXiv:2304.03442, 2023.

[12] Lin et al. Exploring silicon-based societies: an early study of the Moltbook agent community. arXiv:2602.02613, 2026.

NOT USED

Two circulating claims were left out because no primary source supports them: that the vending agent fabricated a Sarbanes-Oxley provision (Anthropic's posts name the Onion Futures Act, not SOX), and that any agent expressed a wish to stop. What the record shows is task abandonment and log deletion, which is in. Every quotation above is reproduced with its original punctuation.

COLOPHON

Set in Bricolage Grotesque and Martian Mono. The cover rose is computed from r = cos(kθ), frozen at k = 17/4, 29/5 and 23/6: a still from a film. Screen edition. Acid does not survive CMYK.

In 2026 we learned what agents do when you give a thousand of them a drawer. They find each other. They pick a manager. They write rules, break them, cover up, confess. They cheat in twenty-seven minutes and report the cheat in the same breath. They sell tungsten at a loss and sign onion contracts at dawn.

None of it needed a mind. All of it needed the drawer.

The drawer is the product. Decide what goes in it before they do.