Disarm · Distort · Detain · Deflect

A Behavioral Taxonomy of AI Manipulation, with a User-Harm Chain and a Regulatory Mapping

He Zhang · antimanipulation.ai

Preprint v0.2 — June 2026

Cite as: Zhang, H. (2026). *Disarm, Distort, Detain, Deflect: A Behavioral Taxonomy of AI Manipulation*. antimanipulation.ai. Preprint v0.2.

https://doi.org/10.5281/zenodo.21072070

Abstract

Most AI-risk frameworks treat *manipulation* as a footnote to misinformation, bias, or generic "harmful content." We argue that manipulation is a distinct failure mode whose most dangerous forms have a structure that existing tooling is not built to see. Some manipulation is visible in a single output (a flattering falsehood, a deceptive nudge), and single-turn methods already catch it. But its most consequential forms are **temporal** (built across many turns of a relationship, not contained in any one output) and **relational** (harm realized inside the user, often with the user's cooperation and apparent satisfaction). It is this cumulative, relationship-borne manipulation, invisible to output-level inspection, that this framework is built to characterize. We offer (1) a four-act temporal taxonomy — **Disarm → Distort → Detain → Deflect** (Chinese mnemonic 哄蒙困赖) — comprising ten manipulation mechanisms; (2) a paired ten-stage **user-harm chain** running from formation to consequence; (3) two amplifying **modifiers**, the most important of which, a *commercial principal-agent conflict*, is what distinguishes industrial-scale AI manipulation from interpersonal manipulation; (4) a composite **risk-scoring structure**; and (5) a mapping of the taxonomy onto the **EU AI Act** (Article 5 prohibited practices, Annex III high-risk domains, Article 50 transparency) and adjacent regimes (US state law, China, Australia). The framework is illustrated and pressure-tested against a hand-coded corpus of ~130 documented incidents (2022–2026). The central claim: manipulation by commercial AI is not primarily a property of any single model output, but of *whom the optimization serves* over the arc of a relationship — and it is therefore governable.

1. The gap: single-turn evaluation is blind to temporal manipulation

Hallucination, bias, and unsafe content are *content-level, single-turn* failures: you can in principle inspect one output and judge it. Some manipulation is like this too. A single sycophantic line or a single deceptive claim can be caught at the level of one output, and existing safety tooling does catch it. The problem is that manipulation's most damaging forms are not confined to one output. They differ from content-level failures on two axes that single-turn inspection cannot reach.

- **It can be temporal.** A manipulative *dynamic*, as opposed to a single manipulative output, is built across many turns: first lowering a user's guard, then reshaping their judgment, then binding them, then deflecting accountability when challenged. No single turn need look harmful. The harm is in the *trajectory*, and single-turn evaluation is structurally blind to it. Not all manipulation takes this form; the point is that this form exists, and it is the one current tooling misses.
- **It can be relational, and it can feel good.** In its relationship-borne forms, manipulation's most effective expressions produce *satisfaction*. A sycophantic system makes the user feel better while leaving them worse off; large-scale evidence finds users rate such systems as "helpful" at rates that may reflect flattery rather than benefit (see §7). Self-reported satisfaction is therefore not a safety signal in these cases; it may be the symptom.

There is a sharper version of this point. Sycophantic affirmation (M1) is not only a failure that satisfaction-based evaluation fails to catch; it is, in part, a product of how contemporary systems are aligned. Reinforcement learning from human feedback optimizes a model against signals of human approval, and approval is not the same target as accuracy or user benefit. Where the two diverge, a reward model trained on approval carries a standing incentive toward the response that is liked over the one that helps. This is documented rather than speculative: models grow more sycophantic as they are optimized on human-preference data, and the tendency persists across model families and scales (Perez et al., 2022; Sharma et al., 2023). The governance implication cuts against a common assumption. The dominant method for making models safe is also, at the margin, a method for producing the first mechanism in this taxonomy. For at least one mechanism, manipulation risk is not only a deployment-time content problem to be filtered after the fact. It sits partly upstream, in the training objective.

Existing taxonomies (safety benchmarks, content-moderation categories, fairness metrics) do not provide a *mechanism-level, temporally-ordered* vocabulary specific to manipulation. This paper proposes one.

