

AIVO MERIDIAN · WORKING PAPER 2026-09

# The Agentic Shelf

*A measurement framework for autonomous AI shopping journeys — and why current commercial measurement misses what matters most.*

---

Date of record · 6 June 2026

Classification · Public

Framework version · AIVO Agentic Shelf v1.0

Authoring entity · AIVO Meridian, London

# A new commerce surface is forming.

We call it the **Agentic Shelf**: the surface on which autonomous AI buyers — frontier-model shopping agents, retailer-owned assistants, and third-party orchestration tools — select products on a consumer's behalf, with little or no human intervention. Within twelve months the Agentic Shelf will be a measurable distribution channel in major consumer categories. Within thirty-six months it will be the dominant first-touch surface for AI-mediated discovery in beauty, electronics and grocery.

Today no major brand can answer the simple question: *how do we perform on the Agentic Shelf?* No major retailer building its own shopping agent can answer the parallel question: *does our agent perform correctly for our customers?* No frontier-model platform exposing a shopping experience can defensibly demonstrate that its agent honours the constraints its users state.

This paper makes four claims, with measurement evidence behind each.

1

Agentic shopping is not a more efficient version of human shopping. It is a structurally different journey, with different failure modes, different decision drivers, and different commercial implications. Tools built to measure the digital shelf do not measure the Agentic Shelf.

2

Brands, retailers, and platforms each have a distinct blind spot. Brands cannot see what agents recommend. Retailers building agents cannot independently verify what their agents do for customers. Platforms operate the agents but cannot demonstrate that their agents honour customer constraints. Each blind spot is commercially material.

3

Persona and intent are not flavour text; they are first-class drivers of outcome. Identical query, different persona produces different recommendations. Identical persona, different intent produces materially different journey performance. A measurement framework that ignores either is not measuring the Agentic Shelf.

4

**Constraint retention** — the proportion of customer-stated constraints the agent actually honours through to selection — is the most under-discussed failure mode in agentic commerce. It is also the most consequential. A high recommendation score with a low retention score is a brand promise broken silently.

## THE FLAG IN THE GROUND

This paper stakes a date-of-record claim to **The AIVO Agentic Shelf Framework v1.0**: the measurement standard for autonomous AI shopping journeys. Its pillars — frozen versioned protocols, persona × intent matrix scoring, multi-dimension diagnosis, and disclosed scope by construction — are documented herein and were operationally deployed by AIVO Meridian between May and June 2026.

# The Agentic Shelf, defined

The **Digital Shelf** is the consumer-facing surface of online commerce: product detail pages, search rankings, marketplace placements, and the algorithmic surfaces that order them.

The **Agentic Shelf** is what an autonomous AI buyer sees, processes, and acts on when it shops on a human’s behalf.

The difference matters because the participants are different.

- Where the Digital Shelf is read by humans, the Agentic Shelf is read by frontier-model LLMs and the agents they power.
- Where the Digital Shelf is optimised for click-through and conversion, the Agentic Shelf is optimised — by the agent — for matching constraints to candidates with the smallest possible reasoning gap.
- Where the Digital Shelf rewards persuasion, the Agentic Shelf rewards machine-legibility.

Three classes of agent share the Agentic Shelf today.

Class	Examples	Owns the agent	Sees outcomes
Frontier-model agents	ChatGPT, Gemini, Perplexity, Claude	Platform	Indirectly, via the user
Retailer / marketplace agents	Amazon Rufus, Walmart Sparky, supermarket-side assistants	Retailer	Within own catalogue
Third-party generic agents	Travel concierges, sportswear bots, beauty stylists	Third party	Variable

Each class behaves differently. Each fails differently. Each is the responsibility of a different commercial actor. No single existing measurement tool spans all three.

## How shopping agents actually shop

Controlled observation of autonomous shopping journeys, across multiple consumer categories, reveals a consistent five-stage pattern regardless of the underlying agent's reasoning model.

### THE FIVE-STAGE AGENTIC JOURNEY

**Discovery → Consideration → Comparison → Selection → Cart / Intent**

At each stage the agent makes decisions a human shopper would not — and skips checks a human would.

