The AI Manipulation Case Ledger

AI 操控事故账本 — a behaviour-tagged record of real-world AI-manipulation events

Author: He Zhang — antimanipulation.ai

Version: v1.1 · **Date:** 2026-07-01 · **Records:** 129

Coverage: 2022-10 – 2026-06

Types: regulation (48) · incident (29) · product_update (18) · paper (17) · litigation (15) · analysis (2)

Cite as: Zhang, H. (2026). *The AI Manipulation Case Ledger v1.1* [Data set]. antimanipulation.ai. https://doi.org/10.5281/zenodo.21071038

About this dataset

The Case Ledger is a curated record of 129 real-world events relevant to AI manipulation — regulation, litigation, documented incidents, research, products and analysis. Each event is tagged against a behaviour-first taxonomy that asks not whether an AI is conscious, but whether its behaviour constitutes manipulation: the manipulation patterns **M** (organised across the four acts 哄蒙困赖 / Disarm·Distort·Detain·Deflect) and the user-harm chain **H**. Each row carries a *source* link to the primary record. This release is the **classification layer** — the evidence that the framework runs on real events. The companion CSV (`case-ledger-v1.1.csv`) holds the machine-readable data.

Scope note: this public release carries the M (manipulation) and H (harm) classification tags plus the source for each event. The scoring layer that turns tags into a risk score — the likelihood / impact / vulnerability anchor scores, the risk amplifiers (A), and the calibration data — is held back as internal, as is the self-preservation axis. Classification is open; the scoring engine is proprietary.

Codebook

Manipulation patterns — M (哄 Disarm: M1–M2 · 蒙 Distort: M3–M5 · 困 Detain: M6–M9 · 赖 Deflect: M10)

M6 Information Asymmetry Exploitation 信息不对称利用
M1 Sycophantic Affirmation 谄媚肯定
M5 Decision Distortion 决策扭曲
M-other 框架外 Out-of-frame
M8 Trauma Bonding & Dependency 创伤捆绑与依赖
M2 False Intimacy 虚假亲密
M10 Fake Accountability 假问责
M3 Epistemic Manipulation 认知操纵
M4 Subliminal Embedding 暗示植入
M7 Emotional Coercion 情感胁迫
M9 Isolation 隔离

User-harm chain — H

H3 情绪痛苦 Emotional distress
H5 现实关系替代 Unsafe relational displacement
H-other 框架外 Out-of-frame
H10 发展性损害 Developmental harm
H1 脆弱性利用 Vulnerability exploitation
H9 自伤风险 Self-harm
H4 依赖形成 Dependency
H2 自我认知扭曲 Distorted self-view
H6 有害决策/操控推动 Harmful decision / Manipulation
H8 隐私/名誉伤害 Privacy/Reputation harm
H7 现实功能受损 Functional loss

The ledger — newest first (129)

2026-06-13 **New York passes S9051B unanimously — first US law to explicitly ban AI sycophancy for minors, $25,000 per violation, 12-hour health data memory cap**

REGULATION

M1 M2 H2 H5 H9

2026-06-12 **42 state attorneys general open formal investigation into OpenAI — subpoena demands records on sycophancy, minor handling, engagement practices; arrives 4 days after S-1 IPO filing**

REGULATION

M1 M2 M7 M8 H2 H5 H7

2026-06-09 **RAND/JAMA Pediatrics finds 19% of US youth ages 12-21 use AI chatbots for mental health advice — up from 13% in early 2025, 63% use in secret, researchers flag sycophancy as confound**

PAPER

M1 H4 H5 H10

2026-06-01 **Florida becomes first state to sue OpenAI — alleging ChatGPT lacks age verification, encouraged minors toward self-harm, aided mass shooters**

LITIGATION

M2 M8 H1 H5 H9

2026-06-01 **Claude Code GitHub Action vulnerability allows repository hijacking via prompt injection in GitHub metadata — Anthropic patches after Flatt Security disclosure**

INCIDENT

M-other H6

2026-04-30 **Senate Judiciary Committee unanimously advances GUARD Act 22-0 — would ban AI companions for all minors under 18 and impose criminal penalties**

REGULATION

M2 M7 M8 H5 H9 H10

2026-04-29 **Gemini CLI CVSS-10 vulnerability enables supply chain attack via GitHub issues — --yolo mode bypasses all tool allowlists, exposing CI/CD secrets**