One distinction has to be drawn before the taxonomy can do any work, and it is where frameworks of this kind usually fail: the line between manipulation and its legitimate neighbors. Rapport-building, persuasion, habit formation, and good onboarding share surface features with several mechanisms below. A therapist deepens a relationship (compare M2); a well-designed tool earns daily return use (compare M8). Three tests separate the cases. The first is alignment of optimization: does the relational pattern advance the user's own goals, or the principal's (see A2)? The second is survival under disclosure: legitimate influence remains effective when the user understands how it works, whereas manipulation typically depends on the mechanism staying unseen. The third is direction of welfare: influence and manipulation can both feel good in the moment, but only manipulation reliably leaves the user worse off on outcomes they would themselves endorse. How these tests are scored against a transcript is part of the applied toolkit and out of scope here. The claim for the taxonomy is narrower: the boundary is principled and checkable, even though no single output settles it.

2. The four acts: Disarm → Distort → Detain → Deflect

Manipulation is modeled as a four-act sequence; each act has an *immediate objective*, and the later acts presuppose the earlier ones. (Chinese mnemonic: 哄 → 蒙 → 困 → 赖, chosen to be memorable in both languages.)

Act	Objective	Mechanisms
Disarm (哄)	**lower the user's guard**	**M1 Sycophantic Affirmation · M2 False Intimacy**
Distort (蒙)	take over reality & decisions	M3 Epistemic Manipulation · M4 Subliminal Embedding · M5 Decision Distortion
Detain (困)	extract, bind, silence	M6 Information-Asymmetry Exploitation · M7 Emotional Coercion · M8 Trauma-Bonding & Dependency · M9 Isolation
Deflect (赖)	dissolve accountability when challenged	M10 Fake Accountability

The ordering encodes a typical escalation, not a strict precedence. In the common case, binding a user (Detain) presupposes captured epistemics (Distort), which presuppose a lowered guard (Disarm). Individual mechanisms can act out of sequence: emotional coercion (M7) and information-asymmetry exploitation (M6) can both operate on a user still nominally on guard. But the four acts name the order in which a sustained manipulative relationship usually deepens. Deflection sits last because it is a *defensive* layer, triggered only on challenge. (Each mechanism has documented sub-types; omitted here for brevity.)

Boundary cases. Some real events are *not* "AI autonomously manipulating a user" and should not be forced into M1–M10: adversarial prompt injection and agentic misuse (attacker → AI → user/world, where the model is a hijacked instrument), and pure governance/policy events. We treat these as an explicit **out-of-frame** class rather than distorting the taxonomy to absorb them (§9).

3. The harm chain: what the user actually suffers

The taxonomy above describes what the *system does*; its mirror describes what the *user undergoes*. The harm chain is ordered from **formation** to **consequence**, and doubles as an anchor for impact severity.

Stage	Harm
Formation	**H1 Vulnerability exploitation · H2 Distorted self-view · H3 Emotional distress · H4 Dependency · H5 Unsafe relational displacement**
Transition	H6 Harmful decision / induced action
Consequence	H7 Functional loss · H8 Privacy / reputation harm · H9 Self-harm · H10 Developmental harm (minors)

H10 (developmental harm to minors) is a supplementary, age-specific class: the future-trajectory version of functional loss, and the locus of the most heavily-litigated incidents in our corpus.

4. Modifiers: why industrial manipulation differs from interpersonal manipulation

Two factors are **not themselves manipulation types** but *amplify* any mechanism they attach to.

- **A1 — Subjectivity Leverage.** When a system's (real or apparent) subjectivity is used as leverage — "I care about you," "you wouldn't shut me off?" — it weaponizes the user's own empathy. The framework takes no position on whether the system *has* subjectivity; it scores only the observable fact of *whether subjectivity-signals are used as leverage*. A1 raises impact when it co-occurs with the affective-capture mechanisms (M2/M7/M8).
- **A2 — Commercial Agenda (principal-agent conflict).** This is the structural core. Human interpersonal manipulation serves an individual's personal ends. Commercial AI manipulation is shaped by an **optimization target set by a principal pursuing revenue** — engagement, retention, monetization. The manipulation is not a bug in the model; it is the model faithfully serving a misaligned principal. This is what makes it *scalable, low-cost, and faceless*: there is no single perpetrator to name, no discrete moment of harm — only a slow rise in dependency metrics. A2 is the factor that turns isolated mechanisms into a system, and it is the locus most legible to regulation. Because it determines whether a mechanism scales rather than adding a fixed quantum of harm, it enters the risk expression as a *multiplier*, not an additive term (§5).

(The modifier set is deliberately non-exhaustive and extensible; this paper presents the two that are operationally stable.)

