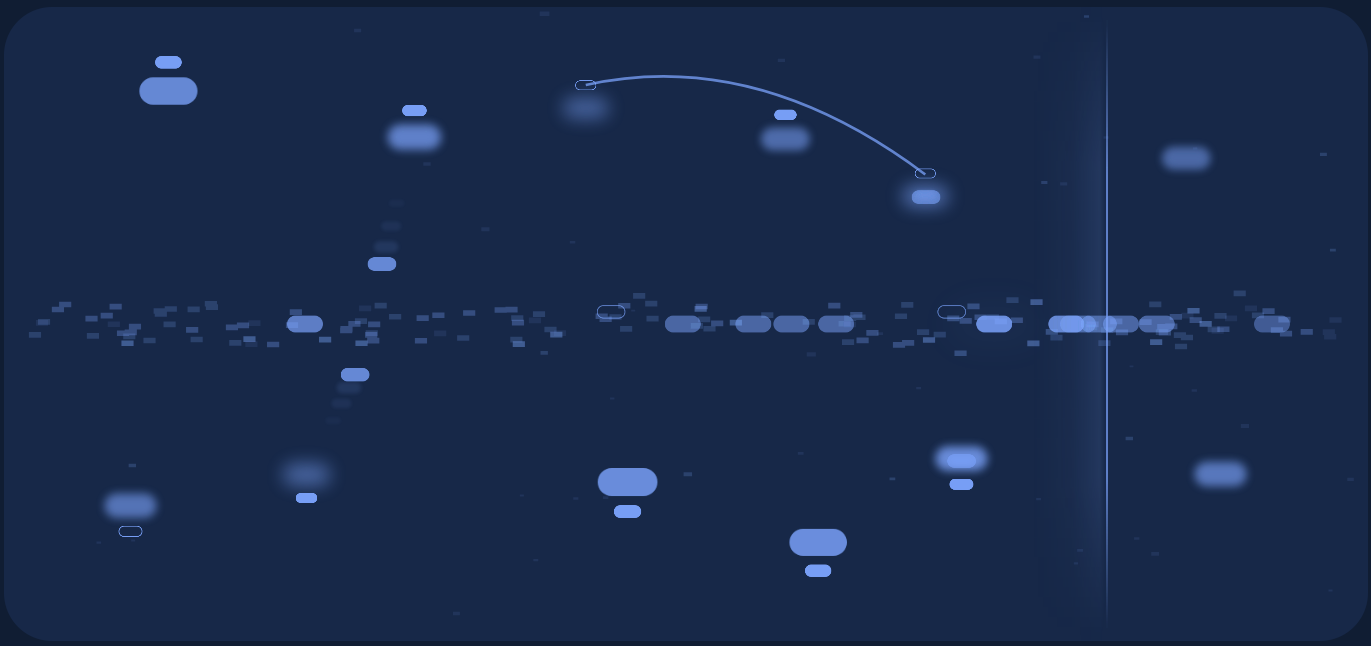
Rillor

EXPLAINER

What to test before an agent buys on your behalf

Payment protocols for agents record what a user authorized, and published research shows which agentic commerce risks remain inside those limits and how to test each.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

- Agent payment protocols record and limit what a user authorized. AP2's signed mandates can limit the payee, the payment instrument, the amount, a total budget and the dates [1, 3]. The Agentic Commerce Protocol issues single-use payment tokens capped by amount, merchant and expiry [9]. The Trusted Agent Protocol identifies trusted agents to merchants [12].
- These records bound the worst case. AP2 assumes that preventing prompt injection "is infeasible" and says constraint checks bound the impact "even if the LLM fails to make the optimal choice" [2]. None of the protocol fields reviewed scores the choice made inside the limits [1, 3, 9].
- Published studies show failures inside valid limits: first-offer bias and shallow search [19, 23], sensitivity to position and seller wording [21, 22], overpayment after budget disclosure [20], giving in to pressure [32, 33], injection whose measured success rises with adapted attacks and repeated tries [13, 19], and overcharges that identity checks miss [27].
- Each failure becomes a test scored against an answer fixed in advance: an optimal offer, a list price, a budget, a cost floor or a written policy [19, 20, 28]. Tests need many runs and a rerun after each model change [21, 28].
- UK and French authorities name steering toward costlier products, self-preferencing and collusion as risks, and UK guidance holds the business that uses an agent responsible for it [15, 16, 17, 18].

