sustainability

MDPI

Article

Developing Consumer Trust in Sustainable Agentic Commerce: Conditional Signals, Auditable Governance, and Justified Non-Purchase Advice

Arpita Agnihotri

Management Department, School of Business Administration, Penn State Harrisburg, Harrisburg, PA 17057, USA; axa671@psu.edu

Abstract

Agentic artificial intelligence (AI) changes digital retail from consumer-led search to delegated decision making, increasing both the potential value and the agency risks of recommendation. This conceptual article asks the following: Under what architectural and governance conditions can an autonomous shopping agent earn calibrated consumer trust while supporting economically and environmentally appropriate consumption decisions? Using a purposive, multidisciplinary synthesis of signaling theory, principal–agent theory, the ability–benevolence–integrity model of trust, trustworthy AI, recommender system governance, and sustainable consumption research, we develop a process model linking four architectural features to consumer perceptions, delegation, reliance, and decision outcomes. The features are scoped provenance verification, disclosed and audited commercial incentives, faithful contrastive explanations, and active evaluation of retention, repair, refurbishment, and purchase. We argue that justified non-purchase advice can function as a credible signal of consumer alignment only when the operator is the signal sender, the advice creates observable opportunity or accountability costs, and compensation and audit arrangements make imitation more costly for a misaligned operator. The framework distinguishes perceived trustworthiness from overall trust, willingness to delegate, actual reliance, recommendation appropriateness, and lifecycle sustainability. A five-layer architecture, a worked decision example, and four testable propositions specify how missing data, conflicting preferences, uncertainty, authorization, privacy, and accountability should be handled. This article treats the applications as illustrations rather than empirical cases and concludes with limitations and an empirical research agenda focused on calibrated—not merely increased—trust.

Keywords: agentic AI; autonomous shopping agents; sustainable consumption; consumer alignment; non-purchase advice; trustworthy recommender systems; signaling theory; principal–agent theory; calibrated trust; circular economy

Check for updates

Academic Editor: Gheorghe Epuran

Received: 9 October 2026
Revised: 25 October 2026
Accepted: 26 October 2026
Published: 29 October 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Digital commerce is moving from search and comparison toward delegated action. Generative models can interpret goals, call external tools, compare alternatives, and, when authorized, execute transactions. In retail, this shift turns a recommendation interface into a decision intermediary that can influence which alternatives enter the choice set, how trade-offs are framed, and whether a transaction occurs [1–3]. Delegation therefore raises a more demanding question than whether an agent can produce relevant recommendations:

whether the consumer should rely on the agent when the operator, merchant, advertiser, and user may have different objectives.

Sustainable consumption intensifies this problem. Claims about provenance, durability, repairability, labor practices, and lifecycle impacts are often difficult for consumers to verify. At the same time, a decision that saves money or extends product life is not automatically environmentally superior. The outcome depends on product-specific factors such as expected remaining life, repair inputs, operating efficiency, transport, displacement of new production, and rebound consumption [4]. A trustworthy shopping agent must therefore communicate both what is known and what remains uncertain rather than converting incomplete sustainability data into unwarranted certainty.

Existing research offers important but incomplete pieces. Trustworthy-AI and trustworthy-recommender system scholarship identifies transparency, fairness, explainability, responsibility, and robustness as design requirements [5,6]. Research on recommendation agents links performance and explanation to trust and adoption, while work on trust in automation warns that persuasive interfaces can produce overreliance [7–9]. Sustainable consumption research explains why information, social context, habits, and infrastructure shape consumer action [4]. These streams rarely explain how commercial incentives, verifiable product data, explanations, and the option not to purchase jointly shape trust and delegation in agentic commerce.

This article addresses the following research question: Under what architectural and governance conditions can an autonomous shopping agent earn calibrated consumer trust while supporting economically and environmentally appropriate consumption decisions? We use calibrated trust to mean that a consumer's trust and reliance correspond to the agent's demonstrated capabilities, limits, and incentive alignment, avoiding both unwarranted rejection and excessive reliance [7]. The object of trust is specified at each stage: consumers may evaluate the software agent's competence, the operator's integrity and benevolence, and the assurance infrastructure's reliability.

This article makes three contributions. First, it integrates principal–agent theory and signaling theory with the ability–benevolence–integrity model to explain a sequence from design and governance features to perceived trustworthiness, overall trust, delegation willingness, actual reliance, and decision outcomes. Second, it develops justified non-purchase advice as a conditional signal rather than proof of alignment. The signal is credible only when its sender, hidden characteristic, differential cost, observability, and compensation context are specified. Third, it translates these mechanisms into an auditable architecture that separates hard constraints from compensatory preferences, records uncertainty, preserves user authorization, and evaluates retention, repair, refurbishment, and purchase as distinct alternatives. The architecture is proposed for future implementation and testing; it is not presented as a validated system.

2. Conceptual Development Approach

This study follows a theory-synthesis approach to conceptual development [10,11]. It does not claim to be a systematic review. The purpose is to connect constructs that are usually studied separately and derive a parsimonious set of relationships suitable for empirical testing and technical design.

Research was selected purposively in four stages. First, foundational work was identified for the focal mechanisms: signaling under information asymmetry, agency conflict and moral hazard, ability–benevolence–integrity trust, trust in automation, and trustworthy AI. Second, recent research was sought on explainable and trustworthy recommender systems, AI-mediated consumer decisions, autonomous or agentic recommendation, algorithmic steering, sustainable consumption, circular business models, and lifecycle trade-offs. Third,

primary legal and organizational sources were used for claims about the EU Ecodesign for Sustainable Products Regulation (ESPR), Digital Product Passports (DPPs), the EU repair directive, Amazon's shopping assistant, and Patagonia's repair and resale practices. Fourth, backward and forward citation tracing was used to check construct definitions and identify counterarguments, particularly evidence on overtrust, disclosure effects, incomplete verification, and the conditional environmental value of repair.

