Personal World Models: Toward Human-Centric Personal Intelligence Systems

Justin Holbrook
Principal Researcher
anywhereintelligence.ai

Swaroop Kallakuri
Principal Researcher
anywhereintelligence.ai

Joshua Holbrook
Principal Researcher
anywhereintelligence.ai

Lucas Burgess
Independent Researcher

Abstract

The dominant paradigm in AI research optimizes for machine-side intelligence: model scale, benchmark performance, reasoning depth, and agentic capability. This framing is correct for many objectives and incorrect for one: systems designed to live with a specific human over years, accumulating a structure of meaning that strengthens the coherence, agency, and adaptive capacity of that person over time. For such systems, the central missing architecture is not more intelligence but a different kind of structure. Specifically, one that encodes not how the world generally behaves, but what the world specifically means to this user, in their specific history, with their specific relationships.

Prior work proposed that achieving Base Level Intelligence (BLI) — goal specification, state measurement, and adaptive difference reduction — requires heterogeneous cognitive architectures rather than scaled language models. This paper extends that argument along a second dimension: even a system exhibiting BLI that operates on a general world model cannot serve a specific human's personal intelligence over time without a categorically distinct structure.

We adopt the term *personal world model* as the locally embodied structure of referents, relationships, and historical associations through which a specific user experiences and interprets the world. We introduce and formalize naive architectural requirements for building a personal world model: permissioned meaning commitment, three-layer knowledge separation (ontology primitives / common-sense priors / meaning commitments), distinction-preserving coherence management (indexicality and injectivity), constitutional continuity as a user-owned alignment process, continuous engagement without session boundaries, and meaning-constrained execution. We present AIOS (Anywhere Intelligence Operating System) as an initial implementation of these principles and contrast the personal world model with existing approaches, such as user profiling, graph RAG, personalized memory, and LLM knowledge bases. We conclude by proposing an evaluation agenda for personal world models and situate our contribution within the human-computer interaction, knowledge representation, and AI alignment literature.

1.1 The Intelligence Maximization Paradigm and Its Limits

Since the earliest neural networks of the 1950s, the dominant framing of AI progress has been machine-centric: how capable is the system, how broadly can it generalize, and how high does it score on benchmarks that abstract over specific domains? This framing has produced extraordinary results within its own terms. Large language models trained at scale demonstrate fluency across domains that would have been implausible ten years ago. Agentic systems complete multi-step

tasks with increasing reliability. Vision models match and exceed human performance on perceptual benchmarks. These are genuine achievements.

Yet the intelligence maximization paradigm has a structural blind spot. It is optimized for performance averaged across users — for general rather than personal capability. A system with state-of-the-art benchmark performance and no record of what a specific user finds meaningful will, on the day of its first use, produce outputs calibrated to statistical expectation instead of individuated context. It cannot know "the project" refers to a specific long-running concern, that a certain name refers to a deceased family member, or that a particular topic is deliberately avoided. After resetting at the end of every conversation, it will produce the same generalized behavior again.

This is not a failure of intelligence. It is a failure of architecture. The system lacks the structure required for a different task than the one for which it was designed.

1.2 The Personal Intelligence Gap

In April 2026, Andrej Karpathy described an influential personal knowledge management workflow [34]: depositing raw source material into a directory where a large language model compiles and maintains an interlinked wiki. He explicitly invoked Vannevar Bush's 1945 Memex [13] as the intellectual predecessor. The response was significant. Practitioners recognized the advance (an always-current, searchable, synthesizing external memory) and critics identified the gap (a compiled wiki that stores what the user has read but not what it personally means).

Bush's original Memex innovation was not storage. It was the *associative trail* — a record of which connections mattered to this person, at which moment, and with what weight [13]. Current compiled-wiki systems inherit the storage without the trail. They optimize documents as helpful referents, but lack the mechanism to record their evolving meaning.

