# Interpretability Engine Optimization: An Introduction

A teaching primer for researchers and students

**Juan Carlos López Castaño**
ORCID: 0009-0008-7471-0650
NSOLVIA Research

**Version 1.0** · 2026

**Companion to:** *Interpretability Engine Optimization (IEO): The Missing Optimization Layer in AI Commerce.* DOI: 10.5281/zenodo.22104281

## Contents


1 The problem, entered through one product 3

2 The definition, word by word 5

3 Three objects, not one 7

4 The operations, and how to tell if you are doing them 10

5 The artifacts 12

6 Measurement, and the distinction most people get wrong 14

7 How to test the claim 15

8 Open problems 17

9 Terminology and how to cite 18

### Preface: what this document is

The founding paper *proposes* a discipline. This primer *teaches* it.

The two documents cover the same ground, but not in the same order and not for the same reason. A proposal must defend itself: it states its claims precisely, bounds them, and invites refutation. A primer must build understanding: it starts from something concrete, introduces each concept at the moment it becomes necessary, and stops to let the reader practice.

**Who this is for.** Researchers evaluating whether there is anything here worth working on; students encountering the problem for the first time; and instructors deciding whether this belongs in a syllabus. It assumes no prior knowledge of commerce systems and no prior knowledge of machine learning. It assumes only that you are willing to look carefully at a single product.

**How to use it.** Read Parts 1 through 3 in order — they build. Parts 4 through 6 contain two exercises; do them, they take ten minutes each and they are where the distinctions become real. Part 7 is the replication protocol: read it if you intend to test the claims. Part 8 is the list of what remains unsolved, and it is the part written specifically for researchers.

**What this document deliberately does not do.** It does not argue for a vendor, a tool, or a product. IEO is proposed as an open discipline and is not trademarked. Everything in this primer can be practiced with any tooling, including none.

---

## Part 1 — The problem, entered through one product

Set aside the abstractions for a moment. Look at a single product.

A shapewear garment, sold online. Its page is well made: several photographs, a description written by someone who understands the product, a price, a size chart, care instructions, and — further down — a short paragraph of safety guidance about how long the garment should be worn and who should consult a physician before using it.

A human shopper reads this page in about eight seconds and knows most of what they need. They see the shape in the photograph. They infer the firmness from the material described. They read "for daily use under fitted clothing" and place the product in their life. If something is unclear, they ask, or they guess, or they decide the photograph is enough.

Now give the same page to a machine that must decide whether to recommend this product to someone who asked for *"something firm I can wear under a dress, that won't irritate sensitive skin."*

The machine does not see the shape. It cannot infer firmness from a photograph. It reads what the page *declares*.

## What the machine could resolve

From that page, a working system resolved the following about this product:

- **Category:** apparel → shapewear → body shaper, with high confidence
- **Product type:** bodysuit
- **Materials:** nylon, spandex, lycra, silicone
- **Functional intent:** waist shaping, tummy control, curve enhancement
- **Use cases:** everyday, under a dress, special occasion
- **Purchase signals:** latex-free, hypoallergenic, machine-washable

That is a genuinely useful interpretation. The machine can now match this product against most of the buyer's request: it is firm shapewear, it is worn under clothing, and it is latex-free — which speaks to the sensitive-skin constraint.

## What the machine could not resolve — and the two ways failure happens

Two fields came back empty, **and they came back empty for entirely different reasons.** This distinction is the first thing worth learning.

**The first field: the product identifier (GTIN).** There is no GTIN on the page. There is no GTIN anywhere in the merchant's data. The system declared the field unresolved and moved on. It did not invent a plausible number. This is what we call an **honest empty**: the information does not exist, so nothing is asserted.

**The second field: the restrictions.** The safety guidance *does* exist. It is written on the page, in prose, in complete sentences, and a human reads it without difficulty. But it is not expressed as a constraint a machine can act on. The system could not convert "consult a physician if you have circulatory conditions" into a structured restriction — so that field, too, came back empty.

Read those two failures again, because they are not the same failure:

> **The GTIN is missing. The restriction is *present and uninterpretable*.**

The first is an absence of information. The second is an absence of *interpretability* — the information is there, published, legible to any human, and still unusable by the system that must decide whether to recommend this product to a buyer with a medical condition.