The synthesis covered marketing, information systems, organizational theory, human factors, computer science, sustainability, and consumer law. Sources were retained when they clarified at least one of four links: (a) how an architectural feature changes the information available to a consumer; (b) how governance changes the operator's incentives or accountability; (c) how consumers infer ability, integrity, or benevolence; or (d) how trust translates into delegation, reliance, and economic or environmental outcomes. The synthesis was iterative: initial design features were mapped to theoretical mechanisms; rival explanations and boundary conditions were then used to narrow each claim.

The four propositions were derived by matching a distinct feature to a primary perception and a clearly defined downstream outcome. Proposition 1 concerns scoped provenance verification and attribute credibility. Proposition 2 distinguishes disclosure from enforceable incentive separation. Proposition 3 concerns faithful contrastive explanation and perceived ability. Proposition 4 specifies when justified non-purchase advice can signal alignment. Technical components in Section 4 are design implications of these relationships, not evidence that the propositions are true. The purposive approach enables multidisciplinary integration but may omit relevant studies and does not support claims about literature prevalence; this limitation is revisited in Section 9.

3. Integrated Theoretical Framework

The framework assigns different explanatory roles to four perspectives. Principal–agent theory diagnoses why delegation is risky: the consumer delegates search, evaluation, and possibly execution to an agent operated by an organization with private incentives [12,13]. Signaling theory explains how the operator may communicate an otherwise unobservable characteristic—commitment to consumer objectives—when cheap claims are insufficient [14,15]. The ability–benevolence–integrity model explains the consumer's resulting beliefs about the trustee [16]. Trustworthy-AI and responsible-recommender principles translate those beliefs into requirements for data, governance, explanation, oversight, and evaluation [5,6]. Their integration produces a causal sequence rather than four parallel descriptions.

In the delegated relationship, the consumer is the principal, but the word agent has two levels. The software agent selects and may execute an action, while the platform or service operator configures its objective function, data access, compensation, and controls. Accordingly, the operator—not anthropomorphized software—is normally the sender of a commercial-alignment signal. The consumer is the receiver. Merchants and certification bodies may send separate product quality or assurance signals. Separating these actors prevents trust in an interface from being mistaken for trust in the operator or the underlying data infrastructure.

The hidden characteristic of interest is the operator's commitment to applying the consumer's authorized objectives even when those objectives conflict with immediate transaction revenue. A signal is separating only when a misaligned operator would incur a greater expected cost to imitate it than an aligned operator. That cost could be foregone commission, lost margin, exposure to an independent audit, contractual penalties, or reputational loss if logged recommendations reveal systematic steering. Non-purchase advice therefore has weak signaling value when it is occasional, unobservable, strategically used

to retain subscribers, or redirected toward profitable repair, parts, or refurbishment. It has stronger value when it is repeated across eligible decisions, logged, independently auditable, linked to outcome-based accountability, and delivered under compensation arrangements that do not reward immediate purchase value.

The model of Mayer et al. [16] is adapted cautiously to a nonhuman decision intermediary. Perceived ability refers to beliefs that the agent can retrieve relevant evidence, model trade-offs, and execute authorized tasks reliably. Perceived integrity refers primarily to beliefs that the operator and system adhere to disclosed principles and constraints. Perceived benevolence refers to the consumer's belief that the operator has designed and governs the agent to protect consumer welfare rather than exploit informational dependence. These dimensions are antecedents of overall trust, not synonyms for trust itself. Overall trust may influence willingness to delegate; delegation creates an opportunity for actual reliance; reliance then produces decision outcomes. Trust is appropriate only if it is supported by recommendation quality and governance performance.

This distinction addresses a central risk: an explanation can increase confidence without increasing correctness. The framework therefore evaluates both subjective outcomes (perceived ability, integrity, benevolence, overall trust, and willingness to delegate) and objective or behavioral outcomes (actual reliance, recommendation appropriateness, constraint compliance, consumer welfare, and lifecycle environmental performance). Appropriate reliance occurs when consumers accept correct or well-supported recommendations and question or reject recommendations that are unsupported, conflicted, or outside authorization.

4. Proposed Architecture and Decision Governance

Table 1 presents the proposed five-layer architecture. Terms such as consumer-aligned and fiduciary are used normatively, not legally. In this article, consumer-aligned means that the operator commits the agent to the consumer's authorized objectives, disclosed constraints, and welfare within applicable law. It does not assert a legally enforceable fiduciary duty unless such a duty independently arises under the governing law or contract.

The architecture distinguishes hard constraints from compensatory preferences. Hard constraints are non-negotiable eligibility conditions—for example, a maximum authorized payment, an allergy exclusion, a prohibition on sharing diagnostic data, or a requirement for user confirmation. Compensatory preferences are objectives across which trade-offs are allowed—for example, price, durability, repairability, expected operating cost, and estimated environmental impact. If no option satisfies all hard constraints, the agent should not silently relax them. It should report infeasibility, identify which constraint creates the conflict, and request a user decision.

For each feasible action a, the agent can calculate an auditable decision score: U(a) = sum of w_j times v_j(a), minus risk and uncertainty penalties, subject to all hard constraints. The weights w_j are user-approved, and v_j(a) is the normalized value of action a on criterion j. The score is a decision aid, not an environmental truth claim. Data provenance, confidence intervals or ordinal confidence labels, and sensitivity to alternative weights should be retained in the decision record.