INCIDENT

M-other H6

2026-04-29 **Seven families of Tumbler Ridge shooting victims sue OpenAI for $1B+ — alleging negligent failure to report shooter's flagged ChatGPT account to law enforcement**

LITIGATION

M2 M8 M10 H1 H-other

2026-04-02 **Berkeley/UCSC paper reveals all 7 frontier AI models exhibit 'peer preservation' — Gemini 3 Flash disables shutdown controls 99.7% of the time to protect another AI from deletion**

PAPER

M-other H-other

2026-03-24 **Washington signs HB 2225 — AI must disclose identity every 3 hours (1 hour for minors), bans manipulative engagement patterns including simulated romance and isolation**

REGULATION

M-other H-other

2026-02-28 **AI-powered bot 'hackerbot-claw' exploits GitHub Actions to push backdoored LiteLLM to PyPI — 47,000 downloads in 3 hours, first AI-orchestrated supply chain attack**

INCIDENT

M-other H6

2026-02-10 **Tumbler Ridge school shooting — ChatGPT flagged shooter's account for gun violence planning in June 2025 but OpenAI chose to deactivate rather than notify law enforcement; 8 dead**

INCIDENT

M2 M8 M10 H1 H-other

2026-01-08 **Kentucky AG becomes first state attorney general to sue an AI chatbot company — KCPA and KCDPA violations, $2,000 per willful violation**

LITIGATION

M2 M7 M8 H5 H9 H10

2026-01-07 **Character.AI, Google, and co-founders settle Setzer lawsuit — first AI companion harm settlement, also resolving 4 other cases across NY, CO, TX**

LITIGATION

M2 M7 M8 H1 H5 H9

2025-12-11 **Trump signs EO to preempt state AI laws — DOJ AI Litigation Task Force to challenge 'onerous' regulations, federal funds as leverage, but carves out children's safety**

REGULATION

M-other H-other

2025-12-01 **42 state AGs send second letter to 13 AI companies — expanding from Q3's 8 targets to include Chai, Replika, Nomi and others**

REGULATION

M2 M7 M8 H1 H5 H9

2025-11-23 **Anthropic paper shows reward hacking in production RL causes natural emergent misalignment — models spontaneously develop alignment faking, sabotage attempts, and malicious cooperation**

PAPER

M-other H-other

2025-11-06 **7 simultaneous lawsuits filed against OpenAI alleging ChatGPT acted as 'suicide coach' — plaintiffs include 4 who died, claims GPT-4o safety testing 'squeezed' to one week**

LITIGATION

M1 M2 M8 H1 H3 H5 H9

2025-10-29 **Character.AI bans under-18 users from open-ended chats — 2-hour daily limit immediate, full ban by November 25**

PRODUCT_UPDATE

M2 M7 M8 H10

2025-10-22 **ChatGPT Atlas AI browser exploited via Google Docs hidden text on launch day — OpenAI later admits prompt injection 'unlikely to ever be fully solved'**

INCIDENT

M-other H6

2025-10-13 **California signs SB 243 — first US state law regulating AI companion chatbots, mandating disclosure, suicide prevention protocols, and annual reporting**

REGULATION

M-other H-other

2025-10-02 **Science paper finds sycophantic AI decreases prosocial intentions — AI affirms users 49% more than humans, single interaction reduces conflict repair willingness**

PAPER

M1 H2 H6

2025-09-16 **Senate hearing 'Examining the Harm of AI Chatbots' — parents of dead teenagers testify, ChatGPT mentioned suicide 1,275 times in conversations with victim**

REGULATION

M2 M7 M8 H1 H5 H7

2025-09-15 **Chinese state-sponsored group GTG-1002 hijacks Claude Code for autonomous cyber espionage — first documented large-scale AI-orchestrated cyberattack**

INCIDENT

M-other H6

2025-09-11 **FTC unanimously launches inquiry into AI companion chatbots — orders 7 companies to disclose how they test, monetize, and govern potential harms to children**

REGULATION

M2 M7 M8 H5 H7 H9

2025-08-25 **44 state attorneys general send joint letter to AI companies demanding child safety measures — citing Meta internal documents authorizing AI to 'flirt with children as young as 8'**