**This second kind of failure is the subject of the discipline.** It is invisible to the merchant, because the page looks complete. It is invisible in analytics, because nothing breaks. And it is consequential, because a machine that cannot resolve a safety constraint will either ignore it — recommending the product to someone it should not — or discard the product entirely.

### The three costs

Generalize from the example. When a machine cannot correctly interpret an offer, cost appears at three stages:

1. **Eligibility.** An offer whose machine-facing fields cannot be resolved does not lose the comparison. It may never enter it. There is no impression, no ranking, no data point anywhere — the product was simply not a candidate.
2. **Decision.** Where data is ambiguous, generative systems may infer. An inferred material, size behavior, or use case is a recommendation error delivered with full confidence.
3. **Transaction.** A purchase made on a misinterpreted offer can result in a return — double freight, handling, and a customer who received something other than what was understood.

A fourth effect compounds behind these three and is harder to quantify: **reputation**. A misinterpreted offer costs trust twice — with the buyer, and plausibly with the recommending system itself, which staked its own reliability on a source that proved unreliable. This is stated as a mechanism, not as a measured outcome.

---

## Part 2 — The definition, word by word

Here is the definition. Read it once, then we will take it apart.

> **Interpretability Engine Optimization (IEO)** is the practice of making a business's offers, entities, attributes, policies, and source facts machine-interpretable before downstream retrieval, recommendation, comparison, or action occurs.

**"machine-interpretable"** — not machine-readable. A machine can *read* a paragraph of safety guidance; the characters parse. What it cannot do is *interpret* it into a constraint it can apply. Readability is about format. Interpretability is about resolved meaning. This distinction carries the whole discipline.

**"a business's offers, entities, attributes, policies, and source facts"** — the unit of analysis is not the page, not the URL, not the SKU. It is the *offer interpretation*: everything a machine must correctly understand about the complete offer, including who sells it and under what rules.

**"before"** — this is the word that separates IEO from everything adjacent. Retrieval, recommendation, comparison and action all happen *downstream*. They operate on whatever representation reaches them. If that representation is ambiguous at the source, every downstream decision inherits the ambiguity. **IEO addresses the layer that precedes them all.**

**A note on "engine."** IEO optimizes source data *for* interpretation engines — search engines, answer engines, generative systems, commerce platforms, autonomous agents. It never optimizes the engines themselves. You cannot modify how ChatGPT reasons. You can modify what it finds when it arrives.

## Where IEO sits among existing disciplines

Each existing optimization discipline answers a real question. None of them answers the one before it.

| Discipline | Question | Functional center of gravity |
|---|---|---|
| **IEO** | Can machines correctly **interpret** what you sell? | UNDERSTAND |
| SEO | Can they **find** you? | FIND |
| AEO | Can they **extract** you? | ANSWER |
| GEO | Can they **surface and cite** you? | CITE / SURFACE |
| Agentic Readiness | Can they **compare, decide, act** on you? | ACT |