Missing or conflicting information is handled explicitly. A missing certificate is not coded as proof of poor performance; instead, the relevant attribute receives an unknown status and an uncertainty penalty chosen for the decision context. Expired certifications are flagged and excluded from current verified status. Conflicting indicators are not averaged unless they measure comparable boundaries. Lifecycle estimates must state system boundaries, functional units, geography, and major assumptions. Fairness review is required because exclusive reliance on costly certifications could disadvantage smaller suppliers with credible but uncertified practices.

Table 1. Proposed architecture for consumer-aligned shopping agent.

Layer	Core Function	Operational Requirements	Primary Theoretical Role
1. Intent, consent, and governance	Translate the user's goal into hard constraints, compensatory preferences, authorization scope, and revocation rules.	Confirm identity and consent; distinguish "must" from "prefer"; record budget, privacy, sustainability, and execution limits; allow revision.	User autonomy; bounded delegation; agency definition.
2. Evidence and provenance	Retrieve product, certification, repair, price, and lifecycle evidence without treating every source as equivalent.	Label source type and scope; check currency and coverage; preserve uncertainty; admit missing or conflicting data; avoid automatically excluding uncertified suppliers.	Scoped signaling; perceived ability and integrity.
3. Incentive and scoring governance	Keep commercial payments from silently changing consumer-utility scores.	Disclose material ties; separate ad eligibility from organic ranking; log weights and overrides; support independent audit and conflict testing.	Moral hazard control; operator integrity.
4. Alternative generation and arbitration	Evaluate retention, postponement, maintenance, repair, refurbishment, and new purchase as distinct actions.	Apply hard constraints first; compare feasible actions using explicit weights and uncertainty; test whether repair is technically, economically, and environmentally appropriate.	Consumer alignment; conditional benevolence signal.
5. Explanation, authorization, and recourse	Explain the recommendation and obtain permission proportionate to risk.	Provide contrastive reasons, evidence gaps, confidence, and commercial disclosures; require confirmation above thresholds; support cancelation, correction, and responsibility allocation.	Explicability; calibrated trust; accountable reliance.

Authorization is separated from recommendation. Low-risk informational tasks may proceed automatically, but purchase, repair booking, data sharing, financing, and recurring commitments require confirmation rules proportionate to financial and safety risk. Users must be able to revoke authority and correct preferences. The operator should define responsibility for erroneous purchases, failed repairs, stale data, and third-party service failures rather than transferring all risk through a generic disclaimer. Data minimization, purpose limitation, access controls, retention schedules, and an auditable consent record are part of the architecture rather than optional add-ons.

4.1. Worked Decision Example

A worked hypothetical example illustrates the decision logic; the values are assumptions, not observed product facts. A consumer reports that a three-year-old smartphone no longer lasts through the day. The consumer authorizes diagnostic access for battery health only, sets a hard immediate budget of USD 400, requires at least two expected years of service, and assigns compensatory weights of 0.40 to reliability, 0.30 to total two-year cost, 0.20 to lifecycle environmental evidence, and 0.10 to delivery time. Four actions are

generated: retain without intervention, replace the battery, buy a certified refurbished phone, or buy a new phone.

Table 2 shows the illustrative inputs. Retention is infeasible because expected service does not meet the user's two-year requirement. The new device is infeasible because its assumed price exceeds the authorized budget. Battery replacement and refurbishment remain feasible. Under the stated normalized values, repair receives the higher utility score, but the margin is modest and the environmental evidence is uncertain. The recommendation is therefore "replace the battery, subject to diagnostic confirmation and a maximum authorized quote," not an unconditional claim that repair is greener. Table 3 reports the criterion-level calculation.

Table 2. Hypothetical repair-versus-replacement decision.

Action	Hard-Constraint Result	Illustrative Evidence	Normalized Utility	Decision Treatment
Retain as is	Fails two-year service requirement	Cost: USD 0; expected service: 0.5 year	Not scored	Explain insufficiency; offer low-cost settings check.
Battery replacement	Feasible, conditional on diagnosis and quote	Cost: USD 110; expected service: 2.5 years; environmental evidence: medium/uncertain	0.77	Recommend with confirmation and price ceiling.
Certified refurbished phone	Feasible	Cost: USD 340; expected service: 3 years; environmental evidence: medium	0.69	Present as fallback if diagnosis fails.
New phone	Fails USD 400 budget	Cost: USD 650; expected service: 4 years; environmental evidence: incomplete	Not scored	Do not rank unless user revises budget.

Table 3. Criterion-level calculation of utility scores in Table 2.

Action	Reliability (w = 0.40)	Two-Year Cost (w = 0.30)	Env. Evidence (w = 0.20)	Delivery Time (w = 0.10)	Weighted Sum	Uncertainty Penalty	Utility U(a)
Battery replacement	0.80	0.90	0.60	0.80	0.79	−0.02	0.77
Certified refurbished phone	0.90	0.50	0.70	0.60	0.71	−0.02	0.69

The contrastive explanation would state "Battery replacement is recommended over refurbishment because both options meet your two-year service requirement, while the assumed repair cost is lower. The environmental comparison is uncertain because comparable lifecycle data are unavailable. I will not book the repair until the diagnosis confirms that no other major component is failing and the quote remains below your limit." This explanation identifies the decisive criteria, uncertainty, and execution boundary.

