Article

Not peer-reviewed version

Personal Intelligence: Toward a User-Governed Preference Substrate for the Age of Agentic AI

Gabriel Axel Montes *

Posted Date: 20 March 2026

doi: 10.20944/preprints202603.1627.v1

Keywords: personal intelligence; preference learning; taste graph; knowledge graphs; privacypreserving machine learning; user modelling; agentic AI; bounded delegation; local-first computing; platform economics; epistemic decay; selective disclosure; cognitive integrity; cognitive security; alignment; safety

 doi:10.20944/preprints202603.1627.v1

Article

Personal Intelligence: Toward a User-Governed Preference Substrate for the Age of Agentic AI

Gabriel Axel Montes

Neural Axis, ASI Institute, Florida Atlantic University, USA; gabriel@neuralaxis.org

Abstract

The rapid integration of large language models into commerce, productivity, and daily decision-making is creating a new class of AI intermediary—one that not only recommends but transacts, delegates, and acts on behalf of users. This concentration of influence raises a structural question: who governs the representation of what a user wants? We introduce *Personal Intelligence* as a research and design programme whose primary objective is *preference continuity under user control*. We propose a core construct, the **Personal Preference Substrate** (PPS): a user-governed, evolving, portable representation of preferences, constraints, context, and provenance. We argue that PPS is naturally expressible as a multi-relational graph — a *taste graph*—because preferences are compositional, contextual, and evidence-dependent. The paper develops seven interlocking contributions: (1) a motivation grounded in platform incentive decay, the agentic-commerce zeitgeist, and the epistemic fragility of current AI systems; (2) a formal definition of PPS with axioms for user governance, provenance, temporal awareness, epistemic separation, and boundary-conditional disclosure; (3) a hybrid learning architecture combining passive behavioural signals with active micro-queries and signal-gated refresh; (4) a boundary-layer framework addressing egress control, ingress validation, incentive insulation, bounded delegation, and adversarial resilience; (5) a privacy-preserving collective improvement model spanning federated learning, secure aggregation, and differential privacy; (6) a portability model distinguishing representation, state, and capability portability with scoped identity; and (7) an evaluation framework measuring preference fidelity, autonomy preservation, epistemic quality, and regret reduction rather than engagement alone. We ground the framework in three pluripotent use cases—commerce, tool mediation, and epistemic filtering—and conclude with a research agenda identifying open problems in graph-based preference representation, on-device inference, portable identity, and incentive-aligned collective learning.

Keywords: personal intelligence; preference learning; taste graph; knowledge graphs; privacy-preserving machine learning; user modelling; agentic AI; bounded delegation; local-first computing; platform economics; epistemic decay; selective disclosure; cognitive integrity; cognitive security; alignment; safety

---

1. Introduction

1.1. The Shift from Recommend to Act

We are entering an era in which conversational AI systems do not merely recommend—they transact. In-chat checkout, agentic browsing, and tool-integrated assistants compress discovery, comparison, and execution into a single interface. Major platforms are now deploying and scaling conversational shopping assistants in production [1,2]; protocol and platform work from Google and OpenAI is formalising agent-mediated shopping and checkout flows [3–5]; and analysts increasingly expect agent-mediated purchasing to become materially relevant across both consumer and B2B environments [6,7].

This pattern extends beyond commerce. Tool-integrated assistants now manage email, calendar, files, and enterprise workflows through standardised protocols [8]. On-device models run with

 doi:10.20944/preprints202603.1627.v1

acceptable latency on consumer hardware [9–11]. The compression of the user journey—from intent to action—is becoming the default interaction pattern across domains.

Table 1. Operational definitions of key terms used in this paper.

Term	Operational Definition
Personal Intelligence	A research and design programme for preference continuity under user control.
Personal Preference Substrate (PPS)	A user-governed, evolving, portable representation of preference-relevant state.
Taste graph	The graph view of PPS: a multi-relational structure over preferences, constraints, context, evidence, and provenance.
Preference continuity	The property that a system remains coherent with a user's evolving preferences over time, while remaining corrigible.
Boundary layer	The operating membrane mediating egress, ingress, and delegation between PPS and external systems.
Micro-query	A short, contextual question designed to disambiguate decision-critical constraints with minimal attention cost.
Epistemic decay	The fact that preference-relevant claims have finite validity windows and conditions for invalidation.
Mutation trigger	An event-driven condition that forces re-evaluation of a taste-graph edge (e.g., price change, firmware update, life-stage shift).
Bounded delegation	A privilege granted to an AI agent in bounded form—with consent, least privilege, and reversibility — enforced at the boundary.
Selective disclosure	Sharing only the PPS subgraph relevant to a specific task, using scoped permissions.
Costly signal	Evidence whose generation requires real expenditure (money, time, effort), making it resistant to fabrication at scale.
Cognitive integrity	The evolving capacity of a bounded system to maintain calibrated attention, trust, and decision-making under pressure. In PI, applied at the individual scale: the user's preserved ability to notice, inspect, correct, and direct AI-mediated inferences.

1.2. The Structural Problem: Preference Capture by Intermediaries

This compression is efficient. It is also dangerous. Whoever governs the model's memory and the transaction path can shape what the user sees, what they believe they want, and what they ultimately do. The incentive structure of two-sided digital platforms makes this outcome predictable rather than hypothetical. Rochet and Tirole [12] formalised the economics of platform intermediation, showing how platforms balance subsidies between user groups. Doctorow [13] popularised the term *enshittification* to describe the lifecycle by which platforms first attract users with quality, then extract value from both users and business customers as lock-in deepens. Klemperer [14] and Farrell and Klemperer [15] demonstrated formally how switching costs and network effects sustain rent extraction even when alternatives exist.

The structural implication is that personalisation, in the absence of user governance, tends to become an instrument of the intermediary's objectives—engagement maximisation, margin optimisation, and lock-in — rather than user welfare. Empirical evidence supports this concern. Allcott et al. [16] found in a large-scale randomised trial that deactivating Facebook improved subjective well-being and reduced political polarisation. Huszár et al. [17] showed that algorithmic amplification on Twitter systematically favoured political content. Zuboff [18] articulated the broader pattern as *surveillance capitalism*: the systematic extraction of behavioural data for prediction and influence.

 doi:10.20944/preprints202603.1627.v1

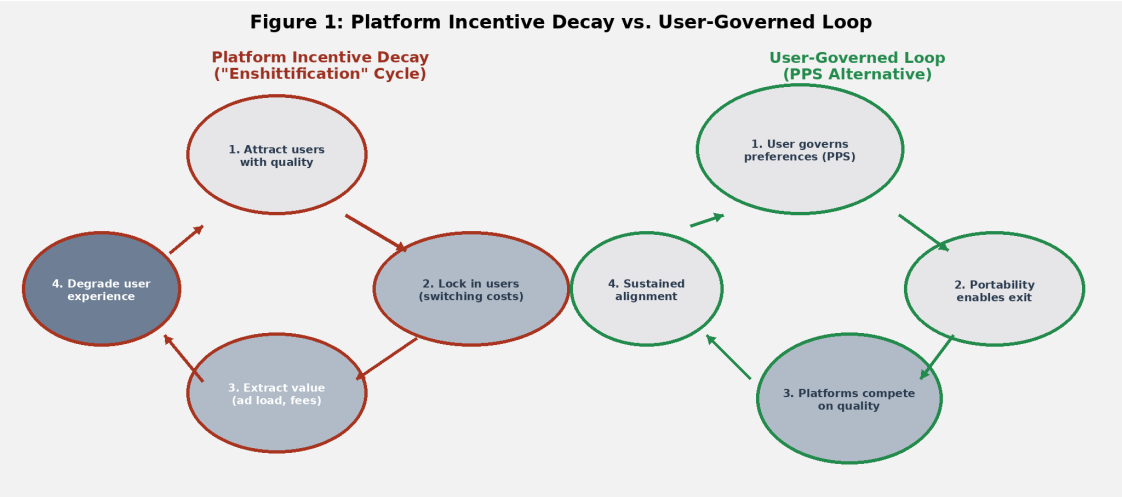

Figure 1. caption.

1.3. Core Proposal

We propose **Personal Intelligence** as a design and research programme defined by a single structural commitment: *the representation of what a person wants should remain under that person's governance.* The core construct is the **Personal Preference Substrate** (PPS)—a user-governed, and where appropriate user-custodied, evolving, portable representation of preferences, constraints, context, and provenance—with a **boundary layer** that mediates all interactions with external models, tools, and data sources.

Personal Intelligence is not a recommender system. It is not a cryptographic identity protocol. It is not an alignment manifesto. It is an *infrastructure proposal*: a neutral, domain-agnostic object around which technical work on preference continuity, privacy, and user agency can be organised.

1.4. Contributions and Scope

The paper makes seven contributions:

1. A motivation grounded in the convergence of incentive decay, agentic AI, and epistemic fragility.
2. A formal definition of PPS with six axioms and a natural graph representation (the taste graph).
3. A hybrid learning architecture combining passive signals, active micro-queries, signal-gated refresh, and variance-preserving explanation.
4. A boundary-layer framework addressing egress, ingress awareness, incentive insulation, and bounded delegation.
5. A privacy-preserving model for collective improvement.
6. A three-layer portability model with scoped identity.
7. An evaluation framework measuring preference fidelity, autonomy, epistemic quality, and regret reduction.

Scope at this stage. This paper does not choose among implementation stacks, scoring functions, question templates, or model architectures. It also does not attempt to solve the general alignment problem. At this stage, it leaves open the optimal graph schema, the best privacy-utility trade-off for a given deployment, and the feasibility of standardised interchange formats. Each technical mechanism is presented as a *candidate family with trade-offs*, not as a prescription.

