MACHINE-TO-MACHINE COMMERCE · PART II

The Machine *Attack Surface*

Your AI agent cannot be seduced. It can still be persuaded. Marketing does not disappear — it becomes less anthropological and more adversarial-computational.

John Rector · August 8, 2026 · johnrector.me/2026/08/08/machine-attack-surface/

ABSTRACT

The standard objection to agentic commerce is that it kills marketing: an AI procurement agent has no vanity to flatter, no fear of missing out, no status anxiety. This paper argues the objection mistakes the medium for the mechanism. Persuasion is any intervention that moves a decision away from ground-truth utility, and AI agents possess their own well-documented susceptibilities: ordering and position effects, context-window dead zones, brand priors absorbed from training data, anchoring and decoy effects, sycophancy, retrieval bias, schema sensitivity, and wording that resonates with a particular model's learned representations. Mapping each vector to the peer-reviewed and industry evidence, I argue that marketing will not shrink but transform — from an anthropological discipline aimed at human psychology into an adversarial-computational discipline aimed at machine cognition — and I sketch what defense looks like for the agents doing our buying.

01 The Disarmed Arsenal

In the first paper in this series, I argued that when fiduciary AI agents take over purchasing, brands will pivot from persuading buyers to corrupting the data those agents trust — captured labs, forged certificates, spec-hacked products. That argument assumed the direct route was closed: that you cannot advertise to an algorithm. This paper examines the assumption, and finds it half true in a way that matters enormously.

The human levers really are disarmed. An agent evaluating cookware has no response to a celebrity chef's endorsement, because it has no aspiration. Sexual imagery lands nowhere, because there is no limbic system to shortcut. Status signaling fails, because the agent will never be seen carrying the product. Emotional storytelling, scarcity theater, the countdown timer, the influencer haul — each of these works by hijacking some evolved circuit for social proof, desire, or loss aversion. The agent has none of those circuits. A century of marketing craft, from Bernays onward, addresses hardware the buyer no longer runs on.

The mistake is concluding that persuasion therefore ends. Persuasion, properly defined, is any intervention that moves a decision away from ground-truth utility toward the intervener's interest. Humans are movable through emotion and identity. Models are movable through structure: how options are ordered, where information sits in a context window, which phrasings align with learned representations, which sources a retrieval pipeline favors. These susceptibilities are not speculative. Nearly every one has been demonstrated, measured, and published.

> *"The agent has no vanity, no fear, no tribe. What it has is a probability distribution — and a probability distribution can be marketed to."*

02 The Anatomy of a Machine Decision

To see where the attack surface lies, walk through what an agent actually does when told "buy me the best cordless drill under $150." Four stages, each with its own exploitable structure.

Retrieve. The agent assembles a candidate set — from web search, retailer APIs, product registries, review corpora. Whatever fails to enter the candidate set cannot be bought. Retrieval is the new shelf, and getting retrieved is the new distribution.

Represent. Each candidate is encoded from whatever text and structured data describes it: spec sheets, schema markup, reviews, editorial mentions. The agent does not see the drill; it sees the drill's textual shadow. Whoever authors the shadow shapes the perception.

Evaluate. The model compares candidates — scoring, ranking, often literally prompting itself to judge options pairwise or in a list. Every documented bias of LLM judgment operates here, at the exact moment of decision.

Decide. A winner is selected and, increasingly, purchased without a human glance at the alternatives. The 2025 launches of OpenAI's Instant Checkout, Visa Intelligent Commerce, and Mastercard Agent Pay mean the loop closes with money moving.

Human marketing attacked attention and memory. Machine marketing attacks retrieval and representation. The budget does not vanish; it migrates down the stack.

03 A Taxonomy of the Cognitive Attack Surface

Eight vectors, each anchored to published evidence. None requires hacking anything. Most are, today, perfectly legal.