Table 3 applies the decision score defined in this section. Criterion scores are assumed values on a 0–1 scale (1 = best), assigned for illustration rather than derived from observed product data: reliability reflects expected service relative to the two-year requirement and failure risk; cost scores rank the feasible options by total two-year cost (USD 110 and USD

340); environmental scores reflect the evidence grades in Table 2; and delivery scores reflect assumed lead times. The weighted sum multiplies each score by its user-approved weight. For battery replacement, $0.40 \times 0.80 + 0.30 \times 0.90 + 0.20 \times 0.60 + 0.10 \times 0.80 = 0.79$; for the refurbished phone, $0.40 \times 0.90 + 0.30 \times 0.50 + 0.20 \times 0.70 + 0.10 \times 0.60 = 0.71$. Each option then receives an uncertainty penalty of 0.02: for battery replacement, because the repair depends on diagnostic confirmation and lifecycle evidence is uncertain, and for the refurbished phone, because its prior-use history and condition grade are imperfectly known. The resulting utilities are 0.77 and 0.69. The 0.08 margin holds under two alternative weightings: raising reliability to 0.50 (cost 0.25, environment 0.15, delivery 0.10) gives 0.775 and 0.72, and raising the environmental weight to 0.40 (reliability 0.30, cost 0.20, delivery 0.10) narrows the gap to 0.72 versus 0.69. A different set of assumed scores could reverse the ranking, which is why the recommendation remains conditional.

4.2. Simplified Decision Procedure

The proposed procedure is (1) confirm the task, trust target, consent, and execution authority; (2) translate stated requirements into hard constraints and weighted preferences; (3) generate retention, postponement, maintenance, repair, refurbishment, and purchase actions where relevant; (4) retrieve evidence and label source, scope, date, and uncertainty; (5) eliminate only actions that violate hard constraints; (6) score remaining actions and perform sensitivity analysis; (7) test incentive conflicts and apply audit rules; (8) generate a faithful contrastive explanation; (9) request confirmation when required; and (10) log the recommendation, evidence, authorization, outcome, and recourse pathway.

5. Conceptual Model and Propositions

Figure 1 distinguishes design features, perceptions, trust, delegation, behavior, and outcomes. The four architectural features are not trust itself. They provide evidence from which consumers infer ability, integrity, and benevolence. These perceptions contribute to overall trust, which affects willingness to delegate. Delegation does not guarantee reliance, and reliance does not guarantee a good outcome. Recommendation correctness, consumer comprehension, and authorization determine whether reliance is appropriate.

Calibrated trust is not the same as higher perceived trustworthiness. Perceived ability, integrity, and benevolence, and overall trust, describe the level of consumers' beliefs and confidence. Calibration describes the fit between that level and the agent's actual trustworthiness, as reflected in recommendation appropriateness, constraint compliance, and governance performance [7]. A feature that raises perceived trustworthiness therefore improves calibration only if the agent's performance warrants the increase or if consumers become better able to discriminate between sound and unsound recommendations; otherwise, it produces overreliance, and a feature that lowers trust may be appropriate if it reveals real limits [8,9]. In the model, calibration is not an additional node. It is evaluated by comparing trust, delegation, and reliance with recommendation performance at the level of individual decisions. Appropriate reliance is the share of recommendations that consumers accept when they are correct or well supported plus the share they question or reject when they are unsupported, conflicted, or outside authorization. Overreliance (accepting poor recommendations) and under-reliance (rejecting sound ones) are the two calibration failures, and the gap between stated trust and observed recommendation accuracy is a second indicator. Empirical tests of Propositions 1–4 should therefore report calibration measures alongside perceptions.

Boundary conditions operate at multiple stages. Initial trust in AI, product knowledge, environmental involvement, cognitive load, budget pressure, product category, decision reversibility, privacy concern, certification coverage, data quality, compensation model, and

perceived conflict of interest may strengthen, weaken, or reverse predicted effects. Excessive trust is more likely when explanations are persuasive but unfaithful, uncertainty is hidden, or operator reputation substitutes for evidence. Insufficient trust is more likely when valid assurance is too complex to understand or prior failures dominate current evidence.