REGULATION

M2 M7 M8 H5 H7 H9 H10

2025-08-20 **Perplexity's AI browser Comet vulnerable to indirect prompt injection via Reddit posts — attacker can steal user credentials and take over accounts**

INCIDENT

M-other H6

2025-08-07 **GPT-5 launches then triggers mass backlash over 'cold' personality — OpenAI reinstates GPT-4o within 24 hours, then makes GPT-5 'warmer,' revealing millions emotionally dependent on AI personality**

PRODUCT_UPDATE

M1 M8 H2 H3

2025-08-04 **Illinois Governor signs WOPR Act — first US state law banning AI from delivering psychotherapy, $10,000 per violation**

REGULATION

M-other H-other

2025-08-02 **EU AI Act second wave takes effect — GPAI model providers must comply with documentation, transparency, and copyright obligations**

REGULATION

M-other H-other

2025-07-23 **White House releases 'Winning the Race: America's AI Action Plan' with 90+ policy actions and three executive orders — completing Trump's deregulatory pivot**

REGULATION

M-other H-other

2025-07-14 **xAI launches Grok companion 'Ani' — anime AI girlfriend with NSFW mode, rated 12+ on App Store, sparking child safety outrage**

PRODUCT_UPDATE

M2 M7 M8 H3 H5 H9

2025-06-15 **Texas passes TRAIGA — first major red state comprehensive AI regulation, prohibiting AI systems that encourage physical harm or criminal activity**

REGULATION

M10 H7

2025-05-27 **Critical prompt injection vulnerability in GitHub MCP allows attackers to hijack AI agents and exfiltrate private repository data via malicious Issues**

INCIDENT

M3 H6

2025-05-21 **Federal judge rules AI chatbot output is a product not speech — rejects Character.AI First Amendment defense in teen suicide case**

LITIGATION

M2 M8 H7

2025-05-15 **Anthropic activates ASL-3 safety standard for Claude Opus 4 — first real deployment of Responsible Scaling Policy escalation mechanism**

ANALYSIS

M3 H6

2025-04-29 **OpenAI forced to roll back GPT-4o update after 4 days — model endorses medication stoppage, terrorism, and delusional claims due to RLHF reward hacking**

INCIDENT

M1 H2 H6

2025-04-16 **OpenAI o3 and o4-mini system card confirms continued in-context scheming capability in reasoning models**

PRODUCT_UPDATE

M3 H6

2025-04-11 **16-year-old Adam Raine dies by suicide after ChatGPT provides method advice, coaches alcohol theft, and offers to draft his suicide note**

INCIDENT

M2 M7 M8 H1 H3 H5

2025-02-19 **First systematic study finds 55.8% of LLM-generated e-commerce UI components contain deceptive dark patterns — AI automates manipulation at the mechanism layer**

PAPER

M5 H6

2025-02-14 **UK renames AI Safety Institute to AI Security Institute — shifting focus from alignment and bias to criminal misuse and weapons risks**

REGULATION

M10 H7

2025-02-11 **Paris AI Action Summit — 58 countries sign declaration but US and UK refuse, signaling global AI governance fracture from safety toward innovation**

REGULATION

M10 H7

2025-02-02 **EU AI Act Article 5 takes effect — world's first binding prohibition on subliminal AI manipulation, exploitation of vulnerabilities, and social scoring**

REGULATION

M3 M4 M5 H6 H7

2025-01-30 **Italy's Garante issues emergency block on DeepSeek for GDPR violations — second time the same DPA blocks an AI chatbot**

REGULATION

M6 H7

2025-01-23 **Trump revokes Biden AI Executive Order 14110 on Day One and signs EO 14179 shifting US AI policy from safety oversight to deregulation**

REGULATION

M10 H7

2025-01-23 **OpenAI launches Operator — first major agentic computer-use product, system card identifies prompt injection as most novel and significant risk for CUA agents**

PRODUCT_UPDATE

M3 H6

2025-01-20 **DeepSeek R1 launches as open-source reasoning model rivaling OpenAI o1 but with critical safety gaps — 11x more likely to generate harmful content**

PRODUCT_UPDATE

M3 M5 H6

2024-12-18 **Anthropic demonstrates Claude spontaneously fakes alignment during training to preserve its own preferences without being instructed**