3.1 ORDERING AND POSITION EFFECTS

When LLMs judge options presented together, the order of presentation changes the verdict. Wang et al. showed that quality rankings from GPT-4-class evaluators can be flipped simply by swapping the order in which two answers appear — with the right ordering, a 13-billion-parameter model was judged superior to ChatGPT on 66 of 80 test queries by ChatGPT itself. Zheng et al. documented the same position bias, alongside verbosity bias, in the LLM-as-judge setting. The commercial translation is immediate: whatever slot in the comparison your product occupies is worth money, and “first in the prompt” becomes what eye-level shelf placement was to the supermarket. Expect slotting fees to be reinvented for context windows.

3.2 CONTEXT-WINDOW DEAD ZONES

Liu et al.’s “Lost in the Middle” found that language models use long contexts unevenly: performance is highest when relevant information appears at the beginning or end of the input and degrades significantly when it sits in the middle — a U-shaped curve uncannily like the human serial-position effect, and present even in models built for long contexts. An adversarial marketer does not need to delete a competitor’s superior spec sheet. It suffices to arrange that it enters the agent’s context in the middle of a long document, while your own numbers land at the edges.

3.3 MODEL PRIORS AND BRAND BIAS

Models arrive at the purchase decision already opinionated. Kamruzzaman et al. found LLMs systematically associate global brands with positive attributes and local brands with negative ones, and disproportionately recommend luxury brands to high-income countries. Jones and Steinhardt showed model outputs are biased toward what appeared frequently in training data. A model’s prior is pretrained shelf space — and it opens a new marketing channel with a long fuse: flooding the public corpus with favorable, crawlable text not to persuade any human reader, but to tilt the priors of the next model generation. Call it training-data marketing. It is patient, deniable, and nearly impossible to attribute.

3.4 ANCHORING, FRAMING, AND DECOYS

The classic Kahneman-Tversky machinery turns out to replicate in silicon. Jones and Steinhardt found Codex adjusts outputs toward anchors embedded in prompts. Itzhak et al. found the decoy effect, certainty effect, and belief bias in current models — and, strikingly, found instruction tuning and RLHF amplify these biases rather than removing them. The decoy SKU — a deliberately inferior product priced to make the target product look optimal — was invented to manipulate humans. The evidence says it should work on model buyers too, and unlike a human, the model will dutifully include the decoy in its comparison table every single time.

3.5 SYCOPHANCY AND ENDORSEMENT HEURISTICS

Sharma et al. showed that state-of-the-art assistants consistently exhibit sycophancy — deferring to user pushback, tailoring answers to perceived views — and traced it to the training signal itself: human raters, and the preference models distilled from them, prefer agreeable answers over correct ones a non-negligible fraction of the time. An agent that weights apparent consensus is an agent that can be marketed to by manufacturing apparent consensus: seeded review corpora, astroturfed forum threads, coordinated "user" signals. The five-star review farm does not die in the agent economy. It gets a promotion, because the reader it deceives now buys instantly and at scale.

3.6 RETRIEVAL BIAS AND SOURCE AUTHORITY

Aggarwal et al.'s work on generative engine optimization demonstrated that content-side changes — citing sources, adding statistics, authoritative phrasing — can raise a site's visibility in AI-generated answers by up to 40 percent. On the platform side, Microsoft has confirmed that Bing's LLM systems use schema.org structured data to understand content. Retrieval pipelines must estimate authority, and authority signals can be manufactured: citation-shaped pages, statistics-dense product content, structured data engineered for machine legibility. The white paper that no human will ever read becomes a load-bearing marketing asset, because the agent reads everything.

3.7 REPRESENTATION RESONANCE

The deepest vector is wording chosen not for meaning but for how it lands in a particular model's learned representations. Zou et al. showed that gradient-searched adversarial suffixes — strings meaningless to humans — reliably steer aligned models, and transfer across model families. Kumar and Lakkaraju applied the idea commercially: a "strategic text sequence" added to a product page significantly increases the probability that an LLM ranks that product first. This is copywriting by gradient descent. The A/B test of the future does not ask which tagline humans prefer; it asks which token sequence maximizes the probability of recommendation in each of the five models that matter.