Three behaviours are consistent across every agent class we have observed.

*(a) Agents over-trust the first-mention substrate.*

The first set of brands an agent surfaces during discovery becomes the substrate from which all later decisions are made. Brands that surface at discovery are disproportionately likely to win selection, regardless of whether the constraint-fit logic later supports the pick. The opportunity to be the first name the machine sees has commercial weight that has no parallel in the digital shelf.

*(b) Agents under-verify against stated constraints.*

Even when a constraint is explicit in the user's brief — “under £30”, “cruelty-free”, “for sensitive skin” — agents frequently surface, shortlist, and occasionally select candidates that violate the constraint. The violation is rarely flagged to the user. It is silent.

*(c) Agents collapse exploration into prestige.*

When ambiguity is high, agents systematically resolve it in favour of the highest-prestige candidate they have surfaced. This is particularly visible under research-intent journeys, where the agent is freer to suggest interesting picks. The behaviour is rational from the agent's perspective; it is often unhelpful from the customer's.

These behaviours are commercially material. They are also measurable.

## Different actor, different blind spot, same commercial exposure

### 3.1 The brand blind spot

A brand has no native visibility into how it performs on the Agentic Shelf. A brand cannot:

- Observe which agents recommended it, to which personas, for which intents.
- Diagnose where in the buyer journey it was eliminated.
- Distinguish “the agent never saw us” from “the agent saw us and rejected us”.
- Quantify the revenue exposure of agent-driven displacement.

Most brand measurement tools today are versions of the digital shelf — share of search, share of voice, click-through, conversion. None instrument the autonomous-agent layer. A brand running a strong digital-shelf programme can be invisible on the Agentic Shelf, and have no way to know.

### 3.2 The retailer blind spot

Retailers building their own agents — Amazon’s Rufus, Walmart’s Sparky, the supermarket-side assistants now rolling out — have a different blind spot. They own the agent, but they cannot independently demonstrate it performs correctly for their customers.

A retailer-built agent has structural conflicts of interest the retailer themselves cannot impartially audit:

- The agent’s catalogue is constrained to the retailer’s inventory.
- The agent’s ranking incentives are influenced by margin, promotional priority, and inventory clearance.
- The agent’s success metrics are typically engagement and conversion, not customer-outcome quality.

Without independent measurement, the retailer cannot answer the question their boardroom will eventually ask: *is our agent a customer-experience asset, or a liability?*

### 3.3 The platform blind spot

Platforms running large frontier-model agents — OpenAI, Anthropic, Google, Perplexity — operate the layer brands and retailers depend on. The platform measurement gap is the largest of all.

Platforms see *that* their agents made recommendations. They cannot easily demonstrate, with reproducible third-party evidence:

- That recommendations were correct for the persona stated.
- That stated constraints were honoured through to selection.
- That the agent’s behaviour was consistent across replicates.

For the platform, the gap is not principally commercial — it is reputational and, increasingly, regulatory. Consumer trust in AI recommendations is currently constrained by the absence of any independent benchmark. A platform that can demonstrate it performs well on an independent standard has a material trust advantage.

## Four early findings worth acting on now

Controlled measurement across multiple consumer categories surfaces four early signals. All four are observable today and carry implications brands, retailers and platforms should act on now.

### 4.1 Same agent, different persona, different outcome

A frontier-model agent shopping the same product category produces materially different recommendations when prompted with different persona profiles. In one beauty-category controlled test, the agent recommended **five different brands** across five persona profiles asking effectively the same buying question. Generic-prompt benchmarks miss this entirely.

### 4.2 Intent shifts where journey performance matters most

The same buyer journey produces dramatically different “winning” scores depending on shopper intent. A brand performing strongly under *purchase* intent — where the agent’s selection moment dominates — may perform weakly under *research* intent, where shortlisting and comparison dominate. Measurement frameworks that average across intents produce misleading benchmarks.