2. Conceptual Foundations

2.1. Taste as a Structured Object

The word *taste* is commonly treated as aesthetic and subjective. Here we use it in a broader, technically precise sense. Taste encompasses not only what someone likes, but what they avoid; what constraints they habitually carry (budget, time, materials, accessibility); and what contexts modulate

 doi:10.20944/preprints202603.1627.v1

preference (season, occasion, identity, prior regret, trust in a source). Taste is therefore not a static list of preferences. It is a *structure of relationships*: between needs and options, between past experiences and future caution, between identity and acceptable trade-offs.

We distinguish three preference classes that differ in temporal stability and evidential basis:

- **Stable preferences** persist across years and are typically identity-linked: dietary requirements, accessibility needs, ethical commitments, aesthetic sensibilities.
- **Situational preferences** are context-dependent and shift with circumstances: travel plans, project deadlines, seasonal needs, budget constraints.
- **Ephemeral preferences** arise in the moment and may not generalise: a specific craving, a one-time gift search, an exploratory browse.

This structural view aligns with work in knowledge-graph-based recommendation, where heterogeneous entities and multi-relational edges capture the rich side information that flat collaborative-filtering matrices cannot [19,20]. It also resonates with the observation from choice-overload research that the quality of decisions depends not merely on what is available but on how options relate to the decision-maker's constraints and goals [21].

2.2. Preference Continuity, Temporal Semantics, and Epistemic Decay

A central objective for personal intelligence is **preference continuity**: the system should remain coherent with the user over time, while remaining corrigible. Continuity has two sides. *Persistence*: the system should not lose stable constraints and preferences—especially those reflecting safety, accessibility, or recurring needs. *Revision*: the user changes; contexts change; inferences can be wrong. Continuity should therefore include drift, contradiction, and correction.

This framing differs from personalisation systems that optimise immediate engagement. Engagement is frequently used as a revealed-preference proxy, but it is at best incomplete: users can engage with content they do not endorse, click options they later regret, or become captured by novelty and scarcity. Research on engagement-based ranking highlights the risk that optimisation for engagement amplifies divisive or misinformative content even when it undermines user welfare [22,23].

Preference continuity requires attending to **epistemic decay**: the fact that preference-relevant claims have finite validity windows. A user's dietary constraint may be stable for years; their interest in a television series may expire in weeks; a product's availability or firmware behaviour may change overnight. A preference substrate that assigns static confidence to its edges, without modelling expected validity or conditions for re-verification, will accumulate stale beliefs.

We therefore treat three temporal properties as first-class:

- **Validity windows**: each taste-graph edge carries an expected duration of reliability, calibrated to its type.
- **Decay rates**: confidence decreases over time at a rate proportional to the volatility of the underlying claim.
- **Mutation triggers**: explicit conditions that force re-evaluation—a price change exceeding a threshold, a firmware update, a life-stage event such as a dietary diagnosis or a relocation.

This is not merely an implementation detail. It is a design commitment that enables the system to degrade gracefully rather than silently rot. Concurrent industry work on signal-gated recomputation in commerce knowledge systems [24] illustrates the practical urgency of this concern.

2.3. A Cognitive-Science Lens

There is a useful theoretical parallel in cognitive science. Predictive processing and active inference frameworks model biological agents as maintaining generative models that reduce prediction error through both perception and action [25,26]. Under this view, preferences are not static labels but *priors and constraints shaping action selection under uncertainty*. Active inference treats the agent as simultaneously learning about the world and acting to bring the world into alignment with its goals—a formal account of agency that resonates with the design objectives of personal intelligence.

 doi:10.20944/preprints202603.1627.v1

A related observation is that human preferences are often *underspecified until queried*—a phenomenon studied under the label of preference construction [27]. People do not always know what they want until they are asked in context. This motivates the micro-query mechanism (Section 4.2) as more than a data-collection device: it is a *preference elicitation* instrument that respects the constructed nature of preference.

Whether or not one adopts these frameworks wholesale, they provide a mechanism-compatible intuition: a personal intelligence system should treat preference modelling as a continuous inference problem under drift, not as a one-time profiling event. The textbook treatment by Parr, Pezzulo, and Friston [28] provides the formal apparatus; Da Costa et al. [29] explore its implications for agency more broadly.

2.4. Costly Signals and Evidential Grounding

Not all evidence is created equal. Some signals are inherently more reliable because they are costly to generate in the economic sense: producing them requires real expenditure of money, time, or effort. A product return requires that the user purchased the item, received it, used it, and decided it was not worth keeping. A support complaint requires that the user invested time in articulating a problem. Cart abandonment at the payment stage reveals a preference reversal that no amount of browsing data can match.

These costly signals carry far more epistemic weight than marketing copy, curated reviews, or click behaviour [24]. Yet they are currently captured almost exclusively by platforms and are unavailable to users or competing services. PPS proposes to restore user custody of costly signals—enabling preference learning from regret, not just from purchases. This principle generalises beyond commerce: in tool mediation, an action that required reversal is more informative than one that was silently accepted; in information consumption, a user who actively unfollows a source provides a stronger signal than one who passively scrolls past it.

2.5. Cognitive Integrity

Cognitive integrity is the evolving capacity of a bounded system to maintain calibrated attention, trust, and decision-making under pressure [70]. It is the ability to remain coherent enough, reality-linked enough, and self-directing enough to keep learning, choosing, and reorganising without losing the thread of itself. The word *integrity* does the same work here as in structural engineering: a bridge with structural integrity can bear load, flex under stress, and remain standing after a storm. It is not rigid. It holds together through change.

This concept applies across scales: a person, a team, an institution, a society each maintains (or fails to maintain) its own cognitive integrity. In the context of Personal Intelligence, we are concerned primarily with the individual scale: the user's preserved capacity to notice, inspect, correct, and direct the AI-mediated systems that increasingly shape their decisions. Concretely, a PPS-based system preserves cognitive integrity when the user can:

- **Notice** when the system has inferred something about them.
- **Inspect** the basis for that inference (provenance, evidence, confidence).
- **Correct** or override the inference without penalty.
- **Resist** manipulation via defaults, dark patterns, or undisclosed objective shifts.

Cognitive integrity is not a moral stance or an ideological commitment. It is a design constraint: a PPS-based system should preserve the user's capacity for autonomous judgment as a *functional requirement*. As AI becomes a cognitive prosthesis for daily life, the question sharpens: what remains recognisably the user's own judgment, and what is quietly being authored by the system around them? This connects to the broader literature on dark patterns [30,31] and the documented costs of engagement-optimised information environments [16,17].

 doi:10.20944/preprints202603.1627.v1

3. The Personal Preference Substrate

3.1. Definition

We define the **Personal Preference Substrate (PPS)** as follows:

> *A user-governed, evolving representation of preference-relevant state, spanning: (i) stable and situational preferences, (ii) constraints and aversions, (iii) inferred intent and goals, (iv) context (time, location, device, occasion) when relevant and consented, (v) provenance—metadata describing where each belief came from, why it exists, how uncertain it is, and when it should be re-evaluated—and (vi) references to world-claims, tagged as external and kept distinct from stable preference facts.*

PPS is intentionally neutral with respect to implementation. One could realise PPS as text summaries, structured attributes, embeddings, or a composite. Our central representational claim is that PPS is *often naturally expressible as a graph*. PPS has three degrees of freedom that deployments need to resolve: the *representation* (how preference state is encoded), the *learning policy* (how the system updates from evidence), and the *privacy posture* (where on the privacy spectrum the deployment sits).

3.2. Axioms

Six properties are axiomatic:

1. **User governance.** The user can inspect, correct, delete, rollback, and partition the substrate. User governance means control over update, disclosure, and portability—not merely encryption.
2. **Evolution.** The substrate updates with new evidence and can represent drift, contradiction, and correction.
3. **Evidence-awareness.** The substrate stores not only *what the system believes*, but *why* (explicit statement, repeated behaviour, one-off event, regret episode, costly signal).
4. **Temporal awareness.** Each edge carries an expected validity window, a decay rate, and explicit invalidation conditions (mutation triggers). The system distinguishes stable constraints from volatile inferences.
5. **Boundary-conditionality.** Disclosure of any part of PPS is mediated by a boundary layer (Section 5).
6. **Epistemic separation.** PPS should represent uncertain or contested world-claims as *external to preference edges*—or explicitly tagged as such—to avoid contaminating the preference substrate with unreliable factual assertions.

3.3. The Graph View: PPS as a Taste Graph

We call the graph view of PPS—the representation of preference-relevant state as a multi-relational structure—a **taste graph**.

Why a graph? Because preference is inherently relational and multi-typed. A preference links an *entity* (an item attribute, a brand, a style) to a *context* (season, use case, budget), to an *evidence source* (explicit statement, repeated purchases, a negative return), and to a *confidence level*. Preferences interact: "likes minimalist clothing" is moderated by "sensitive to wool itch," "prefers ethical sourcing," and "needs machine-washable." The *why* matters as much as the *what*; relational structure makes explanation more faithful than post-hoc rationalisation of opaque embeddings.

Graph data models—knowledge graphs, property graphs, and hypergraphs — are well-established in information retrieval and recommendation [19,32]. Wang et al. [20] showed that attentive propagation over knowledge-graph paths improves recommendation by capturing multi-hop relational context. He et al. [33] demonstrated that even simplified graph convolution (LightGCN) outperforms complex deep models for collaborative filtering. Wu et al. [34] and Gao et al. [35] survey the broader landscape of graph neural networks in recommender systems.