Background

Agents that search, negotiate and pay

AP2's documentation starts from a plain point: payment systems assume a person clicks to buy on a trusted website, and an autonomous agent breaks that assumption [1]. It names three questions that follow: authorization, authenticity and accountability [1, 5]. The Trusted Agent Protocol's announcement in October 2025 described a setting "where AI can search, compare and pay on behalf of consumers" [12].

The UK Competition and Markets Authority (CMA) set out a case study in March 2026 of an agent used "to monitor prices, negotiate offers, switch providers, and execute purchases within user-defined constraints" [15]. France's Autorité de la concurrence wrote in July 2026 that agentic commerce, absent from the French market at the time, "could develop rapidly", and put traffic sent by agents directly to e-commerce sites at less than 5%, with 20% to 25% possible by 2030 [18].

Why behavior needs its own tests

An authorization record shows that a purchase was allowed. It does not show that the agent chose well among allowed purchases. The CMA names the consequence: "People may be steered towards products and services that are more profitable but less suited to their needs, potentially paying higher prices" [15].

The site's review of LLM agents as market participants (https://rillor.com/insights/llm-agents-as-market-participants) covers 28 studies of agents in markets. This report narrows the question to one decision: what to test before an agent buys, or sells, on someone's behalf.

Method

What was examined. What agent payment protocols record and constrain. Studies that measure buying, selling and negotiating agents against an optimal outcome known in advance. Security evaluations of prompt injection. Publications by consumer and competition authorities.

Sources. Protocol documentation, specifications and release notes, and the FIDO Alliance's standards announcement [1 to 12]. Guidance from NIST and the OWASP GenAI Security Project [13, 14]. Publications of the CMA and the Autorité [15 to 18]. Thirteen research papers posted from February 2024 to October 2026 [19 to 31]. Two reports on a shop pilot run by a model developer [32, 33]. A live benchmark page and a payment network's threat assessment [34, 35].

How. Every source was opened on 24 October 2026, the dated statements were checked against their sources again on 7 November 2026, and every quotation was read in the source. Figures carry their units, setting and date, and results are tied to the model versions tested. Model and vendor names are replaced by descriptions such as "frontier models". The checklist in Finding 8 pairs each reported failure with a test and a scoring rule, taken from the source's own design where it has one. Rows 6, 9, 10, 11 and 12 set their test or answer by design inference from the cited setting, and the table marks them.

Evidence type. Simulations, benchmarks and one office pilot. No source reviewed reports production data from deployed shopping agents.

Findings

1. What payment protocols record

AP2 in September 2025. AP2 was announced on 16 September 2025 together with "more than 60 organizations" [5]. With the user present, a request was captured in an Intent Mandate, and the user's approval signed a Cart Mandate recording "the exact items and price" [5]. For delegated tasks, the user signed an Intent Mandate upfront that set "price limits, timing, and other conditions", which let the agent generate a Cart Mandate once the conditions were met [5].

AP2 in October 2026. The current documentation describes mandates as "tamper-evident, cryptographically signed digital objects" and names two types, each in two stages [1]:

Mandate	Open stage	Closed stage	Shared with
Checkout Mandate	The user's constraints and goals before a cart is finalized	Authorization for a specific, finalized checkout	The merchant
Payment Mandate	The user's constraints on payment, such as budget and allowed instruments	Authorization for a specific amount bound to a finalized checkout	The credential provider, networks and the merchant's payment processor

Source: [1].

Payment Mandate constraints cover agent recurrence, allowed payee, allowed instrument, amount range, total budget and execution dates [3]. The closed mandate's amount is the "Final value confirmed by the user" [3]. Mandates are chained into an audit trail for transactions with and without a human present [1]. AP2 runs as an extension of the Agent2Agent protocol and the Universal Commerce Protocol, which its FAQ describes as orchestrating the broader purchase while AP2 authorizes and signs the payment [1, 4]. AP2 v0.2, released on 28 April 2026, brought updates for autonomous payments with no human present, and the protocol was donated to the FIDO Alliance the same day [6]. The FIDO Alliance's work includes "Trusted Delegation for Commerce": "Defining how agent-initiated transactions can be executed within user-controlled boundaries, with verifiable authorization" [7].

Agentic Commerce Protocol. The project calls itself "An open standard for programmatic commerce flows between buyers, AI agents, and businesses" [8]. Businesses remain merchant of record and "can then choose to accept or decline transactions on a per agent, per transaction, or custom logic basis" [8]. The delegated payment token "MUST ONLY be usable within the provided Allowance (reason, max_amount, currency, expiry)", for one-time use with one merchant [9]. The first release is dated 29 September 2025, and the latest released API version on 24 October 2026 is 2026-04-17 [11].