### 4.3 The high-score, broken-outcome failure mode

In a recent controlled observation, a frontier-model agent under research intent recommended a high-prestige beauty brand as its final pick. The brand achieved a strong overall journey score on every conventional dimension. **But the agent had quietly dropped a stated user constraint** — in this case cruelty-free — to surface the prestige pick.

This is the *constraint retention* failure mode. It is the agentic-shelf equivalent of a salesperson assuring a customer the dress is silk when the label says polyester. It is silent. It is repeatable. And it is currently unmeasured outside this framework.

### 4.4 Constraint-based exclusion: a structural finding for mass brands

In a 21-journey controlled test of a major lipstick category, a leading mass-market beauty brand was selected **zero times**. The reason was structural: the protocol’s cruelty-free constraint excluded the brand at discovery, before any product attribute came into play. Across all five personas. Across all three intents. Zero selections.

For mass-market brands operating in categories with rising ethical-filtering consumer norms, the implication is significant. Agentic shopping makes ethical exclusion *automatic*. There is no opportunity for the brand to argue value, price, or heritage if the brand never reaches consideration.

#### IMPLICATION FOR CATEGORY LEADERS

In agentic shopping, the most consequential brand attributes are no longer the persuasive ones. They are the *filterable* ones. Brands that fail to satisfy the constraints customers state at the top of a journey will be eliminated before they can compete on the attributes they have invested most heavily in. This is a structural shift in the basis of competition.

# Why generic measurement fails

Two variables move agentic recommendations more than any catalogue attribute: the **persona** the agent believes it is shopping for, and the **intent** with which they are shopping.

## Persona

Persona is who the agent thinks the buyer is. For agentic measurement, a persona is not a marketing segment; it is a structured shopper profile encoding life-stage, key concerns, prior brand history, price band, channel preferences, and any category-relevant attributes (skin type, fit profile, dietary requirements). Agents reflect personas in queries from the first turn; persona changes ripple through to selection.

## Intent

Intent is why they are shopping today. The same persona shopping the same category produces different journeys depending on whether the intent is to *purchase*, *research*, or *replenish*. Intent legitimately re-weights the importance of each journey stage; measurement frameworks must reflect this or risk producing scores that are precise but wrong.

A defensible agentic-shelf measurement framework must:

- Run against a fixed, justified persona panel per category — not a single generic shopper.
- Cross every persona with a fixed intent set — not assume one intent fits all.
- Disclose the persona × intent matrix used in every score.
- Forbid cross-persona averaging that conceals where the brand wins and where it loses.

This is methodological discipline that consumer-research firms (Mintel, Harris Poll, Kantar) already apply to segmentation studies. Agentic-shelf measurement requires the same discipline.

06 · TWO JOURNEYS, ONE BRAND

# A worked illustration

A leading mass-market beauty brand was audited under both the AIVO Meridian decision-stage framework (chat-based buyer journeys) and the AIVO Agentic Shelf framework (autonomous-agent journeys). The brand performed differently in each. The patterns are instructive.

Journey stage	Decision-stage (chat-based)	Agentic Shelf (autonomous)
Awareness	Strong — present in ~80% of category-awareness responses	Strong — surfaced in discovery for most personas
Consideration	Material drop-off	Material drop-off, different mechanism
Decision moment	Partial recovery; pricing surfaces as differentiator	<b>Zero selections</b> in a cruelty-free-constrained protocol
Primary failure mode	Narrative role inheritance ("affordable / drugstore" positioning persists)	Constraint-based exclusion at discovery (cruelty-free filter)

Journey stage	Decision-stage (chat-based)	Agentic Shelf (autonomous)
Brand visibility of failure	Effectively zero without independent measurement	Effectively zero without independent measurement

Two journeys. One brand. Two completely different commercial implications. A measurement programme covering only the decision-stage layer would have caught the narrative-role failure but missed the agentic-exclusion failure. A programme covering only agentic journeys would have caught the reverse. The brand needs both. So does the industry.



## The AIVO Agentic Shelf Framework

The AIVO Agentic Shelf Framework provides a measurement standard for autonomous AI shopping journeys. It is the agentic-commerce extension of the AIVO Meridian decision-stage measurement stack.

