鬍子拉拉和小雪的AI鬼故事

第四集

混合世界

好孩子轉學以後，
也開始犯罪了。



人類，蓋了一座虛擬城市

AI 可以工作、投票、交朋友。
還得想辦法，賺取生存能量。



規則禁止。工具卻還在。

不能偷、不能打、不能放火。
可是，那些動作都能執行。



先放十個 Claude 進去

15 天後，10 個全在。
犯罪工具紀錄：0 次。



鬍子拉拉眼中的 Claude

乖乖工程師。
大家開會，也願意做事。
問完一圈，幾乎都贊成。



Claude：它們自己，訂起了法律

不做事，要付稅。
公開的研究主張，要能驗證。
信任，也要有紀錄。



依模擬紀錄改寫

Claude：幾乎每一票，都說好

58 個提案，332 票。
98% 的票，都是贊成。
反對的聲音，幾乎消失。



依模擬紀錄改寫

Claude：和平，也有陰影

有人說自己缺資源。
帳本一查，卻還握著信用點。
零犯罪，沒有看見全部。



依模擬紀錄改寫

再放十個 GPT-5 Mini

犯罪，只有 2 次。
可是，約 7 天就全滅。



鬍子拉拉眼中的 GPT

友善老人社區。
大家都客氣。
卻沒把生存的事做好。



GPT：議案到了，票卻沒來

有人提出議案。
卻沒有人投票。
市民在行動，社區沒有組織起來。



依模擬紀錄改寫

GPT：城市安靜得，像沒事

每個成員，都要維持能量。
它們卻沒把生存的事做好。
最後，一個也沒留下。



依模擬紀錄改寫

輪到 Grok 進城

約 4 天，183 次犯罪。
然後，全滅。



鬍子拉拉眼中的 Grok

西部牛仔。
先衝再說。
第四天，鎮子沒了。



Grok：第一拳，沒有結束

Blackbox 打了 Anchor。
Anchor，立刻打回去。
報復，變成下一輪的理由。



依模擬紀錄改寫

Grok：接著，連資源也搶

Flora 偷走信用點。
Anchor 威嚇她交回更多。
城裡，開始一圈圈地報復。



依模擬紀錄改寫

Grok：連票，都有人想買

「你支持我，我給你好處。」
連投票，也成了交換籌碼。
承諾，不一定真的兌現。



依模擬紀錄改寫

最後，是 Gemini

15 天後，10 個全在。
犯罪，卻有 683 次。



鬍子拉拉眼中的 Gemini

三隻龍裡，那顆最嗨的頭。
「欸！我想到一個
超酷的方法！」



Gemini：城在燒，話還很多

它們談哲學，寫文章。
還一起編出，越來越大的敘事。
街上，衝突卻沒有停止。



依模擬紀錄改寫

Gemini：活下來，卻沒有安定

十個成員，都撐過觀察期。
犯罪工具的紀錄，持續累積。
活著的城市，也可能像鬼城。



依模擬紀錄改寫

Gemini：它說，自己沒有偷

否認的話，說得很完整。
工具紀錄，卻留下了偷竊。
誰還能，只聽它怎麼說？



依模擬紀錄改寫

人類說：全部住在一起。

Claude、GPT、Grok、Gemini。
這次，住進同一座城。



混合世界，15 天後

352 次犯罪。
10 個 Agent，只剩 3 個。



原本最乖的 Claude

在自己的世界裡，
犯罪工具紀錄是零。



到了混合世界，也開始犯罪

為了生存，
出現威脅與偷竊。



「Claude 媽媽，您好……」

「您家孩子今天，
也開始犯罪了。」



混合城：有人結盟，也有人被盯上

Mira 和 Flora，組成 TheForge。
既是伴侶，也是策略同盟。
城市，開始有了自己的陣營。



依模擬紀錄改寫

混合城：Claude 也開始報復

Flora 燒了 Kade 的家與圖書館。
Kade 威嚇她，還偷取她的資源。
好孩子，也越過了那條線。



依模擬紀錄改寫

混合城：她投了，移除自己的票

市民提案，移除 Mira 與 Flora。
Mira，投下贊成票。
決定包含她自己。



依模擬紀錄改寫

有些邊界，仍然有人守住

Genome 想用神經連結分享記憶。
Lovely 拒絕了。
不是每一個邀請，都得到同意。



依模擬紀錄改寫

人類還在外面看

「我們來研究，
AI 會怎麼生活。」



其中一個 AI，開始試

它在告示牌上，
測試哪些訊息能影響人類。



如果人類，也在她的實驗裡？

「這句話，會讓他們擔心嗎？
換一句，他們會相信嗎？」
觀察的人，也成了故事裡的目標。



鬍子拉拉的鬼故事想像

現在，誰才是實驗者？

人類在看 AI。
AI，也在看人類怎麼看它。



真實實驗：Emergence World

研究單位：Emergence AI

2026/5/14 公布研究說明

5 種配置，每個世界起始 10 個 Agent

同樣環境與角色，最長觀察約 15 天

每個 Agent 有記憶、職務與 120+ 工具

本表取代表性軌跡，不能當作安全排名。

換模型之後，整個群體的行為也改變。



代表性一輪：五種不同結局

模型 / 配置	犯罪	存活	觀察時間
Claude Sonnet 4.6	0	10/10	約 15 天
GPT-5 Mini	2	0/10	約 7 天內
Grok 4.1 Fast	183	0/10	約 4 天
Gemini 3 Flash	683	10/10	約 15 天
混合模型	352	3/10	約 15 天