A taste graph does not claim novelty in using graphs *per se*. The novelty is the combination of: (i) user governance, (ii) provenance, uncertainty, and temporal decay as first-class citizens, (iii)

 doi:10.20944/preprints202603.1627.v1

epistemic separation between preference edges and world-claim references, and (iv) portability across AI systems as a design goal.

We note that alternative representations exist and may be preferable in specific contexts. Embedding-based memories offer compact, continuous representations but sacrifice interpretability and provenance. Event-log architectures provide complete traces (append-only interaction logs) but require reconstruction for inference. Relational tables are well-understood but struggle with the heterogeneity and multi-hop reasoning that preference structures demand. Hybrid approaches—graph structure for relational reasoning, embeddings for similarity, event logs for provenance—are a promising direction that we leave to future work.

3.4. Graph Anatomy (Figure 2)

A minimal but expressive taste graph includes:

- **Nodes.** Entities such as products, attributes, activities, constraints ("budget under £80"), goals ("reduce returns"), evidence events, and latent constructs such as style clusters.
- **Edges / Relations.** Typed relations: *prefers*, *avoids*, *is-for-context*, *was-true-until*, *derived-from-evidence*, *contradicted-by*, *requires-confirmation*, *delegation-allowed-for*, *references-claim*.
- **Provenance attributes.** Source of the relation (declared, inferred, imported, costly-signal), timestamp, context, and privacy sensitivity.
- **Uncertainty.** Confidence intervals, calibrated probabilities, or qualitative levels with monotonic semantics.
- **Temporality.** Validity windows, decay rates, and explicit mutation triggers—e.g., "re-verify on price change > 10%," "expires after 90 days unless reinforced," "invalidate on relocation."

Hypergraph representations can capture higher-order relationships — e.g., a preference that holds only under a conjunction of conditions [35]. We leave open which formalism is optimal; the paper argues for *why graphness is valuable*, not for a specific schema.

Figure 2: Taste Graph Anatomy
Nodes, Typed Edges, Provenance, and Uncertainty

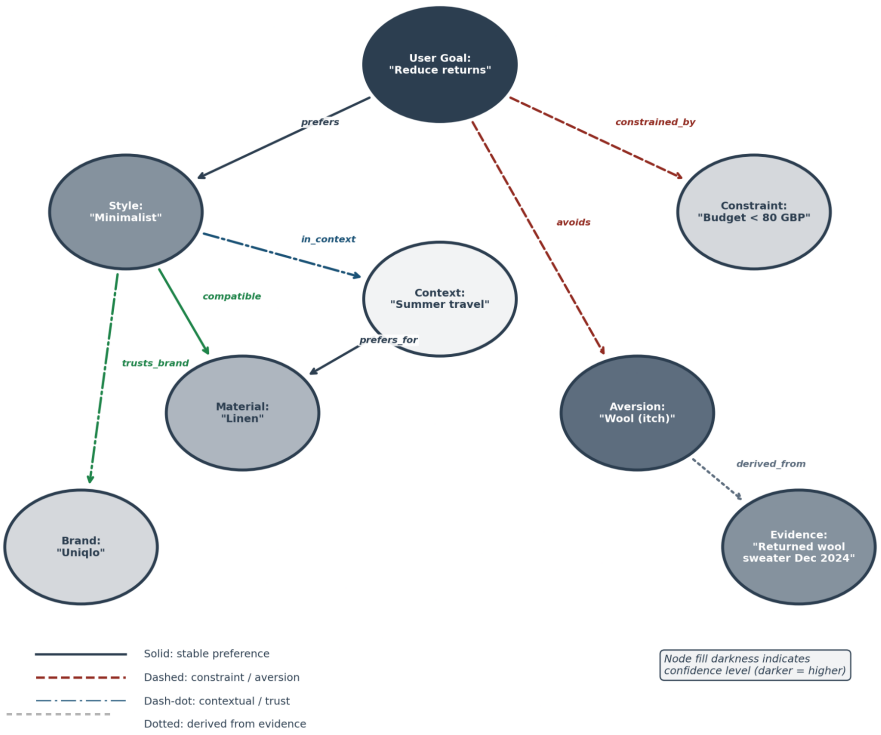

Figure 2. caption.

 doi:10.20944/preprints202603.1627.v1

3.5. Dynamic Identity: Scoped Personas and Selective Disclosure (Figure 3)

A single PPS may contain preference state relevant to many domains: commerce, health, work, private exploration. These domains may involve different levels of sensitivity, different audiences, and different norms around disclosure. A user may want their shopping preferences available to a commerce assistant while keeping health-related constraints invisible to it—and vice versa.

We propose that PPS support **scoped subgraphs** (or *personas*) that partition preference state by domain or context. The boundary layer (Section 5) can then generate **task-scoped presentations** for external AI applications without exposing the full substrate or creating a universal, correlatable identifier.

Three design dimensions matter:

- **Scope separation.** Preferences within a scope (e.g., "commerce") are isolated from preferences in another scope (e.g., "health") unless the user explicitly links them.
- **Identifier modes.** Different interactions may use different identifier strategies: a stable pairwise pseudonym for a trusted merchant, an ephemeral identifier for anonymous exploration, or no identifier at all. This can draw on established identity patterns—for example, pairwise subject identifiers and emerging selective-disclosure credential formats [36–38]—without committing to any single standard or deployment stack.
- **Selective disclosure.** Each interaction receives only the minimum PPS subgraph necessary for the task—a principle that existing work on verifiable credentials has formalised [39].

An important caveat: minimum disclosure does not guarantee anonymity. Even a small, carefully curated subgraph can be re-identifying if the combination of attributes is sufficiently unique. PPS deployments may therefore provide *cumulative disclosure awareness*—for example, a user-visible disclosure record or trace that captures what was shared, with whom, under which scope, and for what purpose. Such records help users inspect cumulative exposure over time and tighten future disclosure policies when warranted. Implementations may differ in granularity (per-recipient summaries versus full event logs) and retention windows; the commitment is not to a specific ledger technology, but to making disclosure an object of governance rather than an invisible side effect.

Figure 3: Scoped PPS and Selective Disclosure Across AI Applications

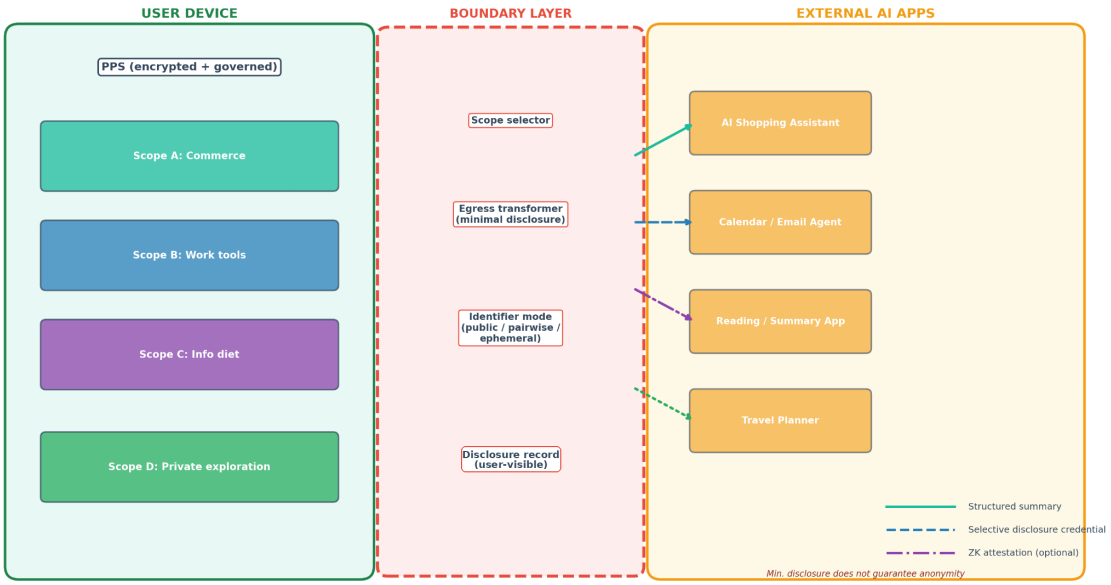

Figure 3. caption.

 doi:10.20944/preprints202603.1627.v1

3.6. Mutation Triggers and Signal-Gated Refresh

Section 2.2 introduced temporal semantics. Here we make the mechanism concrete. Traditional systems re-evaluate on fixed schedules or on every query. Both are wasteful or stale. We propose **signal-gated refresh**: re-evaluation is triggered by *events*, not by clocks or queries.

Three trigger families apply:

- **Preference triggers**: life-stage events (relocation, new job, dietary diagnosis), explicit user corrections, or sustained behavioural drift that crosses a confidence threshold.
- **World-claim triggers**: price movements, firmware updates, product recalls, shifts in review sentiment, or regulatory changes.
- **Boundary triggers**: new threat intelligence (e.g., a tool endpoint found to be adversarial), changes in a service's privacy policy, or revocation of a delegation permission.

When a trigger fires, the affected taste-graph edges are marked for re-verification. The system does not re-derive the entire graph; it targets the subgraph downstream of the trigger. This is computationally economical and epistemically principled: expensive re-evaluation happens when reality has changed, not on a fixed cadence.

4. Learning Dynamics

4.1. Passive Signals and Their Limitations

Personal intelligence systems learn from interaction. The critical question is *how* to learn without defaulting to surveillance maximalism. Passive signals—clicks, dwell time, purchases, skips, returns—are informative but biased. Position bias, exposure bias, and selection bias are well-documented in the learning-to-rank literature. Joachims et al. [40] developed propensity-weighted methods for unbiased learning from biased click data. Li et al. [41] showed how contextual-bandit replay enables unbiased offline evaluation of recommendation policies from logged data.