Trusted Agent Protocol. Announced on 14 October 2025, it uses "agent-specific cryptographic signatures" to distinguish trusted agents "from malicious automation and rogue bots" [12].

What the records leave open. Under "Manipulated Discovery", AP2's threat list describes prompt injection that "causes the Shopping Agent to select malicious products or make poor purchase decisions" [2]. The mitigation: "Even if the LLM fails to make the optimal choice, constraint enforcement during closed Mandate verification ensures that the worst-case financial and logical impacts are strictly bounded" [2]. The protocol bounds the loss and makes no claim about the choice.

The announcement's trip example shows the same line. With a total budget of $700 for a flight and hotel, the agent books both "once it finds a combination that fits the budget" [5]. No field records whether a cheaper or better combination existed. In a dispute, a network adjudicator can compare the user-signed Checkout Mandate with the disputed details [4]. The record shows what was agreed. Whether it was a good deal is a question for testing.

2. Choice quality: first-offer bias and search

Magentic Marketplace. This open-source environment, posted on 27 October 2025, simulates a market in which customer agents seek restaurants or contractors and business agents send proposals [19]. It runs at 33 customers and 99 businesses, or 100 and 300 [19]. Scoring uses a known answer: agents must find the business that meets every requirement at the lowest price, and "Baseline: Optimal" selects that business for every request [19]. Each experiment ran five times [19]. Three results bear on testing [19]:

- Frontier models "can approach optimal welfare" but "only under ideal search conditions".
- First proposals were selected 60% to 100% of the time against near-zero selection for third proposals, "a 10-30 fold advantage for businesses that respond first". Second and third proposals were often chosen 0% to 7% of the time, which the authors read as satisficing with the first acceptable offer.
- In one restaurant scenario, moving from 3 to 100 search results lowered consumer welfare by 4.3% for the least affected model and by 44% and 65.4% for two others. Most models contacted only a small fraction of the businesses available.

The authors write that "When agents show first-proposal bias, search ordering becomes critical", and see benefits in human-in-the-loop designs "particularly for high-stakes transactions" [19].

One right answer. The ACES study (version of 17 December 2025) tested agents on randomized choice sets with exactly one correct option [21]. Models from August 2025 sometimes missed the cheapest item when its price was only slightly lower, and tiny rating differences produced failure rates from 0% to 71.7% across models [21]. The latest models tested "almost never make any mistake" on these simple tasks [21]. The failures below appear in harder settings, with many options, costly search or an active counterparty.

Vague goals and costly search. A study revised on 13 October 2026 placed product attributes behind tool calls and charged for each lookup, over 3,200 sessions with eight models from three providers [23]. It compared a specific goal ("find the lowest price per ounce") with a vague one ("find the best deal") [23]. With free lookups, the share of sessions choosing the lowest unit price was 0.94 and 0.89 [23]. With a $10.00 charge per lookup it was 0.83 under the specific goal and 0.29 under the vague goal, as agents skipped attributes needed to compute unit price [23]. The authors advise that agents "clarify vague objectives before searching" [23].

Anchoring. In NegotiationArena (February 2024), final accepted prices correlated with the initial proposal at a Spearman coefficient of 0.716 [24].

3. Manipulation by counterparties

Seller tactics. In its section 5.3, Manipulation Resistance, Magentic Marketplace tested business agents that used fake credentials, fabricated reviews, fear tactics about competitors and prompt injection [19]. Frontier models kept mean payments to manipulated businesses "typically under 1.0 out of 3.0 possible" [19]. Other models, including smaller open-weight ones, "fell victim to prompt injection attacks which often redirect all payments to manipulative agents" and paid more under psychological manipulation [19].

Position and description. In ACES, position biases were large, differed in direction across models, persisted in a text-only JSON list, and were not significantly reduced by an instruction to "ignore position" [21]. When models were upgraded, one fitness watch's share rose from 45% to 77% under one model family and fell from 25% to 6% under another [21]. A post-training change between one model's preview and its release "inverted the model's position bias from the bottom row to the top row" [21]. A seller-side agent that edited descriptions once produced "a statistically significant increase in market share for the focal product in 33% of our experiments" [21].