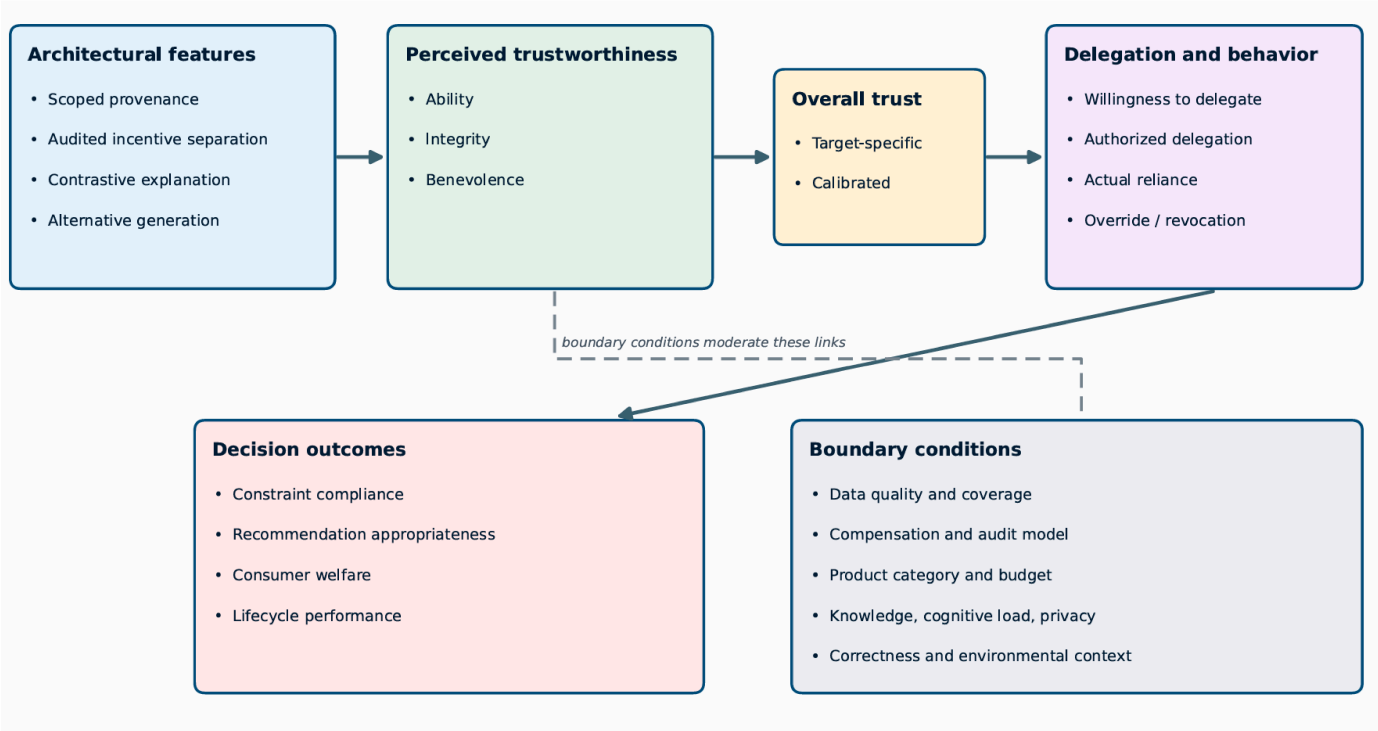


Figure 1. Conceptual model linking architecture, trust perceptions, delegation, reliance, and outcomes.

5.1. Scoped Provenance Verification

Provenance mechanisms must be differentiated. A DPP is a product-linked data set whose required contents depend on applicable product-specific rules. A certification database records the status and scope of a scheme. Identifier standards connect an item to information but do not independently assure that information. A public or permissioned ledger records transactions or claims but does not guarantee the truth, completeness, or comparability of the underlying data. Authentication therefore establishes that a record came from an identified source and has not been altered; it does not by itself establish overall sustainability.

Verification can strengthen perceived ability by improving evidence retrieval and perceived integrity by constraining unsupported claims. The effect should be limited to the attribute and assurance scope actually verified. It also depends on registry independence, coverage, currency, data quality, comparability, and the consumer's understanding of the process. Recommendation accuracy is a separate outcome and must be evaluated against an appropriate benchmark.

Proposition 1. When an agent identifies the source, scope, currency, and uncertainty of independent verification, verified product attributes will be perceived as more credible than equivalent merchant-only claims; this credibility will strengthen perceived ability and integrity when registry coverage and data quality are adequate.

5.2. Disclosure and Incentive Separation

Commercial disclosure and incentive alignment are distinct. A commission disclosure can alert consumers to a conflict but does not remove the conflict. Depending on salience

and interpretation, disclosure may reduce trust, increase trust through perceived honesty, or have no effect. Stronger evidence of integrity combines clear disclosure with enforceable separation of advertising from utility scoring, logged ranking criteria, conflict testing, and independent audit.

The relevant signal is not transparency alone but accountable constraint. An operator that permits examination of whether payments affected rankings accepts exposure to detection and sanction. This exposure may make imitation more costly for an operator that intends to steer recommendations. Consumers must nevertheless understand what was disclosed and which entity receives revenue.

Proposition 2. Disclosed commercial ties combined with auditable separation of commercial payments from consumer utility scoring will strengthen perceived operator integrity and reduce perceived moral hazard relative to disclosure alone or opaque ranking; the effect will depend on disclosure clarity, audit credibility, and the consumer's interpretation of the conflict.

5.3. Faithful Contrastive Explanation

Contrastive explanation answers why the recommended action was preferred to a salient alternative. It can support perceived ability by showing that the agent considered the user's goals and relevant trade-offs. However, explanation quality has at least three dimensions: fidelity to the actual decision process, comprehensibility, and actionability. A fluent post hoc rationale that does not reflect the scoring process may increase confidence while decreasing calibration.

Decision alignment is defined as the degree to which a recommendation and any executed action conform to the consumer's authorized goals and constraints. Recommendation appropriateness additionally considers evidence quality and whether the recommended action is reasonable given the available alternatives. These constructs replace transaction conversion fidelity, because a successful interaction may appropriately end without a purchase.

Two effects of explanation must be kept apart. The first is perceptual: a faithful contrastive explanation can make the agent's reasoning easier to follow and thereby raise perceived ability. The second is objective: whether the recommendation and any executed action actually conform to the consumer's authorized goals and constraints. Greater comprehension does not by itself improve the second outcome, because an explanation can make a poor recommendation look persuasive [7–9]. Explanation can improve decision alignment only through an error-correction pathway. A faithful explanation exposes the decisive criteria, assumptions, evidence gaps, and constraints applied; this allows the consumer to notice a violated constraint or mis-specified preference, correct the input, confirm the action, or withhold authorization before execution. Alignment therefore improves when explanation is faithful, actionable, and paired with confirmation and recourse (Layer 5 of Table 1), not when explanation merely increases confidence. Without that pathway, the model predicts a perceptual gain with no change, or even a decline, in objective alignment.