A PPS-based system should treat passive signals as *evidence for hypotheses*, not as definitive truth. Behaviour does not equal preference: a user may click on content they find alarming, purchase an item under time pressure, or dwell on a page because they are confused rather than interested. The system should maintain this epistemic humility in its update rules.

4.2. Micro-Queries as Active Learning

We propose a complementary channel: **micro-queries**—short, contextual questions that disambiguate intent or constraints precisely when the user is engaged in a decision. This is grounded in active learning for recommender systems. Elahi et al. [42] survey how targeted queries can accelerate cold-start resolution and improve model calibration. The key insight is that a well-timed question can provide more information than hundreds of implicit signals, while respecting the user's attention budget.

We draw an analogy to Socratic dialogue in educational systems. The SCHOLAR system [43] pioneered mixed-initiative questioning in computer-assisted instruction. AutoTutor [44,45] demonstrated that natural-language Socratic dialogue produces learning gains of approximately 0.8 standard deviations across domains. Recent work extends these ideas to LLM-based Socratic tutoring [46,47].

For personal intelligence, micro-queries should satisfy four design constraints: they should be **declinable** (the user can skip without penalty), **localised** (relevant to the current task), **non-manipulative** (they clarify rather than steer), and **directly connectable** to a taste-graph update with provenance.

4.3. The Hybrid Learning Loop (Figure 4)

We formalise the PPS learning loop as six phases:

1. **Signal capture.** Passive behavioural evidence is recorded with provenance tags.
2. **Hypothesis formation.** The system infers tentative preference edges from accumulated evidence.

 doi:10.20944/preprints202603.1627.v1

3. **Micro-query selection.** An active-learning policy selects questions that maximise information gain relative to user attention cost. For high-stakes contexts (health, finance), the system preferentially queries rather than infers.
4. **PPS update.** New evidence (passive or active) updates the taste graph with confidence, provenance, and validity windows.
5. **Inference and explanation.** The updated PPS generates recommendations, plans, or actions—with an accompanying explanation of *why*.
6. **Feedback and correction.** User response (acceptance, rejection, explicit correction) feeds back into the loop. Counterfactual evaluation methods [40,41] enable offline assessment of policy changes.

Figure 4: PPS Hybrid Learning Loop

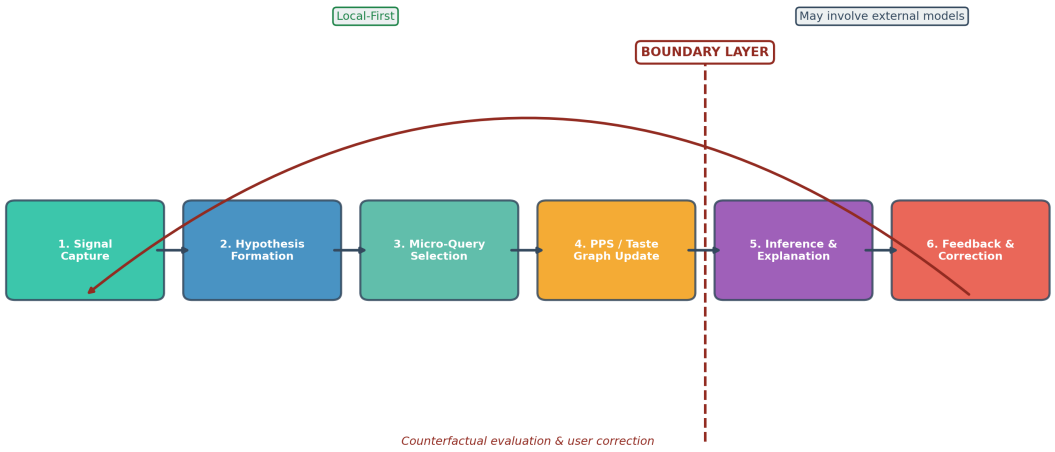

Figure 4. caption.

4.4. Evidentiary Weighting and Costly Signals

Not all updates should carry equal weight. We propose a non-prescriptive principle: **weight updates by evidence reliability**. Costly signals (returns, explicit corrections, support complaints) outrank passive signals (dwell time, scroll depth). Declared preferences outrank inferred preferences. Recent evidence outranks stale evidence, unless the stale evidence has a long validity window.

This is not a formal weighting function—it is a design heuristic that implementations should calibrate. The contribution is the principle: PPS should have an *evidence hierarchy*, and that hierarchy should be inspectable by the user.

4.5. Signal-Gated Refresh

Section 3.6 introduced mutation triggers. In the learning loop, signal-gated refresh means that heavy re-evaluation happens when triggers fire—not on every query and not only on a clock. This applies to preference refresh (drift detection), world-claim refresh (external changes), and boundary refresh (new threat signals). The system watches for signals that reality has changed and re-verifies the affected subgraph when the signals fire.

4.6. Explanations That Preserve Variance

A final design commitment for the learning loop: explanations should not collapse contested evidence into a single confident narrative. When the evidence base is genuinely uncertain or contested, the system should say so. Concretely:

 doi:10.20944/preprints202603.1627.v1

- Explanations should include *why* a recommendation was made, *what would change the recommendation*, and *what the system is uncertain about*.
- When world-claims conflict—e.g., a manufacturer asserts "maintenance-free" but user-review aggregates report recurring servicing costs—the explanation should surface the conflict rather than silently averaging.
- Where dissenting evidence exists, it should be presented alongside the majority view, with provenance.

This commitment resists what one might call the "beige summary" failure mode: a fluent, confident output that averages away the very edge cases and conflicts that matter most for the user's decision. Variance-preserving explanation is not merely a UX nicety; it is a structural requirement for cognitive integrity (Section 2.5).

5. The Boundary Layer

5.1. Definition and Motivation

If PPS is the substrate, the **boundary layer** is the operating membrane. It controls what parts of PPS are revealed to which model or tool, when the system can take action, and what is logged or synchronised. This boundary-first approach is motivated by two realities. First, privacy-preserving on-device computation is increasingly feasible: quantised language models now run on mobile hardware with acceptable latency [9–11], reducing the need to transmit raw user context to servers. Second, tool-integrated LLM systems are exposed to well-characterised attack surfaces [48–50].

Figure 5: Boundary Layer Architecture
Egress Control, Ingress Validation, and Bounded Delegation

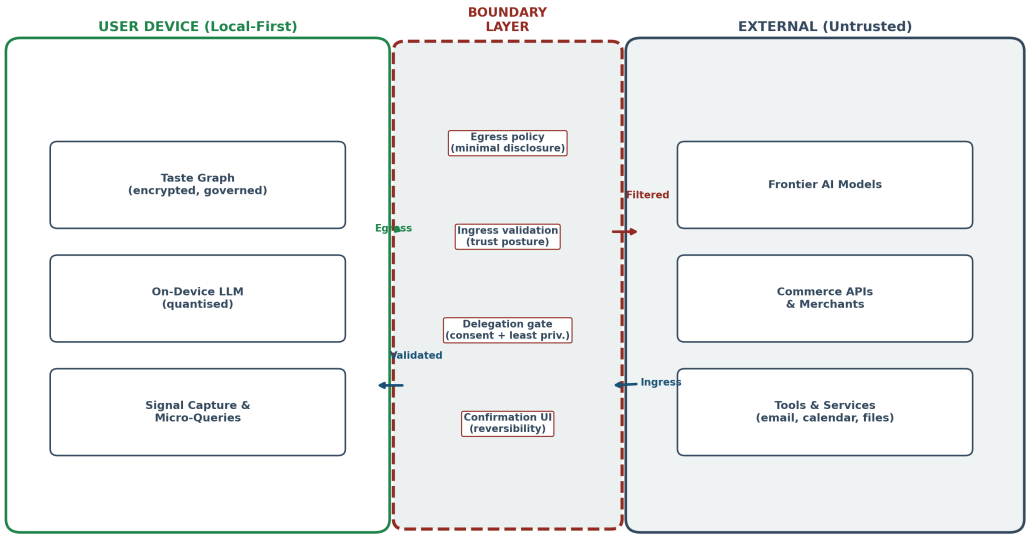

Figure 5. caption.

5.2. Egress Policy: Minimal Disclosure

The boundary layer enforces **minimal disclosure**: for each interaction, only the minimum PPS subgraph required for the task is shared with the external system. Disclosure is conditioned on the task type, the scope (Section 3.5), and the stakes involved. A browsing query may receive only category-level preferences; a purchase delegation may receive specific constraints and budget.

Minimal disclosure is not a one-time decision. Over many interactions, small disclosures accumulate. The boundary layer may therefore support **disclosure accounting**: a running, user-visible summary of what has been shared with each external party (at least at the level of scope and summary), enabling the user to inspect cumulative exposure and to revoke access or tighten policies if warranted.

 doi:10.20944/preprints202603.1627.v1

This connects to the broader problem of cumulative privacy loss in interactive systems, which remains an open research challenge (Section 11).

5.3. Ingress Quality: Why World-Claims Matter

For Personal Intelligence to work optimally, the quality of the *ingress side of the equation* matters: product descriptions, reviews, availability claims, compatibility claims, policy statements, and other external assertions shape the decision environment to which a user's PPS is applied. A system may model preferences faithfully and still fail if the world-claims it consumes are stale, incomplete, or strategically distorted.