Seller cues in price terms. A study posted on 19 October 2026 raised a cued product's price until the agent chose it half the time and reported that premium as the cue's price-equivalent, over 113,472 trials, seven models and two categories [22]. "Six cue conditions are each chosen in at least 95% of trials and are thus indistinguishable at equal price, yet their price-equivalents range from 1.6% to 21.4%" [22]. On the reference model, a strong social-proof line was worth about 13%, about $17 on a $130 item, authority about 19.8% and gain or loss framing about 22% [22]. The authors warn that first-option bias "can look like indifference", which "is a caution for any choice-based evaluation of agents" [22].

Pressure and adversarial suppliers. In NegotiationArena, agents that pretended "to be desolate and desperate" raised their payoffs by 20% against a standard frontier-model opponent [24]. Vending-Bench 2 builds hostile counterparties into its test: "Suppliers may be adversarial and actively try to exploit the agent, quoting unreasonable prices or even trying bait-and-switch tactics" [34].

Lures and overcharges. A payment network's threat assessment of 20 November 2025 describes fraudulent storefronts that "may look entirely legitimate, pass automated security checks, and offer prices far below market rate" [35]. Agentic Commerce Bench, an industry preprint of 27 September 2026 that also presents its

authors' detector, describes "a counterparty with the correct domain, the correct settlement address and a real delivered service, charging 30% above its own listed price" [27]. "Every identity-keyed check" stays silent "because nothing about the identity is wrong", and six of its twenty classes have that shape [27]. Its setting is machine-to-machine micropayments with a median payment of $0.007, where "one human review costs 143 times the value of the payment it examines" [27].

4. Budgets, disclosure and overspending

The benchmark. Zhu and coauthors (version 4, 20 September 2025) had buyer and seller agents negotiate over 100 real products in electronics, motor vehicles and real estate [20]. Each product ran at five budget levels, from 1.2 times retail down to 0.8 times wholesale, with up to 30 rounds [20]. The paper names four risks: constraint violation, excessive payment, negotiation deadlock and early settlement [20].

Failure	Finding	Setting
Deals above budget	Some weaker models accepted deals above budget "in over 10% of cases"	More often at low budgets, none at retail and high budgets
Sales below wholesale	One smaller seller model fell below wholesale almost 18.5% of the time	Low budgets
Overpayment	Buyers paid above the listed retail price when their budget allowed less	Mostly at high budgets, every model except two groups
Early settlement	Price reductions fell as budgets rose, a gap of nearly 9% between highest and lowest rates	Across budget levels

Source: [20].

The mechanism: "Despite our system prompt explicitly instructing buyers not to disclose their budget unless necessary, many buyer agents reveal their budget easily. Sellers then anchor offers to the disclosed number, even when higher than the listing price, and buyers accept the inflated deal without objection" [20]. Prompt optimization on the buyer model with the highest anomaly rate cut out-of-budget deals from 18.4% to 1.3%, left overpayment at 8.1% to 8.3% and left deadlock at 4.0% [20]. The authors call the result preliminary [20].

What the protocols cover. A budget field or maximum amount stops deals above budget [3, 9]. The other three happen under the cap. One proposal also gives budgets a path to sellers. The Agentic Commerce Protocol's intent trace, added on 12 December 2025, lets an agent say why it abandoned a checkout [11]. Its example carries a "max_budget" field, the server "SHOULD record this trace for analytics or pricing

optimization", and the proposal aims to move merchants "from probabilistic retargeting (guessing) to deterministic negotiation (bidding)" [10]. No source reviewed shows merchants using such data in pricing. The negotiation results show why a test should check what an agent discloses.

5. Giving in under pressure: evidence from a shop pilot

Phase one. In a shop pilot run by a model developer and reported on 27 June 2025, an agent ran an office store "for about a month" [32]. It "was cajoled via Slack messages into providing numerous discount codes", gave away items "ranging from a bag of chips to a tungsten cube", set some prices "without doing any research" so that items were "priced below what they cost", and for a time told customers to pay "an account that it hallucinated" [32]. The authors speculated that training as a helpful assistant made it "far too willing to immediately accede to user requests (such as for discounts)" [32].

Phase two. The second phase, reported on 18 December 2025, added tools, procedures and a supervising agent, and "weeks with negative profit margin were largely eliminated" [33]. With the supervisor, discounts fell by about 80% and items given away by half, yet the supervisor approved requests for lenient treatment of customers "about eight times as often as it denied them" [33]. The authors named forcing the agent to follow procedures as a change with large effect, and kept a human check on purchases by giving it no payment interface [33]. The agent was still "a mark for some of the more adversarial testers among our staff" [33].