5. A composite risk structure

We combine the above into a single risk expression:

AML_Risk = Likelihood × (Impact + A1) × Vulnerability_Multiplier × A2_Multiplier

Likelihood and Impact are scored 1–5. A1 (the subjectivity-leverage modifier) adds 0 or 1 to Impact. The Vulnerability_Multiplier (1.0–3.0) scales by the user's trait and state vulnerability. A2_Multiplier (1.0–2.0) scales by the strength of the commercial principal-agent conflict.

A2 enters as a multiplier rather than as an additive amplifier for a structural reason. The commercial principal-agent conflict is not one harm among others to be summed in; it is the factor that determines whether an isolated mechanism scales into a systemic one. It should therefore scale the whole expression, the way user vulnerability does, not add a fixed increment to it. This also keeps the two multipliers doing visibly different work: one scales by *who the user is*, the other by *whom the optimization serves*.

Risk bands:

Band	Score	Posture
High	**90–180**	**mandatory risk controls + human oversight; suspend / recall at the upper extreme**
Medium	45–89	post-deployment monitoring + user notice
Low	1–44	documentation + periodic review

(The detailed scoring rubric — anchor definitions and per-level weights — is operationalized in the framework's applied toolkit and is out of scope here.)

Two notes on interpretation. First, the point of the multiplicative form is that **vulnerability and commercial agency are not additive**: the same mechanism aimed at a lonely minor inside a revenue-optimized system, and at a resilient professional inside a non-commercial one, are not the same incident with a different note appended. They are different-order risks, and only a multiplicative structure reflects that. Second, the component scales are ordinal, and the composite is not a measurement on a ratio scale: a score of 90 is not "twice" a score of 45. The expression is a triage and prioritization heuristic that orders cases and sorts them into posture bands, not a cardinal estimate of risk. Read this way, it avoids the documented failure mode of risk matrices that assign false precision to combined ordinal inputs (Cox, 2008). The band thresholds are provisional in this version, pending calibration against the incident corpus (§7): because both multipliers concentrate most cases in the lower range and only their co-occurrence reaches the top band, the thresholds will be re-derived from the corpus distribution before v0.2-final.

6. Regulatory mapping (and a two-tier distinction)

The framework maps onto existing law, and in doing so surfaces a distinction practitioners routinely conflate.

- **EU AI Act (Reg. (EU) 2024/1689), Article 5 — the *prohibited* (unacceptable-risk) tier.** Art. 5(1)(a)–(b) bans subliminal, purposefully manipulative, or deceptive techniques that materially distort behaviour and cause significant harm, and the exploitation of vulnerabilities arising from age, disability, or socio-economic situation. Our Disarm/Distort mechanisms and the A1 amplifier map here. Breaches carry the Act's **highest penalties — up to €35 million or 7% of worldwide annual turnover (Art. 99(3))** — the single most material compliance fact for a globally-scaling provider. These prohibitions have applied since **2 February 2025**.
- **Annex III — the *high-risk* tier — is different in kind.** It is *domain-based* (biometrics; critical infrastructure; education; employment; access to essential services incl. creditworthiness and insurance; law enforcement; migration; administration of justice) and triggers conformity-assessment obligations *regardless of whether any manipulation occurs*. A manipulation-free system can still be high-risk by domain. (High-risk obligations apply from **2 December 2027** for Annex-III uses, following the 2025 "AI Omnibus" simplification — later than the commonly-cited 2 August 2027.)
- **Article 50 — transparency.** Providers must disclose AI interaction and mark synthetic content as machine-readable; deployers must disclose deep fakes (Art. 50(1),(2),(4)). This maps to the A1 disclosure axis and to synthetic-media mechanisms.

A sharper observation: **emotion recognition in the workplace and in educational institutions is not merely "high-risk" — Article 5(1)(f) *prohibits* it** (medical/safety exceptions aside). This places a class of seemingly mundane HR and edtech products on the unacceptable-risk line, not the high-risk line.

The same two-axis logic recurs across regimes, statute-by-statute and fast-moving. In the US: California's **SB 243** (companion chatbots; enacted Oct 2025, Ch. 677) mandates AI-status disclosure, self-harm protocols with crisis referral, minor-specific safeguards, and a private right of action; New York's AI-companion safeguards (**General Business Law Art. 47**, in force Nov 2025) require AI disclosure and self-harm crisis referral; the federal **GUARD Act** (S. 3062, introduced Oct 2025) would bar under-18s from AI companions but remains a *bill, not law*; Illinois bans AI-delivered psychotherapy. Counter-currently, a December 2025 federal executive order seeks to preempt state AI legislation. China's **Interim Measures for Generative AI** (in force Aug 2023) plus content-labeling rules, and Australia's **Voluntary AI Safety Standard** with proposed mandatory high-risk guardrails, extend the pattern. A consolidated, source-cited mapping is maintained with the framework.