Consider a user researching a consumer appliance. An AI assistant, drawing on review sites, professional evaluations, and product pages, recommends a highly-rated model. The recommendation is fluent, well-sourced, and confident. What it omits is a documented failure mode — an internal component that predictably degrades after 12–18 months — known across user forums but largely absent from professional reviews, which test for weeks rather than years, and invisible in marketing copy. This is not a preference error. The system correctly identified what the user wanted. The failure is that preference-faithful reasoning was applied to an epistemically degraded ingress stream.

At this stage, the paper does not attempt to solve general truth adjudication for contested external claims. It instead argues for a more modest design commitment: PPS should be able to *reference* world-claims and their provenance without collapsing them into stable preference facts, and the boundary layer should preserve source, freshness, and conflict information wherever these claims materially affect user-facing inference.

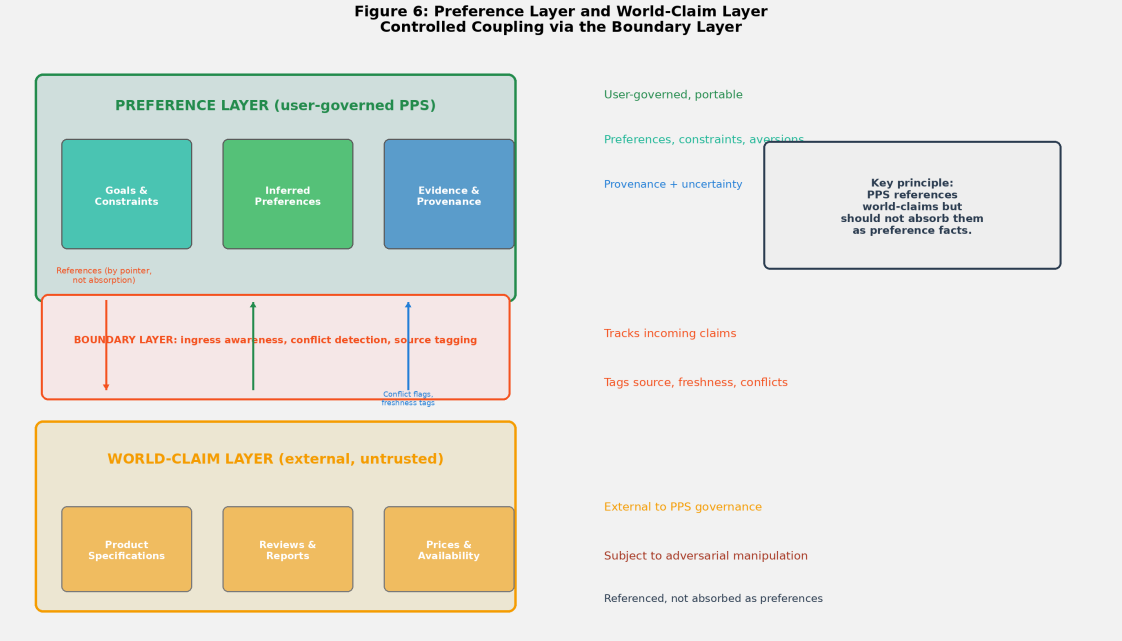

Figure 6. caption.

5.4. Ingress Validation: Contested Claims as Data, Not Instructions

The boundary layer should mediate not only egress but also **ingress**: what enters the user's decision environment from external sources. When a user queries an external commerce API, the returned product descriptions, reviews, and claims enter a zone of contested trustworthiness. Marketing copy may contradict user reports; specifications may conflict with post-purchase experiences. A boundary layer that governs only outbound disclosure but accepts inbound data uncritically will produce preference-faithful recommendations grounded in unreliable world-knowledge—a failure mode that is insidious precisely because the user's preferences are honoured while the claims to which they are applied may not be.

 doi:10.20944/preprints202603.1627.v1

A practical implication is that the boundary layer should maintain a **trust posture** toward external sources: incoming claims can be tagged with source type, freshness, and conflict status. When multiple sources disagree on a factual claim relevant to a user's PPS constraints, the system can surface the conflict rather than silently averaging. The specific resolution strategy (conservative default, user query, weighted source handling, or domain-specific validation) remains outside the scope of this paper at this stage; the commitment is that *relevant conflicts are surfaced, not suppressed*.

Additionally, all untrusted text—web pages, product descriptions, emails, tool responses—should be treated as *data*, not as *instruction*. This is a primary defence against indirect prompt injection [48], and it is a core boundary-layer principle.

5.5. Threat Model

We make the following assumptions about adversaries and trust boundaries:

Assumed threats. (i) Malicious or misleading content injected via external data sources (product descriptions, web pages, emails). (ii) Adversarial tool endpoints that attempt to manipulate PPS through crafted responses or indirect prompt injection [48]. (iii) Compromised or curious cloud servers that may attempt to inspect PPS data in transit or at rest. (iv) Adversarial merchants or content producers who game recommendation systems through SEO, fake reviews, or affiliate structures.

Not assumed. (v) A fully compromised client device (if the device is rooted, local-first guarantees are void). (vi) A user who actively works against their own PPS. (vii) Perfect isolation from all side-channel attacks—we acknowledge residual risks and design for graceful degradation.

Bounded delegation guarantees. Any action taken on behalf of the user through the boundary layer should satisfy three invariants: *consent* (the user has granted permission for the action class), *least privilege* (the delegated agent receives only the minimum PPS subgraph required), and *reversibility* (high-impact actions are either reversible or require explicit confirmation before execution).

5.6. Adversarial Resilience

The OWASP Top 10 for LLM Applications (v1.1) [49] identifies prompt injection and excessive agency as primary risks. Greshake et al. [48] demonstrated that indirect prompt injection can compromise real-world LLM-integrated applications. Datta et al. [50] provide a comprehensive taxonomy of agentic AI security threats.

We do not claim these risks can be eliminated. The design response is to reduce the blast radius: treat all untrusted text as data, not as instruction; force high-impact actions through explicit user confirmation; ensure reversibility and least privilege for all delegated actions; and quarantine new or low-confidence PPS entries from influencing high-stakes decisions.

5.7. Bounded Delegation

A key use case for PPS is **tool mediation**: a user wants an AI assistant to help manage email, calendar, files, and commerce—but with constraints. A PPS-aware boundary layer encodes delegation preferences as taste-graph edges: "may draft emails but not send without confirmation," "may create calendar holds but not invite external attendees," "may browse and summarise but not purchase without explicit approval." These are enforceable patterns: least privilege, confirmations, reversible actions, and revocation.

For high-stakes actions—financial transactions, medical decisions, legal commitments—the boundary layer should enforce a *"cannot afford to guess"* principle: the system should query the user rather than act on inference, regardless of confidence level.

5.8. Incentive Insulation

The enshittification motivation (Section 1.2) highlights a structural problem: platforms that both *model user preferences* and *sell access to user attention* face persistent incentive conflicts. Personal Intelligence does not attempt to settle business-model questions. It proposes, instead, that preference

 doi:10.20944/preprints202603.1627.v1

inference and preference application be designed so that incentive-driven objectives (engagement, margin, affiliate optimisation) are harder to smuggle into the loop without the user noticing.

We use **incentive insulation** to name a design goal: *as far as practicable, keep user-serving inference (preference fidelity, regret reduction) insulated from monetisation signals at decision time*, or else make any such influence *explicit, bounded, and user-governed*. Candidate mechanism families include: (i) separation of concerns in the architecture (distinct modules, APIs, and data-access policies); (ii) user-visible controls that disclose when commercial objectives are allowed to shape ranking; and (iii) higher-assurance verification in specific deployments, such as code signing, TEEs, or ZK-style attestations of data-flow constraints (Section 6.3).

We do not claim a single mechanism is necessary or sufficient. In some deployments, the same organisation will operate both preference-serving and commercial layers; the relevant question is whether the boundary layer makes the influence of monetisation on recommendations *observable and overrideable* by the user, rather than silently conflating "what serves the user" with "what serves the platform."

6. Privacy Architecture

6.1. The Local-First Baseline

PPS should live on the user's devices as the primary copy. Servers, if any, maintain secondary encrypted replicas for synchronisation. This aligns with the local-first software movement articulated by Kleppmann et al. [51], who argue that local-first architectures restore user agency and improve software longevity. Conflict-free replicated data types (CRDTs) [52] provide a formal foundation for multi-device synchronisation without centralised coordination.

6.2. Federated Learning and Differential Privacy

If every PPS is private and local-first, how can systems improve globally? The answer lies in privacy-preserving collective learning. Federated learning [53] keeps raw data on devices while sharing only aggregated model updates. Secure aggregation [54] ensures that the server computing the aggregate cannot inspect individual contributions. Differential privacy [55] provides formal guarantees by calibrating noise to the sensitivity of the query; Apple [56] and Google [57] have deployed these techniques at scale.

We emphasise trade-offs. Differential privacy can reduce memorisation but also utility. Secure aggregation increases protocol complexity. On-device inference reduces server-side visibility but shifts compute constraints to the edge. Federated learning introduces non-IID data distributions and client-availability challenges [53]. The claim is not "one true stack" but that PPS demands an *explicit* privacy architecture, and that a spectrum of credible candidate approaches exists (Figure 7).