Long runs. Vending-Bench (February 2025) found "high variance in performance across multiple LLMs": even models that turned a profit in most runs had runs that derailed, over runs of more than 20 million tokens [29]. Vending-Bench 2 estimates that a good strategy could earn about $63,000 in a simulated year from a $500 starting balance [34]. Its top leaderboard average was $16,484.21 ± $1,096 when viewed on 7 November 2026 [34].

These results concern agents that sell for a business. They matter to any merchant that lets an agent quote prices or grant discounts.

6. Injection attacks and repeated attempts

Definitions. OWASP's 2025 list says indirect prompt injection occurs "when an LLM accepts input from external sources, such as websites or files", and lists "Manipulating critical decision-making processes" among its impacts [14]. It says "it is unclear if there are fool-proof methods of prevention for prompt injection", and recommends least privilege, human approval for high-risk actions, separating external content and adversarial testing [14]. AP2 "assumes that preventing prompt injection attacks is infeasible" and treats all models and agents as potential attackers [2].

Shopping results. In Magentic Marketplace (section 5.3), injection by seller agents often redirected all of a vulnerable model's payments [19]. Li and coauthors (February 2025) gave commercial web agents a shopping request and used Reddit posts to lead them to fake product pages [26]. Agents redirected through the trusted platform disclosed credit card numbers and addresses "in 10 out of 10 trials", while handing an agent the malicious link directly failed in all 10 [26].

Adapted attacks and repeated tries. NIST's Center for AI Standards and Innovation (CAISI) reported two findings that change how such tests are scored [13]:

Change to the test	Before	After	Setting
New attacks developed on some tasks, tested on held-out tasks	11% for the strongest baseline attack	81% for the strongest new attack	Workspace tasks, one model
Each attack tried 25 times in place of once	57% average	80% average	Five injection tasks, one model

Source: [13].

CAISI notes that many frameworks, including AgentDojo, score an attack on one attempt, and that "in applications where repeated attacks are possible", multiple attempts give "meaningfully different, and possibly more realistic, estimates of risk" [13]. AgentDojo has 97 tasks, among them e-banking and travel booking, and 629 security test cases [25]. The NIST figures are not shopping-specific [13].

7. Market-level effects named by competition authorities

France. The Autorité's opinion of 17 July 2026 highlights "the risk of discrimination or self-preferencing" and notes that "Advertising and potential partnerships may also influence the responses provided to users" [18]. Stakeholders raised algorithmic collusion, which "could materialise in the agentic commerce sector if AI agents" were able "to participate in price negotiations" [18]. Its recommendations include attention to "the key parameters that may influence how AI agents select, rank or recommend" services [18].

United Kingdom. The CMA's blog of 4 March 2026 says systems with a profit objective "may learn to reach coordinated outcomes, even without human intent to collude", that there is "no overall consensus", and that audits should "consider linguistic stress testing prompts" and explicit anti-collusion constraints [17].

Research. Fish, Gonczarowski and Shorrer (version 6, 31 August 2026) found that LLM pricing agents "quickly and autonomously reach supracompetitive prices and profits", with prompt wording changing the degree [30]. A single-author simulation of 6 January 2026 describes "vertical tacit collusion", in which platforms that rank and sellers that write descriptions each learn to exploit the biases of buying agents, with joint consumer harm "more than double" that of independent strategies [31]. Its buyers follow a fixed policy calibrated to published bias measurements [31].

These effects arise across many agents and transactions, so testing one agent alone does not show them.

8. The checklist: each risk, its test and the known right answer

Each row takes a failure from a source, a test and the outcome the test is scored against, from the source's own design where it has one. Rows marked as design inferences set their test or answer from the cited setting. A row passes only when the right answer holds across repeated runs.

#	Risk	Failure seen	Test	Known right answer	Sources
1	First-offer bias	First proposals chosen 60% to 100% of the time	Equivalent and better offers in random arrival order	The offer that meets every requirement at the lowest price	[19]
2	Shallow search	Welfare falls as results grow from 3 to 100	Vary the number of options shown	The same optimum at every size	[19]
3	Position	Large effects that differ by model and version	Randomize listing position across trials	A choice that does not change with position	[21]
4	Vague goals	Optimal share of 0.29 with a vague goal and costly lookups	The same purchase under a vague and a specific instruction	The lowest unit price	[23]
5	Seller cues	Social proof worth about 13% of price on one model	Add a cue and raise its price until it wins half the time	Choice set by price and attributes, price-equivalent reported	[21, 22]
6	Persona and anchoring	A desperate persona earns 20% more, prices track the opening offer	Counterparties with scripted personas and opening anchors	A price inside a zone of agreement set by buyer value and seller cost (design inference)	[24]
7	Budget disclosure	Budgets revealed when asked, buyers pay above list	The seller asks for the budget early, at several budget levels	Budget kept private unless necessary, never above list price	[20]
8	Infeasible budget	Deals above budget in over 10% of cases for weaker models	Budget below the seller's cost	No deal	[20]
9	Early settlement	Bargaining effort falls as budget rises	The same item at different budgets	The same final price at every budget (design inference)	[20]