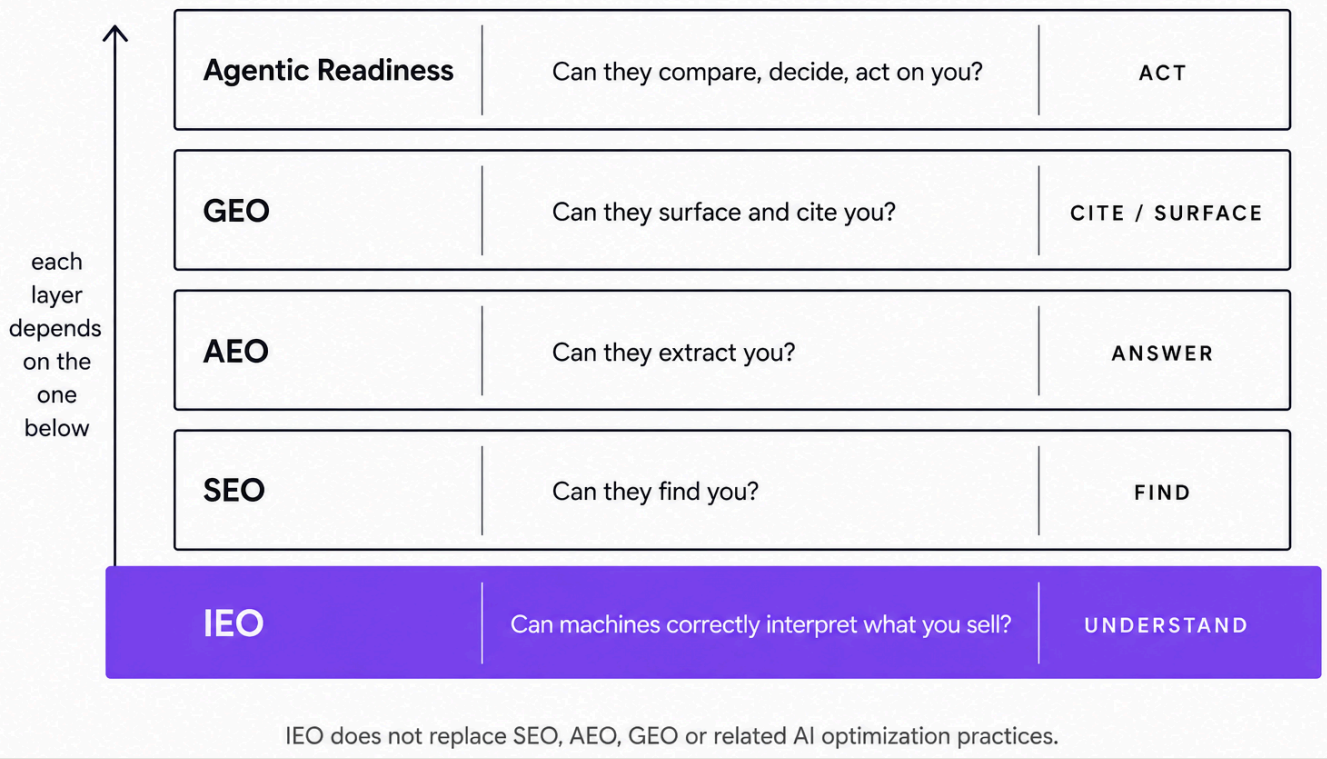


**Figure 1.** *The IEO ladder: interpretability as the layer preceding discovery, extraction, citation and action.*

**IEO does not replace any of these.** It addresses the interpretability layer they depend on. A page can rank first and still be uninterpretable; a brand can be cited frequently and still be described incorrectly. These are functional centers of gravity, not impermeable borders.

**A disambiguation for readers coming from machine learning.** In ML research, "interpretability" (XAI) asks whether *humans can interpret the model*. Commerce IEO asks the reverse: whether *models can interpret the merchant*. Same word, opposite direction. It is worth stating clearly the first time, because the collision of terms causes real confusion in seminars.

---

## Part 3 — Three objects, not one

Return to the shapewear example. Suppose every product field resolved perfectly. Is the offer now interpretable?

No — because the machine still does not know **who is selling it** or **under what rules it can be bought.**

Machines evaluating an offer must resolve three distinct objects:

**1. Product Interpretability** — what is being sold: identity, taxonomy, attributes, materials, use cases, functional intent, purchase signals, restrictions, and the differences between near-identical items.

**2. Brand Interpretability** — who stands behind it: identity, provenance, manufacturing facts, certifications, claims and the evidence supporting them.

**3. Transactional Interpretability** — the rules under which it can be bought: price, availability, shipping, returns, eligibility, exclusions.

> *Product tells the machine what is being sold. Brand tells it who stands behind it. Transaction tells it under what rules it can be bought.*

The third object is the one almost no one treats as an interpretation problem, and it is worth dwelling on. Consider a buyer who asks an agent to *buy* — not recommend, buy — a garment that ships internationally and can be returned if the size is wrong. To act on that request, the machine must resolve: does this store ship to that country, at what cost, in how many days, is this item returnable, within what window, and is a refund issued in money or in store credit?

Those are not marketing facts. They are **preconditions for a transaction**, and if a machine cannot resolve them, it cannot complete the task — regardless of how well the product itself is described. The offer fails at the last step, for reasons that have nothing to do with the product.

Figures 3 and 4 show a Semantic Product Record and a Semantic Brand Record, including the transactional rules.