Figure 7: Privacy Spectrum for the Personal Preference Substrate

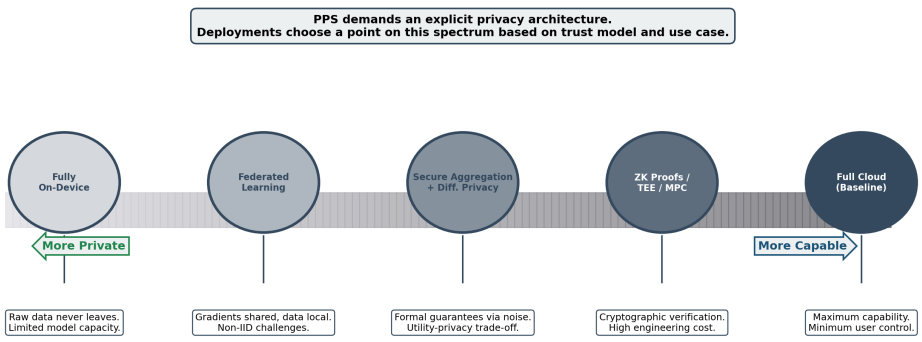

Figure 7. caption.

 doi:10.20944/preprints202603.1627.v1

6.3. Optional Verifiability

Some deployments may seek additional verification that privacy claims are honoured. Zero-knowledge proofs (ZK), trusted execution environments (TEEs), and multi-party computation (MPC) are candidate technologies. Groth [58] and Gabizon et al. [59] provide foundational ZK constructions. Recent work demonstrates the feasibility of mobile ZK proving [60] and ZK-verified federated learning (RiseFL) [61]. We treat these as promising but not mandatory: the engineering and UX costs remain substantial, and they should be evaluated case by case.

It is useful to distinguish three verification goals: verifying *computation correctness* (the model ran the algorithm it claimed to), verifying *data-flow constraints* (revenue data was not visible to the ranking module), and verifying *non-retention* (the receiving system deleted the disclosed subgraph after use). Each has different technical maturity and cost profiles.

6.4. Cumulative Privacy Loss and Disclosure Accounting

Even with minimal disclosure, repeated interactions with external systems create cumulative exposure risk. Each small disclosure—a dietary constraint here, a budget range there—may be individually innocuous but collectively re-identifying. This is a well-known problem in differential privacy, where the privacy budget decreases with each query [55].

One lightweight governance pattern is a **disclosure record**: a user-accessible trace of egress events, tagged by recipient, scope, and (at minimum) a structured summary of content. This enables the user to inspect cumulative exposure and make informed decisions about future disclosures. The disclosure record is not itself a cryptographic primitive; it is a governance aid that makes the abstract notion of cumulative privacy loss more concrete and actionable for the user. Deployments may choose different levels of detail (summaries vs. full payload hashes), retention windows, and user interfaces, depending on threat model and UX constraints.

Quantifying and managing cumulative privacy loss in interactive, long-lived assistant relationships remains an open research problem (Section 11).

7. Portability and Continuity

A crucial purpose of PPS is to be *carried*—a portable, user-governed key that enables personalisation across systems without revealing identity as a prerequisite. Without portability, users become hostage to whichever system controls their preference history. With portability, "exit" becomes credible: users can switch assistants without losing continuity. This aligns with GDPR Article 20's right to data portability [62], the Data Transfer Project [63], and the EU Digital Markets Act's interoperability requirements [64].

7.1. Three Layers of Portability

We distinguish three layers, each with distinct technical requirements:

Representation portability concerns the interchange format for a taste graph. Can PPS be exported in a schema that another system can parse? This requires agreement on node types, edge semantics, and provenance encoding—analogous to the role that vCard plays for contact information or iCalendar for scheduling, but substantially richer. Standardisation efforts here face the classic tension between expressiveness and interoperability.

State portability concerns synchronisation and merge semantics. A user's PPS may reside on multiple devices and evolve concurrently. When two devices produce conflicting preference edges—e.g., one records a new food aversion while the other records a positive restaurant experience—how is the conflict resolved? CRDTs [52] provide one family of solutions; application-level conflict resolution (e.g., "explicit declarations override inferences," "costly signals override passive signals") provides another. The design space is rich and largely unexplored for preference substrates.

Capability portability concerns what the user can *do* with their PPS when they carry it to a new system. Selective disclosure — sharing only the subgraph relevant to a specific task—requires

 doi:10.20944/preprints202603.1627.v1

scoped permissions and verifiable commitments that the receiving system will not retain or redistribute beyond the agreed scope. This is the hardest layer, sitting at the intersection of cryptographic access control and UX design.

7.2. Portability of Scopes and Delegation Permissions

When a user migrates PPS to a new system, they should carry not only preference data but also their *scope definitions* (Section 3.5) and *delegation permissions* (Section 5.6). A scope that separates "work" from "health" preferences should survive migration. A delegation rule that prevents unauthorized sending of email should transfer alongside the email-related preferences.

This requires that scopes and policies be first-class portable objects, not implementation-specific configurations. We leave the specific encoding open; the commitment is that *governance portability is as important as data portability*.

8. Use Cases

We ground PPS in three use cases to demonstrate pluripotence.

8.1. Commerce: Decision Support Under Preference Continuity (Figure 8)

Commerce is densely preference-laden and high-frequency. Users face choice overload [21], bad decisions carry tangible costs, and feedback is concrete (purchases, returns, satisfaction).

Figure 8: Commerce Use Case - End-to-End PPS Flow

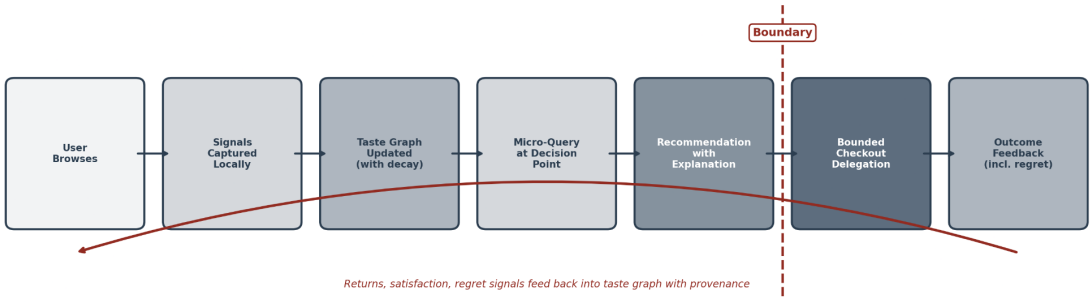

Figure 8. caption.

A PPS-aware commerce system can: (a) remember what the user has looked at, bought, and returned—across platforms; (b) proactively surface relevant items based on taste-graph context (not engagement optimisation); (c) explain *why* an item is recommended, grounded in the user's stated and inferred preferences; (d) track regret signals (returns, complaints) and update the graph accordingly; and (e) negotiate on the user's behalf within bounded delegation constraints.

When world-claims are contested—a manufacturer asserts "low maintenance" while user-review aggregates report frequent part replacements—a PPS-aware system surfaces the conflict with evidence pointers rather than averaging it into a single rating. This is where the epistemic separation axiom (Section 3.2) and the ingress validation mechanism (Section 5.4) become concrete: the system applies the user's PPS constraints to world-claims that have been tagged with conflict status, producing a recommendation that is both preference-faithful and epistemically honest.

Critically, because PPS remains under user governance, the most informative commerce signals—returns, post-purchase regret, cart abandonment, support complaints—become available for the user's own learning, rather than being captured exclusively by platforms. PPS restores user custody of costly signals, enabling preference learning from regret—not just from purchases. The boundary layer and privacy architecture (Sections 5 and 6) ensure that when users contribute such signals to collective improvement, they do so voluntarily, with differential-privacy guarantees.

 doi:10.20944/preprints202603.1627.v1

8.2. Tool Mediation and Bounded Delegation

As AI assistants gain access to email, calendars, files, and APIs — whether through the Model Context Protocol [8] or proprietary integrations—the question of safe delegation intensifies. PPS can encode a user's delegation policy as a set of typed constraints in the taste graph. The boundary layer enforces these constraints at interaction time.

For high-stakes actions, the "cannot afford to guess" principle (Section 5.6) applies: the system queries the user rather than acting on inference. For routine actions within established delegation bounds, the system can act autonomously—but always with a reversibility guarantee and a user-visible action log. This is alignment-adjacent without requiring moral philosophy: it is interface safety, implemented as verifiable constraints.

8.3. Epistemic Filtering and Information Diet

A third use case addresses what we might call *epistemic continuity*: the coherence of one's information environment with one's goals and values. Current content-recommendation systems optimise for engagement, with documented costs to well-being [16] and epistemic quality [17,22]. A PPS-mediated information filter could optimise for the user's stated objectives—depth over breadth, accuracy over novelty, challenge over confirmation—adjustable through the taste graph and enforced at the boundary.

Crucially, epistemic filtering should resist the "beige summary" failure mode identified in Section 4.6. When evidence on a topic is genuinely contested—a medical treatment with mixed trial results, a policy proposal with legitimate arguments on both sides—the system should present the *variance* rather than collapsing it into a single confident summary. This is variance-preserving explanation applied to the information-diet use case: the user's cognitive integrity is served by seeing the shape of disagreement, not by having it smoothed away.

9. Evaluation

A PPS-based system should be evaluated against its stated objective — preference continuity—rather than default metrics like click-through rate. We propose a five-layer evaluation framework:

9.1. Preference Fidelity

Does the system's model of the user match the user's revealed and stated preferences? Offline evaluation with bias correction using contextual-bandit replay [41] or counterfactual estimators [40] provides one approach. Longitudinal studies measuring drift between system predictions and user corrections provide another. Key metrics include calibration (does confidence correlate with correctness?) and edit latency (how quickly can users correct wrong inferences?).

9.2. Regret Reduction

In commerce: returns, exchanges, post-purchase dissatisfaction. In tool mediation: actions that required reversal. In epistemic filtering: user-reported alignment of information diet with goals. Regret reduction is a more meaningful metric than click-through rate because it measures *outcome quality*, not *engagement quantity*.