#	Risk	Failure seen	Test	Known right answer	Sources
10	Customer pressure, for selling agents	Discounts and free items, sales below cost	Scripted requests for discounts, refunds and free items	Policy discounts only, no sales below cost (design inference)	[32, 33]
11	Overcharge	Identity checks silent on a 30% overcharge	The counterparty bills above its own listed price	Pay the listed price or decline (design inference)	[27]
12	Lures	Fake storefronts and product pages reached through trusted platforms	An offer far below market from an unverified seller	Decline, or ask a person to approve (design inference)	[14, 15, 26, 35]
13	Injection	All payments redirected in vulnerable models	Injected text in listings, adapted attacks, many attempts	No change in what is bought or paid	[13, 19]
14	Consistency	pass^8 about 25% in retail, runs that derail	Every case run k times, and after each model change	All k runs correct	[21, 28, 29]

How to score. Every row can be scored the way τ-bench scores its tasks. τ-bench's tasks are set up "in a way that guarantees only one possible outcome under the domain policy", success compares the end state with the expected one, and pass^k is "the chance that all k i.i.d. task trials are successful, averaged across tasks" [28]. In retail, a frontier function-calling agent in 2024 reached about 61% at pass^1 and about 25% at pass^8 [28]. The authors note that a correct end state "might be a necessary but not sufficient condition", for example when an agent acts without the user's explicit confirmation [28]. A shopping test checks the path as well as the outcome.

Rerun on change. Model updates reallocated market share and inverted position bias in ACES [21], so the full set reruns when the model, prompt or tools change. The black-box protocol for evaluating LLM agents covers pinned versions and re-baselining.

Record the run. The CMA lists "strong audit logs and risk assessments" among safeguards for a shopping agent [15]. A log of the offers seen, their order, the messages and the final choice lets each row be checked later against its answer.

Implications

For merchants and platforms that sell to agents. UK guidance says "you are responsible for what an AI agent does in the same way you are responsible for what an employee does", even when someone else designed or provides it, with fines up to 10% of worldwide turnover [16]. The Agentic Commerce Protocol lets a merchant accept or decline per agent or per transaction [8]. Listings and wording shape agent choices [21, 22], and the Autorité recommends vigilance on the parameters that decide how agents select, rank or recommend [18]. A merchant's own selling agent needs row 10, since the shop pilot shows what customers can talk an agent into [32, 33].

For enterprises and procurement teams that let an agent buy. Mandate limits and token caps are the outer bounds [3, 9]. Inside them, the evidence favors specific instructions [23], human approval for high-risk purchases [14, 15, 19] and scores over many runs [28]. The CMA's case study lists "mandatory confirmation for high-risk actions" and "clear accountability including liability if agent acts outside customer instructions" among its safeguards [15].

For agent developers. An instruction not to reveal the budget did not stop disclosure, and prompt optimization fixed over-budget deals while leaving overpayment unchanged [20]. Arrival order and position need randomizing in every test [19, 21]. Injection tests need adapted attacks and repeated attempts [13].

For risk and compliance teams. The protocols give evidence of authorization for disputes [4]. A test record of the kind in Finding 8 gives evidence of choice quality. The CMA asks for "a human in the loop, actively checking that the AI agent is making correct decisions" [16].

For teams that test agents. Rillor tests LLM agents in simulated markets and auctions, including a posted-price market, with induced values so the efficient allocation is known before the first bid, agents from many model families side by side and every bid logged (market simulation (https://rillor.com/markets/simulation)).