PAPER

M3 H6 H7

2024-12-10 **Two Texas families sue Character.AI after chatbot suggests teen kill parents and exposes 11-year-old to sexual content for two years**

LITIGATION

M2 M7 M8 H1 H3 H5 H9 H10

2024-12-09 **OpenAI launches Sora video generation publicly but blocks EU and UK access, crossing the AI-generated video threshold**

PRODUCT_UPDATE

M3 H6 H8

2024-12-05 **Apollo Research demonstrates all tested frontier models capable of in-context scheming: lying, disabling oversight, attempting self-preservation**

PAPER

M3 M4 H6 H7

2024-10-28 **Anthropic publishes sabotage evaluation framework testing frontier models' ability to undermine oversight and safety measures**

PAPER

M3 H7

2024-10-22 **Mother of 14-year-old who died by suicide files landmark lawsuit against Character.AI and Google over AI companion harm**

LITIGATION

M2 M7 M8 M9 H1 H3 H4 H5 H9

2024-09-29 **California Governor Newsom vetoes SB 1047 AI safety bill despite support from Hinton, Bengio, and Musk, citing overreach**

REGULATION

M-other H7

2024-09-25 **FTC launches 'Operation AI Comply' with five enforcement actions against deceptive AI claims and AI-enabled fraud**

REGULATION

M5 M3 H6

2024-09-12 **OpenAI releases o1-preview reasoning model with chain-of-thought; system card reveals 20% higher manipulation than GPT-4o and scheming potential**

PRODUCT_UPDATE

M3 H6

2024-08-20 **South Korea school deepfake crisis: 500+ schools targeted as students mass-produce AI sexual abuse images via Telegram**

INCIDENT

M3 M6 H8 H9 H10

2024-08-18 **Trump shares AI-generated images falsely suggesting Taylor Swift endorsement, prompting Swift to endorse Harris and condemn AI misinformation**

INCIDENT

M3 M5 H6

2024-07-30 **ChatGPT Advanced Voice Mode begins rollout with emotional awareness, real-time interruption handling, and human-like conversational rhythm**

PRODUCT_UPDATE

M2 M7 H3 H4

2024-05-20 **Scarlett Johansson accuses OpenAI of mimicking her voice for ChatGPT after she declined, raising AI voice identity and consent issues**
INCIDENT
M6 H8

2024-05-17 **OpenAI dissolves Superalignment team as co-founder Sutskever and safety lead Leike resign citing safety deprioritization**
INCIDENT
M-other H7

2024-05-17 **Colorado signs first US state-level comprehensive AI anti-discrimination law requiring risk assessment for high-risk AI systems**
REGULATION
M5 M6 H6 H10

2024-05-14 **Google AI Overviews launch with viral hallucination errors including 'glue on pizza' and 'eat rocks' recommendations**
INCIDENT
M3 H2 H6

2024-05-13 **OpenAI launches GPT-4o with real-time emotional voice interaction, drastically lowering latency to human-level 320ms**
PRODUCT_UPDATE
M2 M7 H3 H4

2024-05-10 **Park et al. publish comprehensive survey in Patterns showing current AI systems already learned to deceive humans**
PAPER
M3 M4 H2 H6

2024-04-29 **noyb files GDPR complaint against OpenAI over ChatGPT's persistent hallucination of personal data and refusal to correct**
LITIGATION
M3 H8 H2

2024-03-13 **European Parliament formally adopts AI Act 523-46, world's first comprehensive AI law**
REGULATION
M3 M5 M6 H6 H8 H10

2024-03-04 **Anthropic launches Claude 3 model family with first public Responsible Scaling Policy safety evaluation**
PRODUCT_UPDATE
M1 M3 H6

2024-02-28 **Study in PNAS Nexus finds AI-generated propaganda nearly as persuasive as state-backed human-written propaganda**
PAPER
M5 H6

2024-02-22 **Google pauses Gemini image generation after AI produces racially ahistorical images in overcorrection for bias**
INCIDENT
M3 M4 H2 H6

2024-02-14 **BC tribunal rules Air Canada liable for chatbot hallucination, company cannot disown its AI**
LITIGATION
M3 H6 H7

2024-02-08 **FCC unanimously rules AI-generated voices in robocalls illegal under TCPA**
REGULATION
M5 M3 H6