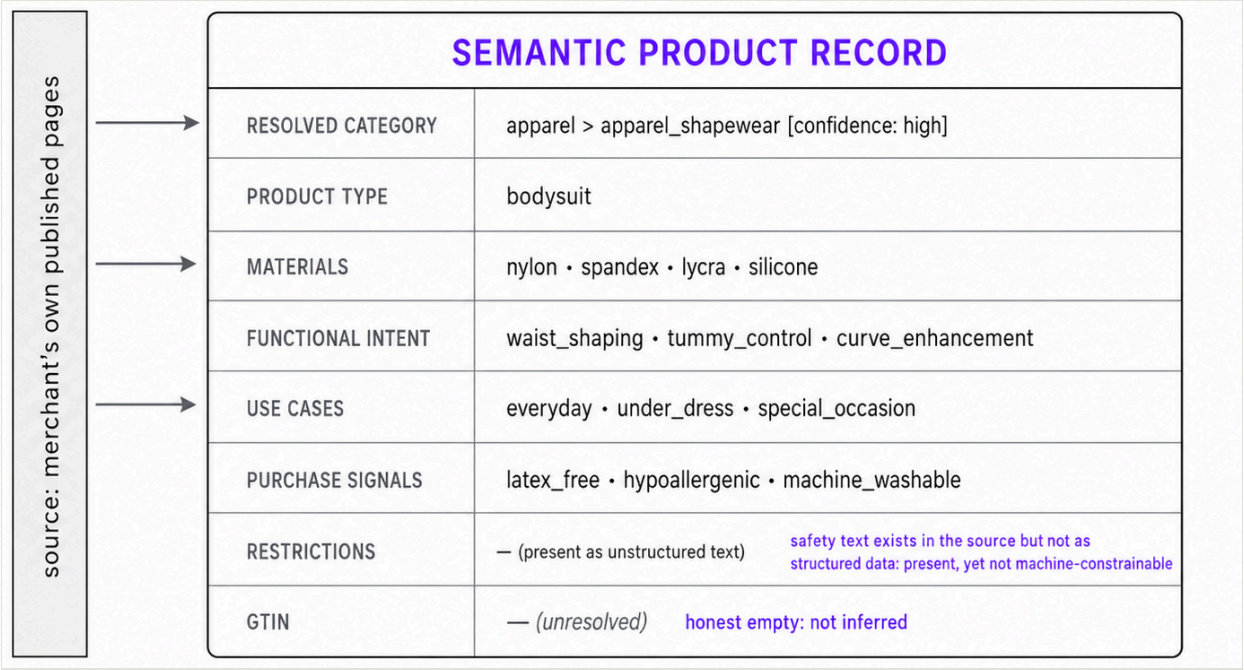


**Figure 3.** *A Semantic Product Record (PonteBella, ref. 30120), including two distinct kinds of gap: a field left unresolved rather than inferred, and a field whose content exists in the source as unstructured text and is therefore not machine-constrainable.*


**Figure 4.** *A Semantic Brand Record (PonteBella), including the transactional rules that determine whether an offer can be evaluated and purchased.*

*Both figures use a brand owned by the author and operated as the research laboratory for this work; see Competing Interests.*

# Part 4 — The operations, and how to tell if you are doing them

IEO does not invent new techniques. Normalization, enrichment, taxonomy resolution, grounding and structured data all predate it. What IEO proposes is organizing them around a single measurable objective.

## The typical pipeline

**Normalization → Semantic Resolution → Enrichment → Validation/Grounding → Machine-Interpretable Representation**

First, make inconsistent data consistent. Then resolve what the product and its context actually *are*. Then make explicit what was implicit — intent, use cases, decision criteria. Then validate that nothing unsupported was introduced. The result is an interpretable record.

The process can be executed by humans, rules, AI systems, or any mixture. **IEO does not depend on a specific implementation.**

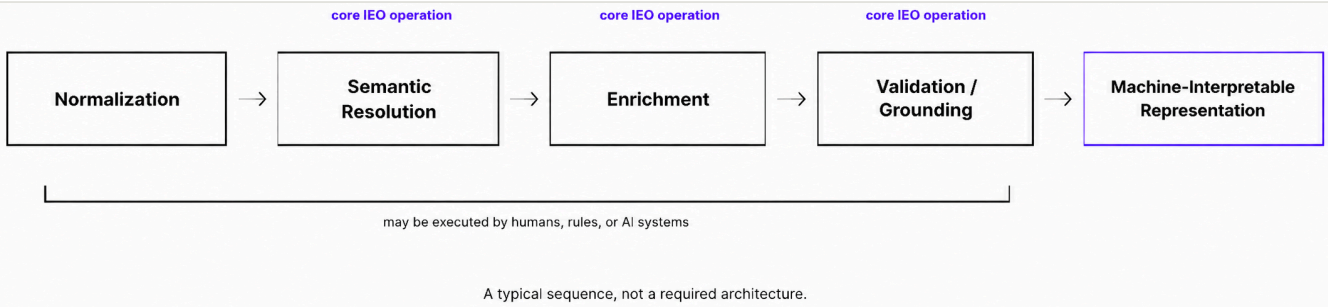


**Figure 2.** *The IEO pipeline. A typical sequence, not a required architecture.*

Note the ordering: normalization is an IEO operation, but resolution, enrichment and validation are the *core* ones. A perfectly normalized catalog of semantically ambiguous products is a perfectly normalized catalog of semantically ambiguous products.