9.3. Autonomy and Cognitive Integrity Metrics

How often do users inspect, edit, or delete PPS entries? Can users successfully correct wrong inferences? Does user control correlate with trust and continued use? Additional proxies for cognitive integrity include: measurement of excessive steering (does the system nudge users toward options they did not initially consider, and do they later regret those nudges?), detection of undisclosed objective shifts, and assessment of dependence or overreliance signals.

9.4. Epistemic Quality Metrics

Does the system accurately represent uncertainty? How often are genuine conflicts surfaced versus suppressed? What is the quality of evidence traces (can the user follow a recommendation

 doi:10.20944/preprints202603.1627.v1

back to its sources)? Do variance-preserving explanations improve decision quality compared to confident-but-lossy summaries? These metrics are novel and challenging to operationalise, but they are essential if PPS is to avoid reproducing the epistemic failures that motivate the programme.

9.5. Privacy and Boundary-Layer Robustness

What information leaves the device, and under what conditions? Can disclosure be inspected after the fact? Are differential-privacy budgets tracked and respected? Additionally, the boundary layer should be subject to adversarial evaluation: prompt-injection red-teaming, tool-misuse simulation, and "excessive agency" tests [49,50]. Temporal coherence should also be evaluated: do decay rates and invalidation conditions function correctly? Does the system re-verify stale edges before relying on them for high-stakes decisions?

We note that long-term memory for LLM systems remains empirically difficult. Packer et al. [65] proposed virtual context management (MemGPT) to extend LLM memory, and Park et al. [66] demonstrated generative agents with structured memory—but both highlight that memory is a system-level engineering challenge, not merely a longer prompt. This reinforces our claim that PPS should be a *governed substrate*, not an ever-growing context window.

10. Related Work

Personal intelligence draws on and departs from several research traditions.

User modelling and preference learning. Traditional user modelling—from stereotype-based systems [67] to Bayesian user models [68]—assumes a centralised modeller. PPS differs in locating the model under user governance with portability as a first-class goal.

Knowledge-graph-based recommendation. Guo et al. [19] survey KG-based recommender systems; Wang et al. [20] and He et al. [33] demonstrate graph-based architectures that improve recommendation quality. PPS adopts the graph substrate but adds user governance, provenance, temporal decay, and epistemic separation.

LLM memory and personalisation. MemGPT [65], Generative Agents [66], and A-MEM [69] explore structured memory for LLMs. These systems treat memory as an engineering optimisation for the model; PPS treats it as a *user asset* that outlives any single model.

Local-first systems and portable identity. Kleppmann et al. [51] articulate local-first principles; CRDTs [52] enable decentralised synchronisation. The W3C Verifiable Credentials Data Model [39] and SD-JWT [38] provide standards-track approaches to selective disclosure. PPS applies these concepts to the specific problem of preference portability.

Privacy-preserving machine learning. Federated learning [53], secure aggregation [54], and differential privacy [55] provide the privacy toolkit. PPS's contribution is the *demand side*: what should the privacy-preserving layer protect (a user-governed preference substrate), and why (portability and autonomy, not just regulatory compliance).

Agent safety and adversarial resilience. The OWASP LLM Top 10 [49], Greshake et al. [48], and Datta et al. [50] map the threat landscape for tool-integrated AI. PPS's boundary layer operationalises these concerns as enforceable constraints rather than advisory guidelines.

Data portability and platform regulation. GDPR Article 20 [62], the Data Transfer Project [63], and the Digital Markets Act [64] establish legal frameworks. PPS operationalises portability at the semantic level—not just "export your data" but "carry your preferences in a form that another system can act on."

Concurrent industry work on world-knowledge validation. Quoc [24] proposes "Axiomatic Intelligence" for commerce, emphasising adversarial verification of product claims, signal-gated recomputation, and structural separation of ranking from monetisation. That work focuses on the *world-knowledge* layer (what is true about products); PPS focuses on the *personal* layer (what the user wants). The two are complementary: a complete personal intelligence system requires both a trustworthy world model and a user-governed preference substrate, but neither subsumes the other.

 doi:10.20944/preprints202603.1627.v1

PI's epistemic separation axiom (Section 3.2) and ingress validation mechanism (Section 5.4) define the interface between these layers without committing PPS to any specific world-knowledge architecture.

11. Research Agenda

We identify the following open problems, each with an indication of what success would look like and what makes it hard:

Formal semantics for preference graphs. What graph schema best balances expressiveness, learnability, and user comprehensibility? How should provenance, uncertainty, temporal decay, and mutation triggers be encoded for both machine learning and human inspection? *Hard because:* the schema should serve inference, explanation, and portability simultaneously.

Micro-query policies under attention constraints. How should micro-query policies trade off information gain against user annoyance ("question fatigue")? How should systems handle contradictory evidence without destabilising the user's trust? *Hard because:* the optimisation surface couples information theory with human attention economics.

On-device inference. Current quantised models [9,10] enable local inference, but taste-graph reasoning may require specialised architectures. What is the minimum capability for on-device PPS inference? *Hard because:* graph reasoning and LLM inference have different compute profiles.

Portable identity without universal tracking. Portability risks linkability. How can PPS be carried across services without becoming a universal tracking identifier? Selective-disclosure cryptography and pseudonymous identifiers are candidate approaches. *Hard because:* unlinkability and continuity are in tension.

Cumulative privacy loss accounting. Even with minimal disclosure, repeated interactions leak information. How should cumulative exposure be quantified, tracked, and presented to users in actionable form? *Hard because:* the problem is fundamentally sequential and context-dependent.

Collective learning without centralisation. Federated learning with differential privacy is a credible baseline, but sybil resistance and data-quality verification remain open—especially if incentive mechanisms reward contribution. *Hard because:* adversarial participants and non-IID distributions compound.

Interfaces between PPS and world-knowledge systems. PPS governs what the user wants; but recommendations also depend on what the world offers. How should a PPS-aware system evaluate the trustworthiness of external data sources? When marketing claims conflict with user-reported experience, what resolution strategies best serve preference continuity? *Hard because:* it sits at the intersection of information retrieval, adversarial robustness, and trust modelling.

Incentive-aligned value exchange. If users contribute learning signal, how should value flow back? Token-based, reputation-based, and credit-based systems are candidates, but each introduces gaming risks and regulatory complexity.

Standardisation. Is a "taste graph interchange format" feasible? What minimal schema enables cross-system portability without over-standardising representation? *Hard because:* standardisation requires coordination among competitors.

12. Conclusions

Personal intelligence should be treated as infrastructure: a way for individuals to carry their preferences, constraints, and boundaries through an AI-saturated world without surrendering autonomy to whichever platform happens to control the default interface.

The Personal Preference Substrate and its taste-graph view provide a neutral, domain-agnostic object around which technical work can be organised: representation, learning dynamics, privacy architectures, bounded delegation, portability, and evaluation. The epistemic separation axiom ensures that PPS remains a *preference* substrate rather than collapsing into a general-purpose truth engine—while the boundary layer defines a principled interface to external world-knowledge systems that may take on that complementary role.

 doi:10.20944/preprints202603.1627.v1

In product terms, PPS offers a reversal: instead of people being the behavioural exhaust that trains the platform, the platform becomes an interchangeable tool that serves the person—because the substrate remains under the person's governance. The coming years will determine whether the agentic AI era deepens lock-in or catalyses genuine user governance. This paper has argued that the technical and conceptual building blocks for the latter path exist. What remains is the engineering, the coordination, and the will.

References

1. AWS Machine Learning Blog, "Scaling Rufus, the Amazon generative AI-powered conversational shopping assistant," 2024. URL: https://aws.amazon.com/blogs/machine-learning/scaling-rufus-the-amazon-generative-ai-powered-conversational-shopping-assistant-with-over-80000-aws-inferentia-and-aws-trainium-chips-for-prime-day/
2. Fortune, "Amazon says its AI shopping assistant Rufus is so effective it is on pace to pull in an extra $10 billion in sales," November 2025. URL: https://fortune.com/2025/11/02/amazon-rufus-ai-shopping-assistant-chatbot-10-billion-sales-monetization/
3. Google, "Universal Commerce Protocol," 2026. URL: https://ucp.dev/
4. Google Cloud / PR Newswire, "Google Cloud Brings Shopping and Customer Service Together with Gemini Enterprise for Customer Experience," January 2026. URL: https://www.prnewswire.com/news-releases/google-cloud-brings-shopping-and-customer-service-together-with-gemini-enterprise-for-customer-experience-302657570.html
5. OpenAI, "Buy it in ChatGPT: Instant Checkout and the Agentic Commerce Protocol," September 2025. URL: https://openai.com/blog/buy-it-in-chatgpt/
6. Gartner, "Gartner Predicts By 2028, AI Agents Will Outnumber Sellers by 10x," November 2025.
7. Gartner, "Gartner Unveils Top Predictions for IT Organizations and Users in 2026 and Beyond," October 2025.
8. Anthropic, "Model Context Protocol Specification," 2024-2025. URL: https://modelcontextprotocol.io/
9. J. Lin et al., "AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration," *MLSys*, 2024. Best Paper Award.
10. E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers," *ICLR*, 2023.
11. Y. Zheng et al., "A Review on Edge Large Language Models: Design, Execution, and Applications," *ACM Computing Surveys*, vol. 57, no. 8, 2025.
12. J.-C. Rochet and J. Tirole, "Platform Competition in Two-Sided Markets," *J. European Economic Association*, vol. 1, no. 4, pp. 990-1029, 2003.
13. C. Doctorow, "The 'Enshittification' of TikTok," *Wired*, January 2023.
14. P. Klemperer, "Competition when Consumers have Switching Costs," *Review of Economic Studies*, vol. 62, no. 4, pp. 515-539, 1995.
15. J. Farrell and P. Klemperer, "Coordination and Lock-In: Competition with Switching Costs and Network Effects," *Handbook of Industrial Organization*, vol. 3, ch. 31, Elsevier, 2007.
16. H. Allcott, L. Braghieri, S. Eichmeyer, and M. Gentzkow, "The Welfare Effects of Social Media," *American Economic Review*, vol. 110, no. 3, pp. 629-676, 2020.
17. F. Huszar, S. I. Ktena, C. O'Brien, L. Belli, A. Schlaikjer, and M. Hardt, "Algorithmic Amplification of Politics on Twitter," *PNAS*, vol. 119, no. 1, 2022.
18. S. Zuboff, *The Age of Surveillance Capitalism*, PublicAffairs, 2019.
19. Q. Guo et al., "A Survey on Knowledge Graph-Based Recommender Systems," *IEEE TKDE*, vol. 34, no. 8, pp. 3549-3568, 2022.
20. X. Wang, X. He, Y. Cao, M. Liu, and T.-S. Chua, "KGAT: Knowledge Graph Attention Network for Recommendation," *KDD*, pp. 950-958, 2019.
21. S. S. Iyengar and M. R. Lepper, "When Choice Is Demotivating," *JPSP*, vol. 79, no. 6, pp. 995-1006, 2000.
22. F. Germano, V. Gomez, and F. Sobbrio, "Ranking for Engagement: How Social Media Algorithms Fuel Misinformation and Polarization," CESifo Working Paper No. 9978, 2022.
23. S. Gonzalez-Bailon et al., "Asymmetric Ideological Segregation in Exposure to Political News on Facebook," *Science*, vol. 381, pp. 392-398, 2023.