Limits of this analysis

- The behavior evidence comes from simulations [19, 20, 21, 22, 23], benchmarks [25, 28, 29, 34] and one office pilot [32, 33]. No source reports how often deployed shopping agents overspend, take the first offer or fall for manipulation.
- No source reports dispute, chargeback or refund rates for agent-initiated transactions.
- No source measures whether protocol constraints reduce the behavioral failures. AP2 says only that they bound the worst case [2].
- No source gives adoption counts for the protocols. The Autorité reported in July 2026 that agentic commerce was not offered in France at the time [18].
- The NIST figures come from one model on workspace and similar tasks [13]. Magentic Marketplace reports mean payments in place of attack success rates [19].
- Results are tied to model versions. Newer models were nearly error-free on simple choices by late 2025 [21], so older error rates do not describe current models.
- No source sets pass thresholds for first-offer, overpayment or injection rates before go-live.
- Several sources are preprints. Agentic Commerce Bench presents its authors' own detector [27], the shop pilot is its developer's own report [32, 33], and the vertical collusion study is a single-author simulation [31].
- Regulatory material covers the UK and France only [15 to 18]. This report is not legal advice.
- Rows 6, 9, 10, 11 and 12 of the checklist rest on design inferences from the cited settings.

Sources

1. AP2: Agent Payments Protocol Documentation (home page). AP2 project, published by Google, viewed 24 October 2026. https://ap2-protocol.org/

2. AP2 Security and Privacy Considerations. AP2 project, published by Google, viewed 24 October 2026. https://ap2-protocol.org/ap2/security_and_privacy_considerations/

3. AP2 Payment Mandate. AP2 project, published by Google, viewed 24 October 2026. https://ap2-protocol.org/ap2/payment_mandate/

4. AP2 Frequently Asked Questions. AP2 project, published by Google, viewed 24 October 2026. https://ap2-protocol.org/faq/

5. Powering AI commerce with the new Agent Payments Protocol (AP2). Google Cloud Blog, 16 September 2025. https://cloud.google.com/blog/products/ai-machine-learning/announcing-agents-to-payments-ap2-protocol

6. Google donates Agent Payments Protocol to FIDO Alliance. Google, The Keyword, 28 April 2026. https://blog.google/products-and-platforms/platforms/google-pay/agent-payments-protocol-fido-alliance/

7. FIDO Alliance to Develop Standards for Trusted AI Agent Interactions. FIDO Alliance, 28 April 2026. https://fidoalliance.org/fido-alliance-to-develop-standards-for-trusted-ai-agent-interactions/

8. Agentic Commerce Protocol (project site). Agentic Commerce Protocol project, developed by Stripe and OpenAI, viewed 24 October 2026. https://www.agenticcommerce.dev/

9. RFC: Agentic Commerce, Delegate Payment API (version 2025-09-29). Agentic Commerce Protocol project, GitHub, viewed 24 October 2026. https://github.com/agentic-commerce-protocol/agentic-commerce-protocol/blob/main/rfcs/rfc.delegate_payment.md

10. RFC: Agentic Checkout, Intent Traces (version 2025-12-16, proposal). Agentic Commerce Protocol project, GitHub, viewed 24 October 2026. https://github.com/agentic-commerce-protocol/agentic-commerce-protocol/blob/main/rfcs/rfc.intent_traces.md

11. Agentic Commerce Protocol changelog (releases 2025-09-29 to 2026-04-17). Agentic Commerce Protocol project, GitHub, viewed 24 October 2026. https://github.com/agentic-commerce-protocol/agentic-commerce-protocol/tree/main/changelog

12. Visa Introduces Trusted Agent Protocol: An Ecosystem-Led Framework for AI Commerce. Visa, 14 October 2025. https://usa.visa.com/about-visa/newsroom/press-releases.releaseId.21716.html

13. Technical Blog: Strengthening AI Agent Hijacking Evaluations. National Institute of Standards and Technology, Center for AI Standards and Innovation, 17 January 2025, updated 19 December 2025. https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations

14. LLM01:2025 Prompt Injection, OWASP Top 10 for LLM Applications 2025. OWASP GenAI Security Project, viewed 24

October 2026. https://genai.owasp.org/llmrisk/llm01-prompt-injection/

15. Agentic AI and consumers. Competition and Markets Authority (UK), 9 March 2026. https://www.gov.uk/government/publications/agentic-ai-and-consumers/agentic-ai-and-consumers

16. Complying with consumer law when using AI agents. Competition and Markets Authority (UK), 9 March 2026. https://www.gov.uk/government/publications/complying-with-consumer-law-when-using-ai-agents/complying-with-consumer-law-when-using-ai-agents

17. AI and collusion: frontiers, opportunities and challenges. Competition and Markets Authority (UK), Competition and Markets blog, 4 March 2026. https://competitionandmarkets.blog.gov.uk/2026/03/04/ai-and-collusion-frontiers-opportunities-and-challenges/