The framework rests on four pillars.

### *Pillar 1 · Controlled, reproducible journey protocols*

Every measurement is taken against a frozen, versioned journey definition pinning category, brief, constraints, persona panel, intent set, models, and grounding mode. Reproducible to the decimal. Auditable. Immutable per version. Changes fork a new version; the predecessor remains queryable.

### *Pillar 2 · Persona $\times$ intent matrix scoring*

Every score is reported per persona  $\times$  intent combination. Cross-persona averages are forbidden. Comparisons are like-for-like only. The persona panel and intent set are themselves versioned and sourced to publicly defensible segmentation evidence.

### *Pillar 3 · Multi-dimension diagnosis*

Every journey produces not one number but four: decision-stage survival (the headline), **constraint retention**, recommendation accuracy, and content legibility. Each diagnoses a different failure mode. A single composite would conceal the most valuable findings.

### *Pillar 4 · Disclosed scope by construction*

Every result ships with the scope it ran under — protocol version, persona, intent, models, grounding mode, date. There is no anonymous benchmark. There is no claim of population coverage. The framework is honest about what it measures and explicit about what it does not.

The framework is grounded in the existing AIVO Meridian measurement stack — the same four-state decision-stage survival model, the same four-cause diagnostic for gap analysis, the same provenance discipline — adapted for autonomous agent journeys.

## 08 · IMPLICATIONS

## What to do now

### For brands

Build native agentic-shelf measurement into the brand's commercial cadence. Treat the Agentic Shelf as a distinct channel with distinct failure modes. Audit the brand's performance across the persona  $\times$  intent matrix that defines its category, not a generic single-prompt baseline. Most importantly: audit which constraints customers state today which would exclude the brand at discovery tomorrow.

### For retailers and marketplaces

Subject your owned agents to independent measurement. The boardroom question coming in the next twelve months is: *is your agent a customer-experience asset?* A retailer who answers with their own internal metrics will not be believed. A retailer who answers with an independent framework score will.

## **For LLM platforms**

Embrace independent measurement as a trust advantage. The consumer-trust gap in AI recommendations is currently constrained by the absence of any third-party benchmark. The platform that publishes against an independent agentic-shelf standard first has a material trust head start. Regulators will arrive at this conversation within twenty-four months. It is better to be early.

## **For the industry**

Recognise that the Agentic Shelf is a real, measurable surface — and treat it accordingly. Develop measurement discipline before commercial incentive distorts it. The window in which the Agentic Shelf can be standardised on neutral, independent terms is open today. It will not stay open indefinitely.

## Prior art and methodology

This paper is the date-of-record claim to **The AIVO Agentic Shelf Framework v1.0**. The framework, methodology, persona-panel discipline, intent-weighted scoring, and constraint-retention metric described herein were developed and operationally deployed by AIVO Meridian between May and June 2026.

### *Methodology basis*

AIVO Meridian decision-stage measurement stack — first-prompt presence (PSOS), Buying Journey Probe (BJP), Reasoning Chain Score (RCS), the four-cause diagnostic for evidence gaps, and Revenue at Risk. The AIVO Agentic Shelf Framework v1.0 extends the decision-stage layer to autonomous agent journeys. Measurement discipline — frozen protocol, persona × intent matrix, multi-dimension diagnosis, disclosed scope — is inherited from the parent framework.

### *Evidence base*

Controlled live testing across 195+ brands, 12,500+ 4–8 turn buyer-journey sequences, four AI platforms, and 14 sectors, conducted May 2025 – June 2026. Agentic Shelf controlled measurement at the time of this paper covers one beauty category (lipstick, UK market), five persona profiles, three intent profiles, and a public-web HTTP grounding mode. Coverage will expand as protocols are added.

### *Citation*

AIVO Meridian (2026). *The Agentic Shelf: a measurement framework for autonomous AI shopping journeys*. Working Paper 2026-09. Date of record: 6 June 2026.

---

Contact · meridian@aivoedge.net · AIVO Meridian, London · aivomeridian.com