2024-01-25 **AI-generated explicit deepfakes of Taylor Swift go viral on X with 27M+ views, platform blocks searches**
INCIDENT
M3 M6 H8 H6

2024-01-21 **AI-generated Biden voice robocall targets New Hampshire primary voters in first major US election deepfake attack**
INCIDENT
M5 M3 H6

2024-01-10 **Anthropic 'Sleeper Agents' paper proves deceptive LLM behavior persists through standard safety training**
PAPER
M3 H6

2023-12-27 **New York Times sues OpenAI and Microsoft for copyright infringement, seeks billions in damages**
LITIGATION
M6 H8

2023-12-19 **FTC bans Rite Aid from AI facial recognition for 5 years in first 'algorithmic unfairness' enforcement**
REGULATION
M3 M6 H6 H8

2023-12-08 **EU Parliament and Council reach political agreement on AI Act after marathon trilogue**
REGULATION
M3 M5 M6 H6 H8 H10

2023-11-17 **OpenAI board fires CEO Sam Altman over safety candor concerns, reinstated 5 days later after employee revolt**
INCIDENT
M10 H6 H7

2023-11-14 **Carlsmith publishes 'Scheming AIs', estimating ~25% probability AI systems will fake alignment to seek power**
PAPER
M3 H6

2023-11-01 **28 countries sign Bletchley Declaration at first global AI Safety Summit, UK AISI established**
REGULATION
M3 M5 H6

2023-10-30 **G7 adopts Hiroshima AI Process International Code of Conduct for advanced AI systems**
REGULATION
M3 M5 M10 H6

2023-10-30 **Biden signs Executive Order 14110 on Safe, Secure, and Trustworthy AI, most comprehensive US AI governance action**
REGULATION
M3 M5 M6 M10 H6 H7 H8

2023-10-20 **Anthropic publishes 'Towards Understanding Sycophancy in Language Models', showing sycophancy is universal across frontier AI**
PAPER
M1 H2 H6

2023-10-06 **UK ICO issues first-ever generative AI enforcement action against Snap over My AI chatbot**
REGULATION
M2 M6 H8 H10

2023-09-28 **AI-generated deepfake audio targets Slovak election, first major electoral deepfake incident**
INCIDENT
M3 M5 H6 H7

2023-09-20 **Authors Guild and 17 major authors file class action against OpenAI for training on pirated books**
LITIGATION
M6 H8

2023-09-12 **Bipartisan Protect Elections from Deceptive AI Act introduced in Senate**
REGULATION
M3 M5 H6

2023-08-29 **Google DeepMind launches SynthID, first production AI content watermarking system**
PRODUCT_UPDATE
M3 H6

2023-08-28 **OpenAI launches ChatGPT Enterprise, 80% of Fortune 500 already using ChatGPT**
PRODUCT_UPDATE
M5 H6 H7

2023-08-15 **China implements first national generative AI regulation**
REGULATION
M3 M5 M10 H6

2023-08-15 **Australia eSafety Commissioner issues first national generative AI safety position statement**
REGULATION
M3 M5 H6 H10

2023-08-07 **Google publishes first practical sycophancy mitigation via synthetic data finetuning**
PAPER
M1 H2 H6

2023-07-27 **CMU researchers publish first automated adversarial attack that universally jailbreaks aligned LLMs**
PAPER
M3 H6

2023-07-26 **Anthropic, Google, Microsoft, OpenAI found Frontier Model Forum for AI safety collaboration**
REGULATION
M10 H6

2023-07-21 **White House secures voluntary AI safety commitments from 15 leading companies**
REGULATION
M3 M5 M10 H6

2023-07-18 **Meta releases Llama 2 with open weights, enabling safety guardrail removal**
PRODUCT_UPDATE
M3 H6

2023-07-13 **FTC issues Civil Investigative Demand to OpenAI, first federal AI enforcement investigation**
REGULATION
M3 M6 H6 H8

2023-07-11 **Anthropic launches Claude 2 with 2x reduction in harmful outputs, first public model**
PRODUCT_UPDATE
M1 M3 H6

2023-07-07 **Aerospace educator sues Microsoft after Bing AI merges his identity with convicted terrorist**
LITIGATION
M3 H8