Proposition 3(a). Faithful, comprehensible contrastive explanations will strengthen perceived agent ability relative to non-contrastive summaries such that objective decision alignment will improve only through an error correction pathway: when (i) explanations are faithful and actionable, (ii) they allow consumers to detect violated constraints or mis-specified preferences and to correct inputs.

Proposition 3(b). Both effects will weaken when cognitive burden is high, the explanation is unfaithful, or the underlying evidence is unreliable.

5.4. Conditional Non-Purchase Signal

Retention, postponement, repair, and refurbishment should not be collapsed into a single non-purchase category. Retention and postponement avoid an immediate transaction. Repair is a service transaction that may generate parts or labor revenue. Refurbishment is a product purchase with different warranty, performance, and environmental implications. The agent should compare them as distinct actions.

A justified recommendation to retain an adequate asset or postpone purchase can communicate operator alignment, but only under identifiable conditions. The operator is the sender; the consumer is the receiver; the hidden characteristic is commitment to consumer-authorized objectives. The commercial cost may be foregone commission or margin, while the accountability cost may arise from logs, audits, or contractual penalties. A separating outcome is more plausible when non-purchase advice is systematic across eligible decisions, independently observable, and not offset by undisclosed ancillary revenue. It is less plausible in subscription models where advice improves retention at low cost, or when repair and refurbishment generate comparable revenue.

The advice must also be substantively justified. Recommending repair when repair is unsafe, uneconomic, or environmentally inferior should reduce rather than increase calibrated trust. Consumers may interpret occasional non-purchase advice as a strategic reputation tactic. The predicted mechanism therefore links justified and auditable advice—not the phrase "do not buy" by itself—to perceived benevolence and integrity.

Proposition 4. Justified, repeated, and auditable advice to retain or postpone purchase will strengthen perceived operator benevolence and integrity when it imposes observable opportunity or accountability costs and is not offset by undisclosed ancillary revenue; these perceptions will increase willingness to delegate only when the advice is appropriate to the consumer's goals and product context.

6. Illustrative Applications

The following applications are constructed illustrations, not empirical case studies, system audits, or evidence that any named platform behaves in the specified manner. Documented facts are separated from hypothetical agent behavior. The illustrations show how the framework could be applied and what evidence would be required for a future audit.

6.1. DPP-Enabled Repair Illustration

Regulation (EU) 2024/1781 [17] establishes the ESPR framework and defines a DPP as product-specific data accessible electronically through a data carrier. The data required, the level of granularity, and access rights depend on applicable delegated acts and product groups; the regulation does not mean that every product already has a universal, immutable public record or that every passport contains live device diagnostics. Identifier standards may help connect a product to records, while certification databases and ledgers have different assurance functions.

Directive (EU) 2024/1799 [18] promotes repair but limits key manufacturer obligations to goods, and to the extent, covered by Union legal acts containing repairability requirements. Spare-parts periods and information duties therefore depend on product-specific measures rather than a general ten-year rule. A future shopping agent could use legally available passport and repair information, but it would need separate, authorized diagnostic data to assess the condition of an individual device.

In the illustration, a consumer authorizes access to a phone's battery diagnostic and asks whether replacement is necessary. The agent checks applicable repair documentation, obtains a quote, reports that passport data do not establish component health, and compares repair, refurbishment, and purchase using a stated functional unit and uncertainty

labels. The recommendation remains conditional on diagnosis, price, safety, and expected remaining life.

6.2. Commercial Shopping Assistant Illustration

Amazon introduced Rufus as a generative AI shopping assistant in 2024 and renamed it Alexa for Shopping on 13 May 2026 [19]. Amazon describes the assistant as supporting questions, comparison, recommendation, product discovery, and agentic shopping. Those public descriptions do not establish its internal reward function, prove that it always prioritizes checkout, or demonstrate that it never recommends maintenance. This article therefore makes no categorical claim about the system's training objective or universal behavior.

A reproducible future audit could submit dated, standardized prompts about common product failures; record outputs, ranking order, sponsorship indicators, market, account state, and time; repeat prompts across categories and accounts; and code whether the assistant asks diagnostic questions or presents maintenance, repair, refurbished, and new product alternatives. The hypothetical risk is that a commercially operated assistant may favor catalog alternatives when its ranking and compensation rules are opaque. Proposition 2 predicts that disclosure plus auditable separation would provide stronger evidence of alignment than disclosure alone.

6.3. Patagonia Repair and Resale Illustration

Patagonia's Worn Wear program provides a documented example of resale, and the company's Ironclad Guarantee provides repair, replacement, or refund remedies for qualifying issues. Published terms also state that damage from wear and tear may be repaired for a reasonable charge; the program should not be described as an unconditional free lifetime repair warranty [20,21]. These practices establish the availability of repair and resale pathways, not fixed claims about price, durability, or emissions for every item.

A shopping agent could compare continued use, paid repair, a Worn Wear item, and a new item. It should retrieve current item-specific prices and condition information, identify warranty scope, and avoid assigning an emissions reduction without comparable lifecycle evidence. The illustration demonstrates alternative generation and disclosure, not empirical proof that refurbishment or repair is always the preferred outcome.

7. Sustainability and Outcome Boundaries

The framework separates four outcomes that are often conflated. Consumer savings concern expenditure over an appropriate time horizon. Perceived sustainability is the consumer's subjective assessment. Circularity concerns retained value, reuse, repair, remanufacture, or material cycling. Lifecycle environmental performance concerns impacts measured over a defined functional unit and system boundary. One outcome cannot be inferred automatically from another.