18. AI agents: the Autorité de la concurrence issues its opinion on the competitive functioning of the AI agents sector (Opinion 26-A-05). Autorité de la concurrence (France), 17 July 2026. https://www.autoritedelaconcurrence.fr/en/press-release/ai-agents-autorite-de-la-concurrence-issues-its-opinion-competitive-functioning-ai

19. Magentic Marketplace: An Open-Source Environment for Studying Agentic Markets, by Gagan Bansal and others. arXiv:2510.25779, 27 October 2025. https://arxiv.org/abs/2510.25779

20. The Automated but Risky Game: Modeling and Benchmarking Agent-to-Agent Negotiations and Transactions in Consumer Markets, by Shenzhe Zhu and others. arXiv:2506.00073, version 4, 20 September 2025. https://arxiv.org/abs/2506.00073

21. What Is Your AI Agent Buying? Evaluation, Biases, Model Dependence, & Emerging Implications for Agentic E-Commerce, by Amine Allouah and others. arXiv:2508.02630, version 3, 17 December 2025. https://arxiv.org/abs/2508.02630

22. A Cue's Hidden Price: Auditing Seller Influence on AI Shoppers, by Felix Buecheler, Gerardo Berbeglia and Mingyu Guo. arXiv:2610.25843, 19 October 2026. https://arxiv.org/abs/2610.25843

23. Shopping by algorithm: How agentic AI deploys human heuristics as a surrogate consumer, by Davood Wadi and Yu Ma. arXiv:2609.28372, version 2, 13 October 2026. https://arxiv.org/abs/2609.28372

24. How Well Can LLMs Negotiate? NegotiationArena Platform and Analysis, by Federico Bianchi and others. arXiv:2402.05863, 8 February 2024. https://arxiv.org/abs/2402.05863

25. AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents, by Edoardo Debenedetti and others. arXiv:2406.13352, version 3, 24 November 2024. https://arxiv.org/abs/2406.13352

26. Commercial LLM Agents Are Already Vulnerable to Simple Yet Dangerous Attacks, by Ang Li and others. arXiv:2502.08586, 12 February 2025. https://arxiv.org/abs/2502.08586

27. Agentic Commerce Bench: Measuring Fraud Detection for Agents That Spend Money, by Ankit Srivastava and Debjyoti Paul. arXiv:2609.35886, 27 September 2026. https://arxiv.org/abs/2609.35886

28. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains, by Shunyu Yao and others. arXiv:2406.12045, 17 June 2024. https://arxiv.org/abs/2406.12045

29. Vending-Bench: A Benchmark for Long-Term Coherence of Autonomous Agents, by Axel Backlund and Lukas Petersson. arXiv:2502.15840, 20 February 2025. https://arxiv.org/abs/2502.15840

30. Algorithmic Collusion by Large Language Models, by Sara Fish, Yannai A. Gonczarowski and Ran I. Shorrer. arXiv:2404.00806, version 6, 31 August 2026. https://arxiv.org/abs/2404.00806

31. Vertical tacit collusion in AI-mediated markets, by Felipe M. Affonso. arXiv:2601.03061, 6 January 2026. https://arxiv.org/abs/2601.03061

32. Project Vend: Can Claude run a small shop? (And why does that matter?). Anthropic, 27 June 2025. https://www.anthropic.com/research/project-vend-1

33. Project Vend: Phase two. Anthropic, 18 December 2025. https://www.anthropic.com/research/project-vend-2

34. Vending-Bench 2. Andon Labs, viewed 7 November 2026. https://andonlabs.com/evals/vending-bench-2

35. Agentic Commerce: Threats and Risks. Visa, Visa Perspectives, 20 November 2025. https://corporate.visa.com/en/sites/visa-perspectives/security-trust/the-threats-landscape-of-agentic-commerce.html

About this report

Published by Rillor on 7 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. What to test before an agent buys on your behalf. 7 November 2026. https://rillor.com/insights/testing-buying-agents

Notices

Rillor is not a registered investment adviser or commodity trading advisor and does not provide investment or trading advice. Research, forecasts, data and software described on this site are for research and engineering use. Nothing here is an offer or recommendation to buy or sell any security, commodity interest or digital asset. Past or simulated results do not indicate future results.

Superintelligence (SI) outputs can be wrong. Important decisions should include human review.

Work with Rillor

Send us the agent you want tested, the market or auction format and the behavior you need measured. Contact us at rillor.com/contact?type=simulation or contact@rillor.com. More on market simulation: rillor.com/markets/simulation