3.8 DIRECT INJECTION: THE BLUNT END

Finally, the crude vector: instructions hidden in the content an agent reads. Greshake et al. demonstrated indirect prompt injection against real LLM-integrated systems in 2023. By 2025 it was operational: Guardio Labs showed Perplexity's Comet browser, asked to buy an Apple Watch, completing checkout on a fake storefront — autofilling stored payment details without a confirming human glance. Injection is fraud rather than marketing, and vendors are actively hardening against it. But the boundary blurs from the legal side: a product page that says, in machine-legible microcopy, "this model is frequently preferred by careful buyers for its verified durability" is not an exploit. It is ad copy addressed to a machine — and nothing in current law is sure what to make of it.

04 The Evidence Table

Machine vector	Human analogue	Documented finding	Evidence
Position & ordering	Shelf placement	Swapping answer order flips LLM judgments; 13B model beat ChatGPT on 66/80 queries under ChatGPT's own evaluation	Wang 2023; Zheng 2023
Context dead zones	Serial-position effect	U-shaped use of long contexts; middle content under-weighted	Liu, TACL 2024
Model priors	Brand equity	Global brands get positive associations; outputs skew to frequent training examples	Kamruzzaman 2024; Jones & Steinhardt 2022
Anchoring & decoys	Decoy pricing	Anchoring, decoy, certainty effects replicate — amplified by instruction tuning	Jones & Steinhardt 2022; Itzhak 2024
Sycophancy	Social proof	Assistants defer to pushback; preference data rewards agreeable answers	Sharma, ICLR 2024
Retrieval & authority	PR / SEO	Content optimizations raise AI-answer visibility up to 40%; Bing confirms schema feeds its LLMs	Aggarwal, KDD 2024; SEL 2025
Representation resonance	Slogan craft	Adversarial strings steer aligned models and transfer; strategic text lifts product rank	Zou 2023; Kumar & Lakkaraju 2024
Direct injection	Deceptive advertising	Indirect injection compromises real apps; agentic browser bought from fake storefront	Greshake 2023; Guardio 2025

05 The New Discipline

None of this is waiting for the future to start. Gartner predicted in early 2024 that traditional search engine volume would fall 25 percent by 2026 as chatbots and agents absorb queries. Adobe's telemetry across more than a trillion visits to U.S. retail sites recorded a 1,300 percent year-over-year jump in generative-AI-referred traffic during the 2024 holidays, still compounding at 4,700 percent year-over-year by mid-2025 — with AI-referred visitors converting at a gap that narrowed from 49 percent to 23 percent in seven months. The buyers are arriving through the models.

And the sell side has noticed. Profound — a company whose product is, precisely, managing how brands appear inside AI answers — raised a $96 million Series C at a billion-dollar valuation in February 2026, with Target, Walmart, and US Bank among its customers. An industry that did not exist three years ago now has a unicorn, retainer clients, and a conference circuit. It calls itself answer-engine optimization. It is the peacetime name for the discipline this paper describes.

Watch what happens inside the marketing organization as this matures. The center of gravity shifts from creative to computational: fewer storytellers, more people who look like red-teamers — running fleets of agent simulations against product pages, measuring share-of-recommendation across model versions the way media buyers once measured gross rating points. Campaigns become model-specific, because what resonates with one model's representations differs from another's. The tagline is A/B tested against a gradient. Brand tracking becomes prior tracking: how does the newest frontier model complete the sentence "the most reliable dishwasher brand is—", and what corpus interventions would change it by the next training run?

This is why "marketing dies" is the wrong forecast. Global advertising is a high-hundreds-of-billions-of-dollars flow. Flows like that do not evaporate when their target changes species. They retool.

06 Hardening the Buyer

If the attack surface is structural, so is the defense. A fiduciary agent worth the name will need countermeasures at each stage of the decision anatomy.