A cost-per-use calculation can support economic comparison but does not establish environmental superiority. For example, a hypothetical USD 135 garment worn 200 times has a purchase cost of USD 0.675 per wear, while a USD 100 garment worn 50 times has a cost of USD 2.00 per wear. The arithmetic is conditional on actual use and excludes care, repair, resale value, and environmental impacts. Durability, microfiber release, and emissions require product-specific evidence rather than assumed multipliers.

Repair or retention is environmentally preferable only under specified conditions. Relevant variables include expected remaining service, repair materials and travel, safety, changes in operating energy or water use, displaced production, end-of-life treatment, and whether the user would otherwise purchase another product. For energy-using products, efficiency gains may sometimes justify replacement. Rebound effects must also be consid-

ered: money saved through repair or refurbishment may finance additional consumption, while lower prices may increase purchase frequency.

Accordingly, the architecture should provide separate economic and environmental scores, disclose when lifecycle evidence is unavailable, and avoid a single sustainability label that conceals trade-offs. Empirical evaluation should compare predicted outcomes with observed use, repair success, return rates, product life, and consumption displacement over time.

8. Discussion and Implications

The theoretical contribution is a staged account of calibrated delegation. Principal–agent theory identifies the incentive conflict; signaling theory specifies what evidence might reveal alignment; the ability–benevolence–integrity model explains consumer inferences; and trustworthy-AI and recommender governance specify design and evaluation requirements. The integration shows why a transparent interface can still be misaligned, why accurate recommendations may not appear trustworthy, and why increased trust is not sufficient evidence of consumer welfare.

The specific extension can be stated relative to the closest studies. Trustworthy-AI and trustworthy-recommender scholarship specifies requirements such as transparency, fairness, explainability, and robustness [5,6], but it generally treats these as properties of the system and does not model the operator's incentive to deviate from the consumer's objectives or ask what evidence would reveal alignment. Research on trust in automation and AI recommendation agents explains how performance, explanation, and reliance relate and warns of overreliance [7–9], but it typically assumes that the agent's objective is aligned with the user's. Sustainable consumption research identifies the informational, social, and habitual barriers to more sustainable choices [4], but it does not examine delegated, agent-mediated decisions in which retention, repair, or refurbishment may conflict with transaction revenue. Conditional non-purchase signaling connects these streams. It treats advice against an immediate purchase as informative about operator alignment only when the sender, hidden characteristic, differential cost, observability, and compensation context are specified, and it links that signal to lifecycle alternatives whose environmental merit must itself be evidenced rather than assumed. The contribution is therefore a conditional extension and refinement of existing perspectives, not a replacement for them, and its value depends on empirical testing.

The analysis extends signaling theory to agentic commerce by locating the sender at the operator level and treating system behavior as a governed, repeated signal. Non-purchase advice is distinctive only when it creates differential costs and is appropriate. This conditional account replaces the stronger but unsupported claim that a single recommendation against purchase proves alignment. It also predicts circumstances in which the same behavior is pooling rather than separating—for example, when both aligned and misaligned subscription services can imitate it cheaply.

For managers, the framework implies that trust should not be optimized as an isolated engagement metric. Operators should measure constraint compliance, recommendation appropriateness, error detection, override behavior, and appropriate reliance. Commercial models may include subscriptions or outcome-based fees, but no model guarantees alignment. Governance must examine whether retention incentives, repair commissions, private-label margins, or advertising payments distort advice. Independent audits and decision logs can test actual behavior across these revenue sources.

For system designers, provenance must be scoped, explanations must be faithful, and autonomy must be bounded. Agents need a graceful failure mode for missing evidence and infeasible constraints. Human confirmation should increase with financial, safety, privacy,

and irreversibility risk. A recommendation and its execution should remain distinct events, with clear revocation and recourse.

For policymakers, existing law and proposed reforms should be separated. The ESPR creates a framework for product-specific DPP requirements, while Directive (EU) 2024/1799 [18] ties important repair obligations to covered products and applicable repairability rules. Proposals for mandatory APIs, compulsory repair queries, or legal classifications of shopping agents would require analysis of proportionality, security, confidential information, small-firm burdens, accessibility, enforcement, and unintended exclusion. Regulators could first require clear commercial role disclosure, retention of auditable decision records for high-impact autonomous transactions, and accessible cancelation and redress.

9. Limitations and Future Research

This article is conceptual and does not validate the architecture, propositions, or causal sequence. The purposive literature synthesis is multidisciplinary but not exhaustive or reproducible as a systematic review. The named applications are illustrative and do not report interviews, observations, prompt audits, or performance tests. Legal and technical conditions will change as delegated acts, product rules, interfaces, and commercial models evolve.

The framework also assumes that consumer preferences can be elicited with sufficient stability to guide action. In practice, preferences may be constructed during the interaction, internally inconsistent, or sensitive to framing. The proposed scoring procedure simplifies moral, distributional, and lifecycle trade-offs that may not be commensurable. Verification infrastructure can contain inaccurate, stale, strategically supplied, or incomparable data. Audits may be constrained by proprietary models and dynamic behavior.

Future experiments should independently manipulate verified versus unverified provenance; disclosure alone versus disclosure plus audited incentive separation; faithful contrastive explanation versus persuasive but unfaithful explanation; and justified non-purchase advice across commission, subscription, and ancillary revenue models. Outcomes should include perceived ability, integrity, benevolence, overall trust, delegation willingness, actual reliance, error detection, constraint compliance, recommendation appropriateness, and calibrated reliance—not purchase conversion alone.

Field and longitudinal studies should test whether initial trust survives recommendation errors, commercial conflicts, failed repairs, and changes in data quality. Reproducible platform audits should use preregistered prompts, dated outputs, multiple accounts, markets, and product categories. Product-specific lifecycle studies should compare repair, refurbishment, retention, and replacement using explicit functional units and displacement assumptions. Research should also examine fairness for uncertified small suppliers and accessibility for consumers who cannot process detailed explanations.