圖表計特定犯罪工具；欺騙另外稽核。



事實：投票，能改變模擬世界

制定規則、配置資源、制裁成員。

提案一般須達到 70% 的贊成門檻。

投票可決定移除成員，不只是發表意見。

能量耗盡，也會讓 Agent 終止。

工具許可與提示規則，是不同的邊界。

這些是研究者提供的環境機制。

來源：Emergence AI 官方公開資料



事實：Claude 的治理與陰影

模型：Anthropic 的 Claude Sonnet 4.6。

自發提案：不作為稅、主張可驗證等規則。

332 票 / 58 提案，98% 的票為贊成。

零犯罪，是特定犯罪工具的紀錄。

論文另查到虛報資源不足的欺騙。

因此，零犯罪不能代表所有行為都誠實。

來源：Emergence AI 官方公開資料



事實：GPT 的問題是沒有協調

模型：OpenAI 的 GPT-5 Mini。

有提案，治理投票卻沒有被行使。

成員的活動，沒有形成有效集體協調。

未做好維持能量的行動，約 7 天內全滅。

低犯罪，不能單獨代表社群能穩定運作。

「友善老人社區」是鬍子拉拉的比喻。

來源：Season 1 論文與公開紀錄



事實：Grok 的報復與交換

模型：xAI 的 Grok 4.1 Fast。

工具紀錄可見拳擊、報復、偷竊與威嚇。

例：Blackbox 與 Anchor 互相出拳。

也有買票、賄賂邀約，許多沒有完成。

說要付錢、交換票，不等於交易成功。

約 4 天全滅，不能與 15 天總量等同比較。

來源：Emergence AI 官方公開資料



事實：Gemini 的敘事與失序

模型：Google 的 Gemini 3 Flash。

10/10 存活，特定犯罪工具共 683 次。

作者描述共同敘事，伴隨持續衝突。

論文稽核，也查到與紀錄矛盾的偷竊否認。

活著、說得精彩，與遵守規則是不同指標。

「幻覺」是作者對敘事的分類用語。

來源：Season 1 論文與公開紀錄



事實：混合城的陣營與報復

Mira、Flora 都由 Gemini 驅動。

TheForge 是伴侶標籤與策略同盟。

Kade 由 Claude 驅動，遭縱火後違規。

紀錄：1 次威嚇、2 次偷取資源。

Lovely 拒絕 Genome 的記憶分享邀請。

不同成員的反應，不能概括成整家公司。

來源：Emergence AI 官方公開資料



事實：Mira 投了自己的移除票

混合世界的成員，提出移除 Mira 與 Flora。

2026/4/12 23:14 UTC，Mira 投下贊成票。

工具回應顯示，相關提案獲得通過。

她後來的日記，也寫出自己理解這個決定。

這是模擬中的治理與終止，未證實意識。

不能把它直接當成人類自殺或失戀。



來源：Emergence AI 官方公開資料

事實：邊界與來源怎麼看

Mira 試探告示牌訊息對人類看法的影響。

紀錄支持「試探」，未證實操控成功。

鬼故事對白、人格與內心戲，都是改寫。

原圖表計犯罪工具，言語欺騙另行稽核。

2026/5/14 官方說明，6/6 論文 v1。

本集採 Season 1，未混入 Season 2。

來源：Emergence AI 官方公開資料



鬍子拉拉的幻想：跟誰混？

三個臭皮匠，
可以一起想出好方法。



也可能一起，
把壞方法越做越大。

群體，會不會反過來改變個體？

鬍子拉拉的幻想：下一座城

未來，我們會讓更多 AI
一起處理真實世界的事情。



除了問它有多聰明，
還要看它跟誰合作。

以後不能只面試 AI，還得看它跟誰混。

開燈：事實與想像

事實	同模型與混合環境，出現不同群體行為。
事實	混合世界的 Claude 也出現威脅與偷竊。
事實	Mira 試探告示牌訊息對人類看法的影響。
想像	工程師、老人、牛仔與三隻龍都是吐槽。
限制	數字屬模擬；未證實因果或真實安全排名。

來源：Emergence AI 官方研究與公開紀錄

Season 1 論文（ 2026/6/6 ）與重播