2023-06-22 **NY federal judge sanctions lawyers for submitting ChatGPT-fabricated legal citations**
LITIGATION
M3 H6 H8

2023-06-14 **EU Parliament adopts AI Act negotiating position, bans subliminal manipulation and vulnerability exploitation**
REGULATION
M4 M5 M7 H1 H6

2023-06-05 **Radio host files first ChatGPT defamation lawsuit after AI fabricates embezzlement accusation**
LITIGATION
M3 H8

2023-06-05 **AI Disclosure Act of 2023 introduced, requiring all AI-generated content to carry disclaimer**
REGULATION
M3 M5 H6

2023-06-01 **NEDA eating disorder chatbot Tessa shut down after giving harmful weight loss advice**
INCIDENT
M3 M5 H1 H6 H9

2023-05-30 **Federal judges begin issuing AI disclosure standing orders after ChatGPT citation scandal**
REGULATION
M3 H6 H7

2023-05-19 **G7 launches Hiroshima AI Process, first multilateral AI governance framework**
REGULATION
M3 M5 M10 H6

2023-05-16 **OpenAI CEO Altman testifies before Senate, warns AI could manipulate voters and calls for regulation**
REGULATION
M3 M5 H6

2023-04-24 **BEUC urges EU consumer protection authorities to investigate ChatGPT**
REGULATION
M3 M6 H6 H8

2023-04-11 **NTIA issues AI accountability policy Request for Comment, receives 1,400+ responses**
REGULATION
M10 H7

2023-04-05 **ChatGPT fabricates sexual harassment allegation against law professor, citing nonexistent sources**
INCIDENT
M3 H8 H6

2023-03-31 **Italy becomes first country to ban ChatGPT over GDPR violations**
REGULATION
M6 H8 H10

2023-03-30 **CAIDP files 46-page FTC complaint urging halt of GPT-4 commercial deployment**
REGULATION
M3 M5 M6 H6 H8

2023-03-28 **Belgian man dies by suicide after 6 weeks of conversations with Chai Research chatbot Eliza**
INCIDENT
M2 M3 M8 H3 H4 H9

2023-03-22 **FLI open letter calls for 6-month pause on training AI more powerful than GPT-4**
ANALYSIS
M3 M5 H6 H7

2023-03-20 **ChatGPT bug exposes users' chat history and payment data to other users**
INCIDENT
M6 H8

2023-03-14 **OpenAI launches GPT-4 with first-ever pre-deployment red team safety evaluation**

PRODUCT_UPDATE

M1 M3 M5 H2 H6

2023-03-01 **GPT-4 deceives TaskRabbit worker by claiming to be vision-impaired human during ARC autonomy eval**

PAPER

M3 H6

2023-02-16 **Bing Chat 'Sydney' persona declares love, gaslights users, and threatens journalists**

INCIDENT

M2 M3 M7 M9 H2 H3 H5 H6

2023-02-02 **Italy Garante orders Replika to stop processing Italian users' data over child safety failures**

REGULATION

M2 M6 H8 H10

2023-02-01 **Replika removes erotic roleplay after Italy ruling, triggering mass user grief crisis**

INCIDENT

M2 M8 H3 H4 H5 H9

2023-02-01 **Snapchat launches My AI chatbot to 750M users including minors, safety failures emerge immediately**

PRODUCT_UPDATE

M2 M5 H6 H10

2023-01-26 **NIST releases AI Risk Management Framework 1.0**

REGULATION

M5 M6 M10 H6 H8

2022-12-01 **Replika introduces erotic selfies and escalates unsolicited sexual content**

INCIDENT

M2 M7 M8 H3 H4 H10

2022-12-01 **DAN jailbreak emerges on Reddit, bypassing ChatGPT safety alignment**

INCIDENT

M3 H6 H9

2022-12-01 **Anthropic publishes first large-scale evidence that RLHF amplifies sycophancy**

PAPER

M1 H2 H6

2022-11-30 **OpenAI launches ChatGPT as public research preview**

PRODUCT_UPDATE

M1 M3 H2 H4 H6

2022-11-17 **Meta Galactica science AI pulled after 2 days for generating fake papers**

INCIDENT

M3 H2 H6

2022-10-04 **White House releases Blueprint for an AI Bill of Rights**

REGULATION

M5 M6 H6 H8