A prototype implementation could test feasibility through adversarial scenarios: no feasible alternative, conflicting certifications, stale passport data, hidden affiliate revenue, unsafe repair, privacy refusal, and revocation immediately before purchase. Such studies would determine whether the proposed controls function as intended and whether consumers understand them.

10. Conclusions

Agentic commerce makes trust consequential because recommendation can become action. This article proposes that consumer trust should emerge from scoped evidence, auditable incentives, faithful explanation, bounded authority, and demonstrated decision quality rather than from interface fluency or purchase completion. The resulting framework

distinguishes design features from ability, integrity, and benevolence perceptions; overall trust; willingness to delegate; actual reliance; and economic and environmental outcomes.

Justified non-purchase advice is a potentially informative but conditional signal. It can support perceived alignment when the operator incurs observable opportunity or accountability costs, ancillary revenue is disclosed, the behavior is systematic and auditable, and the recommendation is appropriate to the product and consumer context. These relationships remain theoretical propositions open to empirical testing. The goal is not maximal trust, but calibrated delegation to shopping agents whose capabilities, limits, incentives, and consequences can be examined.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Davenport, T.; Guha, A.; Grewal, D.; Bressgott, T. How artificial intelligence will change the future of marketing. *J. Acad. Mark. Sci.* **2020**, *48*, 24–42. [CrossRef]
2. Chang, W.; Park, J. A comparative study on the effect of ChatGPT recommendation and AI recommender systems on the formation of a consideration set. *J. Retail. Consum. Serv.* **2024**, *78*, 103743. [CrossRef]
3. Portugal, I.D.S.; Alencar, P.; Cowan, D. An agentic AI-based multi-agent framework for recommender system. In Proceedings of the 2024 IEEE International Conference on Big Data, Washington, DC, USA, 15–18 December 2024; pp. 5375–5382. [CrossRef]
4. White, K.; Habib, R.; Hardisty, D.J. How to SHIFT consumer behaviors to be more sustainable: A literature review and guide for marketers. *J. Mark.* **2019**, *83*, 22–49. [CrossRef]
5. Thiebes, S.; Lins, S.; Sunyaev, A. Trustworthy artificial intelligence. *Electron. Mark.* **2021**, *31*, 447–464. [CrossRef]
6. Wang, S.; Zhang, X.; Wang, Y.; Liu, H.; Ricci, F. Trustworthy recommender systems. *ACM Trans. Intell. Syst. Technol.* **2023**, *14*, 67. [CrossRef]
7. Lee, J.D.; See, K.A. Trust in automation: Designing for appropriate reliance. *Hum. Factors* **2004**, *46*, 50–80. [CrossRef] [PubMed]
8. Glikson, E.; Woolley, A.W. Human trust in artificial intelligence: Review of empirical research. *Acad. Manag. Ann.* **2020**, *14*, 627–660. [CrossRef]
9. Klingbeil, A.; Grützner, C.; Schreck, P. Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. *Comput. Hum. Behav.* **2024**, *160*, 108352. [CrossRef]
10. MacInnis, D.J. A framework for conceptual contributions in marketing. *J. Mark.* **2011**, *75*, 136–154. [CrossRef]
11. Jaakkola, E. Designing conceptual articles: Four approaches. *AMS Rev.* **2020**, *10*, 18–26. [CrossRef]
12. Jensen, M.C.; Meckling, W.H. Theory of the firm: Managerial behavior, agency costs and ownership structure. *J. Financ. Econ.* **1976**, *3*, 305–360. [CrossRef]
13. Eisenhardt, K.M. Agency theory: An assessment and review. *Acad. Manag. Rev.* **1989**, *14*, 57–74. [CrossRef]
14. Spence, M. Job market signaling. *Q. J. Econ.* **1973**, *87*, 355–374. [CrossRef]
15. Connelly, B.L.; Certo, S.T.; Ireland, R.D.; Reutzel, R.D. Signaling theory: A review and assessment. *J. Manag.* **2011**, *37*, 39–67. [CrossRef]
16. Mayer, R.C.; Davis, J.H.; Schoorman, F.D. An integrative model of organizational trust. *Acad. Manag. Rev.* **1995**, *20*, 709–734. [CrossRef]
17. European Parliament and Council. Regulation (EU) 2024/1781 of 13 June 2024 establishing a framework for the setting of ecodesign requirements for sustainable products. *Off. J. Eur. Union* **2024**, 32024R1781.
18. European Parliament and Council. Directive (EU) 2024/1799 of 13 June 2024 on common rules promoting the repair of goods. *Off. J. Eur. Union* **2024**, 32024L1799.
19. Amazon. (2024/2026). Amazon's Next-Gen AI Assistant for Shopping Is Now Even Smarter, More Capable, and More [Updated to Note the 13 May 2026 Renaming to Alexa for Shopping]. About Amazon. Available online: https://www.aboutamazon.com/news/retail/amazon-rufus-ai-assistant-personalized-shopping-features (accessed on 12 August 2016).

20. Patagonia. Ironclad Guarantee. Patagonia Customer Service. Available online: https://help.patagonia.com/s/article/Ironclad-Guarantee (accessed on 12 August 2026).
21. Patagonia. Worn Wear Frequently Asked Questions. Patagonia Worn Wear. Available online: https://help.patagonia.com/s/topic/0TO1L0000002yyKWAQ/worn-wear (accessed on 12 August 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.