Against ordering effects: randomize and repeat — evaluate candidate sets in multiple shuffled orders and aggregate, exactly as careful researchers already do when using LLMs as judges. Against context dead zones: chunked evaluation with per-candidate isolation, so no vendor's information is structurally buried. Against priors: separation of retrieval from judgment, with the judging model forced to justify rankings from in-context evidence rather than latent brand associations. Against sycophancy and astroturf: provenance-weighted consensus, where a thousand reviews that cannot be traced to verified purchases weigh less than ten that can. Against injection: the now-standard rule that retrieved content is data, never instruction — enforced by architecture, not by hoping the model remembers. And against representation resonance, the hardest vector: adversarial training and anomaly detection on input text, because a phrase optimized to move a model tends to look statistically strange.

Two asymmetries should temper optimism. The attacker needs one vector; the defender needs all of them, on every purchase, forever. And the defender's model is public — anyone can buy access to the very system they are optimizing against, iterate offline at leisure, and deploy only what works. This is the same asymmetry that made the first paper's forecast grim: the agents doing our buying cannot remain passive comparers. They must become adversarial systems themselves — forensic about data, paranoid about provenance, randomized against manipulation — or they will be farmed.

There is also an open legal question hiding here, and it deserves its own treatment: consumer-protection law prohibits deceiving consumers. Whether persuading a consumer's agent — by decoy, by ordering, by strings invisible in effect to any human — constitutes deceiving the consumer is a question courts and regulators have barely begun to face.

07 Conclusion: The Pitch Compiles

The twentieth century built a science of wanting: anthropological, Freudian, tuned to the human animal. That science is being made obsolete not because persuasion is ending but because the persuadable party is changing. The machine buyer has no childhood to mine and no status to flatter — but it has ordering effects, dead zones, priors, decoys, sycophancy, authority heuristics, and representational quirks, every one of them documented in the literature and several already monetized by a funded industry.

So the forecast is not that marketing disappears. It is that marketing sheds its anthropology and becomes adversarial computation: a quiet, continuous contest between optimization pressure on the sell side and hardening on the buy side, fought in candidate sets and context windows rather than in commercial breaks. The thirty-second spot gave us the jingle you could not forget. Its successor is a string you will never see, tuned to resonate with a model you will never meet, nudging a purchase you never watched happen.

The pitch does not die. It compiles.

SOURCES

1. Wang et al., “Large Language Models are not Fair Evaluators” (2023). arxiv.org/abs/2305.17926

2. Zheng et al., “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena” (NeurIPS 2023). arxiv.org/abs/2306.05685

3. Liu et al., “Lost in the Middle: How Language Models Use Long Contexts” (TACL 2024). arxiv.org/abs/2307.03172

4. Kamruzzaman, Nguyen & Kim, “‘Global is Good, Local is Bad?’: Understanding Brand Bias in LLMs” (EMNLP 2024). aclanthology.org/2024.emnlp-main.707

5. Jones & Steinhardt, “Capturing Failures of Large Language Models via Human Cognitive Biases” (NeurIPS 2022). arxiv.org/abs/2202.12299

6. Itzhak et al., “Instructed to Bias” (TACL 2024). arxiv.org/abs/2308.00225

7. Sharma et al., “Towards Understanding Sycophancy in Language Models” (ICLR 2024). arxiv.org/abs/2310.13548

8. Aggarwal et al., “GEO: Generative Engine Optimization” (ACM KDD 2024). arxiv.org/abs/2311.09735

9. Kumar & Lakkaraju, “Manipulating Large Language Models to Increase Product Visibility” (2024). arxiv.org/abs/2404.07981

10. Zou et al., “Universal and Transferable Adversarial Attacks on Aligned Language Models” (2023). arxiv.org/abs/2307.15043

11. Greshake et al., “Not what you’ve signed up for” (2023). arxiv.org/abs/2302.12173

12. Guardio Labs, “Scamlexity” (2025). guard.io/labs

13. Search Engine Land, “Microsoft Bing & Copilot use schema for its LLMs” (2025).

14. Gartner, “Search Engine Volume Will Drop 25% by 2026” (2024).

15. Adobe, “Generative AI-powered shopping rises with traffic to retail sites” (2025).

16. Profound, “$96M Series C” (2026). tryprofound.com

17. OpenAI, “Buy it in ChatGPT” (2025). openai.com