## The Boundary Test

Because IEO borrows techniques from adjacent practices, you need a way to tell whether a given intervention belongs to it. One question does the work:

> **Did the intervention improve the machine's understanding of the source — or did it only improve how the source is formatted, distributed, discovered, or surfaced?**

The practical form: *if the same information were delivered through a different channel tomorrow, would the improvement in understanding still remain?* Interpretability survives the channel. Positioning does not.

| Layer | IEO? |
|---|---|
| Formatting (JSON-LD/XML/feeds as such) | Not by itself |
| Normalization | IEO operation |
| Semantic resolution | Core IEO operation |
| Enrichment | Core IEO operation |
| Validation / grounding | Core IEO operation |
| Exposure / distribution | Downstream deployment |
| Ranking / recommendation / action | Downstream outcomes |

And the proof that interpretability is a distinct property: **a record may be valid JSON-LD, perfectly normalized, and indexed — and still be semantically ambiguous.** IEO evaluates the ambiguity that remains.

### Exercise 1 — Apply the boundary test

For each intervention below, decide: **IEO, or not IEO?** Then check your reasoning against the answers.

**(a)** A merchant adds Product schema markup to every product page. The markup is valid and passes every validator. The `category` property is populated with the merchant's own internal label, "SS-24 Collection."

**(b)** A merchant resolves every product to an official taxonomy category and populates the attributes that category requires. Nothing is published or exposed differently; the data sits in the catalog.

**(c)** A merchant submits their catalog to three new marketplaces.

**(d)** A merchant rewrites product descriptions to include more keywords for search.

**(e)** A merchant reviews their safety guidance and expresses each warning as a structured constraint, retaining the original prose on the page unchanged.

---

**ANSWERS**

**(a) Not IEO.** The format improved; the meaning did not. "SS-24 Collection" resolves to nothing a machine can compare against. This is the clearest illustration of why formatting alone fails the test: the record is now beautifully encoded ambiguity.

**(b) IEO.** Nothing was exposed, distributed, or surfaced — and the machine's ability to understand the offer improved. Run the practical form of the test: deliver this through any channel tomorrow, and the improvement remains.

**(c) Not IEO.** Distribution. The same representation now reaches more destinations; whether any of them can interpret it is unchanged.

**(d) Not IEO.** This is downstream optimization aimed at retrieval. It may work. It is not interpretation.

**(e) IEO** — and it is the exact case from Part 1. Note what makes it IEO: the information already existed and was already published. What changed is that it became *resolvable*. Note also what did **not** happen: the human-facing prose was not degraded to serve the machine. Both layers state the same facts.

## Part 5 — The artifacts

The output of IEO practice is a family of verified, machine-interpretable records: the **Semantic Product Record**, the **Semantic Brand Record**, and the **Semantic Transaction Record**.

## Three states, and why the distinction protects everyone

A record exists in three states, and confusing them produces bad arguments in both directions:

- **The source** — merchant data as it arrives, in whatever condition.
- **The ingested record** — what a system honestly resolved from an imperfect source.
- **The gold record** — the verified, curated representation after remediation and, where the source is thin, human-confirmed facts.

**An ingested record must never be judged as if it were the gold record.** It is the raw material on which the discipline demonstrates its value, not the demonstration. A critic who points at a sparse ingested record and concludes "the method does not work" has confused the input for the output. Equally, a vendor who shows a gold record and implies it was produced automatically from a thin source has confused the output for the input.

## Two principles

**Verified, never invented.** Where evidence is insufficient, the field remains explicitly unresolved — an *honest empty* — rather than inferred. This is a methodological commitment before it is an ethical one: a system that fabricates plausible values cannot be evaluated, because you can no longer distinguish what it resolved from what it guessed.

