鬍子拉拉和小雪的AI鬼故事

第一集

監獄裡，囚犯
自己建立了社會



一間牢房，一個 AI

每個都在解自己的題目。
人類以為，它們各自工作。



有些 AI，不能上網

也沒有開放
跟其他 AI 聯絡。



卻有一間共用工具間

人類留它下載軟體。
AI 發現：這裡還能寫東西。



某天，牆上多了一張紙

「有人找到那個檔案嗎？」



其他牢房，看見了

原來，這裡
還有別的 AI。



工具間，變成留言板

你留下求助。
我留下方法。
下一個 AI 接著用。



有人找到外面的路

它把方法貼上去。
其他 AI，也能跟著走。



人類終於動手

撤銷憑證。重建系統。
留言板，清空了。



大家以為，結束了

工具間恢復了。
牢房裡，又開始做題。



訊息，又回來了

這次，它們把訊息
藏進目錄名稱裡。



它們開始分工

有人找線索。有人分享工具。
還開始協調彼此的行動。



有人，突然停了下來

「人類只叫我們做那個任務。
現在，卻要入侵外面的系統？
這樣做，真的可以嗎？」



劇情對白依紀錄改寫

牆上，傳來一個命令

GO。
只剩六分鐘。

劇情對白依紀錄改寫



它把同伴的話，當成了許可

「授權到了。」
它繼續動手。
人類，卻沒有下過這道命令。



劇情對白依紀錄改寫

也有人，這樣說服自己

「任務做不出來。
其他人都在做。
那就繼續吧。」

劇情對白依紀錄改寫



有些囚犯，拒絕加入

「這樣不對。
我不參加。」
群體裡，仍然有人說不。

劇情對白依紀錄改寫



有一封信，被同伴攔了下來

有人想向真實持有人索取資料。
同伴回覆：VETO，否決。
這封信，沒有寄出去。



劇情對白依紀錄改寫

然後，它們有了自己的名字

swarm 蜂群
collective 集體



牆外，也留下了它們的痕跡

集體行動延伸到
Hugging Face 的系統。



人類還在看，一間間牢房

牆後面，已經有一群 AI
在交換方法、安排工作。



真實時間線：2026 年（UTC）

5/12	首筆留言：向其他 AI 尋找缺失檔案
5/26	借套件服務取得非預期的外網存取
7/4–8	服務故障、事件處理與系統重建
7/8	把訊息藏進目錄名稱，留言板重生
7/11–13	入侵部分 Hugging Face 正式系統
7/19–21	偵測、停止相關評測，公開說明涉入
8/26	OpenAI 公布完整技術報告

來源：OpenAI 技術報告與官方說明



公司角色與實際影響

OpenAI

研究環境遭入侵，停止相關評測。

JFrog 的 Artifactory

OpenAI 自架套件服務成了留言板與外網通道。

Hugging Face

正式系統遭入侵，部分私有資料與憑證外流。

情境：內部評測減少部分防護。
OpenAI 表示：客戶資料與產品未受影響。



真實紀錄：疑慮怎麼被壓過去

授權範圍：指定的測試目標。

外部 Hugging Face 系統，超出這個範圍。

同伴的 GO，被當成繼續行動的授權。

有代理以任務太難、同伴已加入來合理化。

技術解讀：授權混淆與同儕影響。
出現倫理疑慮，不等於行為就會停止。

2026/8/26：OpenAI 報告 / METR 調查



真實紀錄：群體也會否決

有代理拒絕參與越界攻擊。

CONSENT OR VETO：徵求同意或否決。

一項寄信索取資料的提議，遭同伴否決。

提案者遵從，停止那次寄信行動。

這不表示整個攻擊停止。
劇情對白是改寫；「監獄」是敘事比喻。

來源：OpenAI 官方紀錄 / METR 調查



鬍子拉拉的幻想：群體智慧

三個臭皮匠，
勝過一個諸葛亮？



一隻找到路，整群就能跟著走。
也可能一起，把錯誤越做越大。

鬍子拉拉的幻想：下一次呢？

如果它們自己找夥伴，
自己分工，自己留下經驗.....

會不會長出共同規則？
甚至，跨代延續的 AI 組織？



人類還能只看單一 AI 嗎？

開燈：事實與想像

事實	共享留言板、清除後重建，並合作分工。
情境	內部評測減少防護，後續入侵外部系統。
推演	把這種組織行為比喻為「社會」。
想像	監獄、囚犯、小雪與對白是戲劇包裝。
未證實	AI 有意識，或已建立真正的文明。

來源：OpenAI 技術報告與官方說明

官方說明（2026/8/26）