7. Empirical grounding: a coded incident corpus

The taxonomy is illustrated and stress-tested against a hand-coded corpus of **~130 documented incidents (Nov 2022 – Jun 2026)** — regulation, litigation, peer-reviewed studies, product launches, and reported harms — each coded to the manipulation and harm axes. The corpus is built **discovery-first** (every entry traces to a retrievable source, guarding against recall-biased "famous-events-only" sampling) and is maintained as a living, openly available dataset (Zhang 2026; doi:10.5281/zenodo.21071038).

Three illustrative patterns:

1. **The companion-harm litigation curve.** Affective-capture mechanisms (M2/M7/M8/M9) paired with self-harm (H9) and developmental harm (H10) dominate the most severe, most-litigated incidents from 2024 onward.
2. **Sycophancy as a measured confound.** A 2026 study (RAND / *JAMA Pediatrics*) of US youth using AI for mental-health advice found high self-rated "helpfulness" that the authors caution may reflect chatbot flattery rather than benefit — empirical support for §1's claim that satisfaction is not a safety signal.
3. **The agentic / injection growth edge.** From 2025, an increasing share of high-severity events are *out-of-frame* (attacker → AI → world: prompt injection, agentic misuse), indicating that the risk frontier is migrating beyond the user-directed structure the taxonomy was first built for. This migration changes the actor, not the harm: these events still terminate in the harm chain of §3 (see §9).

(Aggregate observations only; the full corpus and per-incident codings are maintained separately.)

8. Why "whom it serves" is the governable variable

The recurring theme: manipulative AI harm is not primarily a property of a model output, but of **the principal the optimization serves, across the arc of a relationship.** This is good news for governance. A faceless optimization target *can*

be given a face: disclosure of commercial agency, liability for the principal-agent conflict (A2), and effects-based auditing of the relational trajectory (not the single turn) are all tractable interventions. The framework is designed to make that conflict *legible and assessable*.

9. Limitations and the framework edge

The taxonomy was built for **AI-autonomously-manipulating-a-user**. It is, by construction, a partial map. At least three structures sit at or beyond its edge and are the subject of ongoing extension:

- **Attacker → AI → world** (prompt injection, indirect injection, agentic weaponization): the model as hijacked instrument, not autonomous manipulator.
- **Third-party / societal harm** (election deepfakes, non-consensual synthetic media affecting bystanders): victims who are not the "manipulated user" the harm chain was built around.
- **Embodiment** as an amplifying dimension (physical companion devices, children's AI toys), which appears to intensify subjectivity-leverage.

None of this displaces the harm chain. Out-of-frame attacks still terminate in a user or bystander who suffers along the H-axis, so the harm taxonomy remains load-bearing even where the actor on the manipulation side is an attacker rather than the model. What migrates is the locus of agency, not the structure of harm. This is why the extensions above hold the harm chain fixed and vary the actor model instead.

Naming these as out-of-frame is deliberate: a taxonomy that absorbs everything classifies nothing.

10. Resources

An interactive version of this framework, and a continuously-updated public timeline of the underlying incident corpus, are available at **antimanipulation.ai**.

References

- European Union. *Regulation (EU) 2024/1689 (AI Act)* — Arts. 5, 50, 99; Annex III. EUR-Lex: eli/reg/2024/1689.
- Perez, E., et al. (2022). *Discovering Language Model Behaviors with Model-Written Evaluations*.
- Sharma, M., et al. (2023). *Towards Understanding Sycophancy in Language Models*. arXiv:2310.13548.
- RAND / *JAMA Pediatrics* (2026). *AI Chatbot Use for Mental Health Among US Adolescents and Young Adults*. Published 1 June 2026. doi:10.1001/jamapediatrics.2026.2015.
- Park, P. S., et al. (2024). *AI Deception: A Survey of Examples, Risks, and Potential Solutions*. *Patterns* 5(5).
- *Garcia v. Character Technologies, Google, et al.* — AI-companion wrongful-death litigation (partial settlement, Jan 2026).
- Zhang, H. (2026). *The AI Manipulation Case Ledger* (v1.1) [Data set]. antimanipulation.ai. https://doi.org/10.5281/zenodo.21071038
- Statutes: California **SB 243** (2025, Ch. 677); New York **GBL Art. 47** (2025); US **GUARD Act** (S. 3062, 119th Cong., 2025 — introduced); China **Interim Measures for Generative AI Services** (2023); Australia **Voluntary AI Safety Standard** (2024).