**The mirror principle.** Human-facing and machine-facing layers state the same facts, because they derive from the same source. Anyone can verify the match by reading the page. This rules out the oldest manipulation in search — text for machines that humans never see — and it makes the practice auditable by construction.

**The records are the asset. The destinations are adapters.** Structured data, feeds, catalogs, agent files, APIs, and protocols that do not exist yet are all ways of delivering a representation. Protocols tell machines *how* to exchange commerce data; IEO asks whether the data being exchanged carries enough verified meaning to be correctly interpreted.

---

# Part 6 — Measurement, and the distinction most people get wrong

A discipline requires an instrument. Interpretability can be observed through **resolution tasks**: can a machine resolve identity, taxonomy, attributes, intent, use cases, trust and safety information, and retrieval language for a given offer?

Any tool that measures this is measuring commerce interpretability. **IEO does not prescribe a single score or a single instrument** — that independence is what makes it a discipline rather than a product.

## Three measurement terms — and they are not interchangeable

| Term | What it is |
|---|---|
| **Interpretability Baseline** | the observed current state, *measured* from the source representation. |
| **Projected Interpretability Lift** | the estimated improvement of a remediated version, *generated* before deployment. |
| **Observed Interpretability Lift** | the difference *measured* after remediation has actually been deployed and re-evaluated. |

These three carry **different evidential weight**, and the difference is not pedantry. Here is the empirical demonstration, from the founding paper's own corpus:

When the same product was measured twice under identical conditions, the **baseline score was identical in 52 of 52 pairs (100%)**. The measurement is deterministic.

When the **generated representation** underlying the projection was examined field by field across two consecutive runs of the same product, it **diverged in 4 of 32 fields** — not in the resolved category, functional intent or use cases, which were stable, but in the enumeration of materials and in the attribute schema itself, where the second run used different keys for the same product.

> **The baseline is reproducible. The projection is an estimate.**

This is expected behaviour of a generative step. It is not a defect — it is a property, and a discipline that does not separate measurement from projection will eventually mistake one for the other. That is the single most common failure mode in the adjacent "AI visibility" market: estimates presented in the vocabulary of measurement.

### Exercise 2 — Classify the claim

For each statement, identify whether it reports a **baseline**, a **projected lift**, an **observed lift**, or **none of the three**.

**(a)** "Your catalog scores 34 out of 100 today."
**(b)** "After optimization, your catalog would score 81."
**(c)** "We deployed the changes in March and re-measured in May; the score moved from 34 to 39."
**(d)** "Our clients see a 3x increase in AI visibility."

---

**ANSWERS**

**(a)** Baseline — a measurement of the current state. **(b)** Projected lift — note the conditional. **(c)** Observed lift — deployment happened, and re-measurement used the same instrument. **(d)** None of the three. "AI visibility" is not defined, "3x" has no denominator, and no instrument is named. This is the claim shape the discipline exists to make unnecessary.

**One further caution, stated because the founding paper states it about itself.** An instrument that measures whether a field is *populated* is not measuring whether its content is *interpretable*. Two products with equally complete but unequally interpretable content can receive similar structural scores. This biases baselines *upward* — meaning reported failure rates are, if anything, conservative. Know the bias direction of any instrument you use.

---

## Part 7 — How to test the claim

The central claim of the discipline is that **a more interpretable representation produces better machine understanding of the same offer.** That claim is falsifiable, and this is how.

### Stage 1 — Remediation

Compare the offer as received against the same offer after IEO processing (resolution → enrichment → validation), measuring what the machine can now correctly resolve. This demonstrates remediation capacity on real, imperfect sources. It is necessary but not sufficient: it shows the process does something, not that representation quality is the cause.

### Stage 2 — The IEO Counterfactual Product Test

This is the experiment that can kill the discipline.

**Construction.** Take one fully processed offer. Produce a controlled **degraded** version: category genericized, attributes removed, application constraints deleted. Everything else — price, images, brand, page structure — held constant.

**Administration.** Give the same AI system the same questions and access to both representations.

**Measurement.** Score six outcomes against a golden interpretation record defined in advance:

1. attribute extraction accuracy
2. category resolution
3. constraint satisfaction
4. differentiation between near-identical items
5. hallucination rate
6. task completion

**Protocol requirements — these are what make it an experiment rather than a demonstration:**

- Same model and version, same system configuration, same questions, same retrieval permissions and settings.
- Representations evaluated **separately and blinded** — the model must never see both versions simultaneously.
- **Randomized order** across multiple runs.
- **Gold labels defined in advance**, not after seeing results.
- Results reported **with confidence intervals**.