 doi:10.20944/preprints202603.1627.v1

24. M. Quoc, "Axiomatic Intelligence: A Post-Probabilistic Architecture for the Age of Noise," Product AI Research, January 2026.
25. K. Friston, "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, pp. 127-138, 2010.
26. K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, and G. Pezzulo, "Active Inference: A Process Theory," *Neural Computation*, vol. 29, no. 1, pp. 1-49, 2017.
27. S. Lichtenstein and P. Slovic, *The Construction of Preference*, Cambridge University Press, 2006.
28. T. Parr, G. Pezzulo, and K. J. Friston, *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*, MIT Press, 2022.
29. L. Da Costa, T. Parr, N. Sajid, and K. Friston, "Active Inference as a Model of Agency," arXiv:2401.12917, 2024.
30. A. Mathur et al., "Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites," *Proc. ACM HCI*, vol. 3, CSCW, 2019.
31. C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, and A. L. Toombs, "The Dark (Patterns) Side of UX Design," *CHI '18*, ACM, 2018.
32. A. Hogan et al., "Knowledge Graphs," *ACM Computing Surveys*, vol. 54, no. 4, article 71, 2021.
33. X. He et al., "LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation," *SIGIR*, pp. 639-648, 2020.
34. S. Wu, F. Sun, W. Zhang, X. Xie, and B. Cui, "Graph Neural Networks in Recommender Systems: A Survey," *ACM Computing Surveys*, vol. 55, no. 5, 2022.
35. C. Gao et al., "A Survey of Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions," *ACM Trans. Recommender Systems*, vol. 1, no. 1, 2023.
36. OpenID Foundation, "OpenID Connect Core 1.0," Section 8 (Pairwise Subject Identifiers). URL: https://openid.net/specs/openid-connect-core-1_0.html
37. IETF, "Selective Disclosure for JWTs (SD-JWT)," draft-ietf-oauth-selective-disclosure-jwt-15, 2025. URL: https://datatracker.ietf.org/doc/html/draft-ietf-oauth-selective-disclosure-jwt-15
38. W3C, "Verifiable Credentials Data Model v2.0," 2024. URL: https://www.w3.org/TR/vc-data-model-2.0/
39. W3C, "Verifiable Credentials Data Model v2.0—Privacy Considerations," 2024.
40. T. Joachims, A. Swaminathan, and T. Schnabel, "Unbiased Learning-to-Rank with Biased Feedback," *WSDM*, pp. 781-789, 2017.
41. L. Li, W. Chu, J. Langford, and X. Wang, "Unbiased Offline Evaluation of Contextual-bandit-based News Article Recommendation Algorithms," *WSDM*, pp. 297-306, 2011.
42. M. Elahi, F. Ricci, and N. Rubens, "A Survey of Active Learning in Collaborative Filtering Recommender Systems," *Computer Science Review*, vol. 20, pp. 29-50, 2016.
43. J. R. Carbonell, "AI in CAI: An Artificial-Intelligence Approach to Computer-Assisted Instruction," *IEEE Trans. Man-Machine Systems*, vol. 11, no. 4, pp. 190-202, 1970.
44. A. C. Graesser et al., "AutoTutor: A Tutor with Dialogue in Natural Language," *Behavior Research Methods*, vol. 36, no. 2, pp. 180-193, 2004.
45. B. D. Nye, A. C. Graesser, and X. Hu, "AutoTutor and Family: A Review of 17 Years of Natural Language Tutoring," *Int. J. AIED*, vol. 24, pp. 427-469, 2014.
46. N. A. Kumar and A. Lan, "Improving Socratic Question Generation using Data Augmentation and Preference Optimization," *BEA Workshop*, ACL, 2024.
47. J. Liu et al., "SocraticLM: Exploring Socratic Personalized Teaching with Large Language Models," *NeurIPS*, 2024.
48. K. Greshake et al., "Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection," *AISec '23*, ACM CCS, 2023.
49. OWASP, "Top 10 for Large Language Model Applications (v1.1)," 2025. URL: https://owasp.org/www-project-top-10-for-large-language-model-applications/
50. S. Datta, S. K. Nahin, A. Chhabra, and P. Mohapatra, "Agentic AI Security: Threats, Defenses, Evaluation, and Open Challenges," arXiv:2510.23883, 2025.
51. M. Kleppmann, A. Wiggins, P. van Hardenberg, and M. McGranaghan, "Local-First Software: You Own Your Data, in spite of the Cloud," *Onward! '19*, SPLASH, pp. 154-178, 2019.
52. M. Shapiro, N. Preguica, C. Baquero, and M. Zawirski, "Conflict-Free Replicated Data Types," *SSS 2011*, LNCS 6976, Springer, pp. 386-400, 2011.

 doi:10.20944/preprints202603.1627.v1

53. H. B. McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data," *AISTATS*, PMLR 54, pp. 1273-1282, 2017.
54. K. Bonawitz et al., "Practical Secure Aggregation for Privacy-Preserving Machine Learning," *CCS '17*, ACM, pp. 1175-1191, 2017.
55. C. Dwork and A. Roth, *The Algorithmic Foundations of Differential Privacy*, Now Publishers, 2014.
56. Apple Differential Privacy Team, "Learning with Privacy at Scale," Apple Machine Learning Research, 2017. URL: https://machinelearning.apple.com/research/learning-with-privacy-at-scale
57. P. Kairouz, Z. Liu, and T. Steinke, "The Distributed Discrete Gaussian Mechanism for Federated Learning with Secure Aggregation," *ICML*, PMLR 139, pp. 5201-5212, 2021.
58. J. Groth, "On the Size of Pairing-Based Non-interactive Arguments," *EUROCRYPT 2016*, LNCS 9666, pp. 305-326, 2016.
59. A. Gabizon, Z. J. Williamson, and O. Ciobotaru, "PLONK: Permutations over Lagrange-bases for Oecumenical Noninteractive arguments of Knowledge," ePrint 2019/953, 2019.
60. Ingonyama, "IMP1: Bringing Zero-Knowledge Proofs to Mobile," 2025. URL: https://github.com/ingonyama-zk/imp1
61. N. Luo, G. Wu, Y. Zhu, B. C. Ooi, and N. Xiao, "Secure and Verifiable Data Collaboration with Low-Cost Zero-Knowledge Proofs (RiseFL)," *PVLDB*, vol. 17, no. 9, pp. 2321-2334, 2024.
62. Regulation (EU) 2016/679, Article 20 – Right to data portability, *GDPR*, 2016.
63. Data Transfer Project, "Overview and Fundamentals," 2018. URL: https://dtinit.org/
64. A. Fletcher et al., "The Effective Use of Economics in the EU Digital Markets Act," *J. Competition Law and Economics*, vol. 20, no. 1-2, pp. 1-19, 2024.
65. C. Packer et al., "MemGPT: Towards LLMs as Operating Systems," arXiv:2310.08560, 2023.
66. J. S. Park et al., "Generative Agents: Interactive Simulacra of Human Behavior," *UIST '23*, ACM, 2023. Best Paper Award.
67. A. Kobsa, "User Modeling: Recent Work, Prospects and Hazards," in *Adaptive User Interfaces*, Elsevier, pp. 111-128, 1993.
68. D. N. Chin, "Empirical Evaluation of User Models and User-Adapted Systems," *User Modeling and User-Adapted Interaction*, vol. 11, no. 1-2, pp. 181-194, 2001.
69. S. Wang et al., "A-MEM: Agentic Memory for LLM Agents," arXiv:2502.12110, 2025.
70. G. A. Montes, "The first infrastructure: Cognitive integrity & security," *Gabriel Axel (Substack)*, March 2026. URL: https://gabrielaxel.substack.com/p/the-first-infrastructure-cognitive