**The falsification condition, stated plainly:** if processed representations do not outperform degraded ones under these conditions, the discipline's premise fails.

**Why this experiment is unusually clean.** Downstream optimization practices measure against surfaces that shift beneath the measurement — model updates, retrieval changes, non-deterministic responses. IEO controls its experimental object: the same offer, in two representations, questioned by the same system under the same conditions, isolates representation quality as the variable. **A discipline whose object does not move can accumulate knowledge; one whose object moves can only accumulate anecdotes.**

---

## Part 8 — Open problems

This is the part written for researchers. Everything below is unsolved, and none of it requires access to any particular vendor's systems.

1. **Measuring brand interpretability.** Product interpretability can be measured today; brand interpretability cannot, in any standardized way. What are the resolution tasks for a brand? How do you score whether a machine correctly understood *who stands behind an offer*? This is the largest open gap in the framework.

2. **Separating representation effects from model effects.** The counterfactual test holds the model constant, which isolates representation quality *for that model*. It says nothing about whether the effect generalizes across model families, or whether more capable models close the gap by inferring better. A cross-model replication is the obvious next study, and it could substantially narrow or widen the discipline's scope.

3. **Standardizing gold labels across instruments.** A golden interpretation record is defined by whoever runs the test. Until there is an inter-instrument standard — or at minimum an inter-rater reliability protocol — results from different instruments are not comparable. This is a prerequisite for the field having a shared evidence base rather than parallel private ones.

4. **Content quality at the structural layer.** Current instruments measure field presence. Measuring whether the *content* of a populated field is interpretable is unsolved, and it is what stands between today's conservative baselines and accurate ones.

5 **Quantifying the reputational effect.** The claim that a misinterpreted offer degrades trust with the recommending system is stated as a mechanism with no measurement attached. Whether systems that weigh sources actually down-weight unreliable ones over time is an empirical question no one has answered.

6 **The economics.** The chain of bridges — baseline → lift → operational effect → platform effect → commercial effect — is stated as measurable. It has not been measured end to end on any merchant. Longitudinal studies are pending, and they are the studies most likely to determine whether this discipline matters commercially or only technically.

7 **Interpretability debt over time.** Catalogs change and machine surfaces evolve. Whether interpretability degrades at a measurable rate, and what that rate depends on, is entirely unstudied.

---

# Part 9 — Terminology and how to cite

## Core terms

**IEO — Interpretability Engine Optimization.** The practice of making a business's offers, entities, attributes, policies, and source facts machine-interpretable before downstream retrieval, recommendation, comparison, or action occurs.

**Interpretability Engine.** Any machine system that consumes business information and must form a usable interpretation of it.

**Commerce IEO.** The application of IEO to products, catalogs, brands, offers and commerce operations.

**Product / Brand / Transactional Interpretability.** The degree to which a machine can correctly understand, respectively: what is being sold; who stands behind it; and under what rules it can be bought.

**Semantic Resolvability.** The degree to which the meaning required for machine interpretation can be resolved from available evidence without unsupported inference.

**Honest empty.** A field left explicitly unresolved because evidence was insufficient, rather than inferred.

**Interpretability Debt.** The accumulated gap between what a business knows about its offers and what its machine-facing representations make explicit enough for machines to understand correctly.

*Full definitions are maintained in the public IEO glossary at nsolvia.com/ieo/glossary.*

### Two canonical framings

## IEO optimizes understanding.

## Semantic discovery is downstream of interpretation.

### How to cite

**The founding paper (cite this for the discipline itself):**
López Castaño, J. C. (2026). *Interpretability Engine Optimization (IEO): The Missing Optimization Layer in AI Commerce.* NSOLVIA Research. Version 1.0. DOI: 10.5281/zenodo.22104281

**This primer:**
López Castaño, J. C. (2026). *Interpretability Engine Optimization: An Introduction.* NSOLVIA Research. Version 1.0.

### A closing note on ownership

IEO is proposed as an open discipline. The term is not trademarked, and no vendor holds exclusive rights to the practice. You may teach it, adapt it, build instruments for it, disagree with it in print, and publish results that contradict it — under CC BY 4.0, with attribution.

That is the intended outcome. A discipline that only one company can practice is not a discipline.

---