The gap between what a compiled knowledge base produces (addressable, well-structured information) and what a personal intelligence system requires (a structure of committed personal meaning) is the central problem this paper addresses. We term this missing structure a *personal world model*, arguing it is not a more detailed instance of a general world model but a categorically different structure with alternative primitives, temporal dynamics, authority relationships, and failure modes.

1.3 Contributions

This paper makes five distinct contributions:

1. **The personal world model as a formal construct.** We formally distinguish the personal world model from the general world model [28,37], from user profiling [21], from persistent LLM memory [41,51,73], and from graph RAG [52]. The distinctions are structural, not cosmetic: different units of storage, different authority relationships, different handling of silence and contradiction, different trust properties.
2. **The three-layer knowledge separation.** We introduce and argue for a principled architectural separation of ontology primitives (the grammar of meaning), common-sense priors (defeasible probabilistic defaults), and personal meaning commitments (user-authorized, provenance-tagged, reversible referents). We show that conflating any two of these layers is the principal architectural error in existing approaches.
3. **The physics of coherence: indexicality and injectivity.** We formalize two governing constraints for long-lived personal intelligence systems: *indexicality* (every system state must have a traceable causal history) and *injectivity* (distinct meanings must not collapse to the same representation). We derive continuous engagement — the abolition of session boundaries — as a necessary architectural response to these constraints.
4. **Constitutional continuity as a design primitive for personal alignment.** We introduce *constitutional continuity*: an alignment approach that preserves the structure through which a user continues to revise the system's normative commitments, rather than targeting a fixed alignment objective. We argue this is the correct framing for systems designed to accompany a human life across years, during which values and priorities genuinely evolve.
5. **A concrete evaluation agenda.** We propose several experiments designed to measure properties that distinguish a personal world model from its nearest competitors, including the first prediction we are aware of that specifically tests the bootstrapping tradeoff: conservative

early commitment outperforms aggressive early inference at T+90 despite underperforming at T+1.

1.4 Relationship to Prior Work

This paper is part of a research program developing the theoretical and architectural foundations for personal intelligence systems. A prior paper in this program [30] proposed that Pinker's three subroutines for goal attainment require heterogeneous cognitive architectures rather than scaled language models [42], that the "barrier of meaning" [60,49] is the central unsolved problem in AI, and that achieving understanding-as-capability requires structurally different mechanisms for different cognitive tasks. This paper assumes that framework as background and extends it: even a system that fully achieves goal attainment for an aligned capability and responsibility set fails at personal intelligence if it has no architecture for accumulating and actioning what the directing user has declared meaningful.

A companion paper [31] to this one establishes the IPU (Intelligence Processing Unit) as an architectural primitive for accountability and trust: every interaction produces an immutable, inspectable execution trace that gates all durable system effects. This paper grounds that architecture in the personal world model: the IPU trace must carry not only routing and retrieval provenance but the record of which meaning commitments governed execution, which scope boundaries were enforced, and which operations on the personal world model occurred during the interaction.

1.5 Paper Organization

The paper proceeds as follows. Section 2 establishes the personal intelligence framing and its theoretical basis in extended mind theory and the BLI framework. Section 3 formally distinguishes general world models from personal world models and introduces the three-layer knowledge architecture for AIOS. Section 4 examines why LLM knowledge bases are structurally insufficient. Section 5 addresses the cold-start bootstrapping problem. Section 6 develops the physics of coherence: indexicality, injectivity, and continuous engagement. Section 7 addresses the normative dimension: plural human realities, constitutional continuity, and the architectural account of safety. Section 8 describes meaning-constrained execution and the IPU-grounded trust architecture. Section 9 addresses objections and proposes an evaluation agenda. Section 10 concludes.

2. Personal Intelligence vs. Intelligent Systems

2.1 Intelligence as a Functional Property

Any claim about "personal intelligence" must begin with a workable definition of intelligence that does more than restate anthropomorphic intuitions. Pinker (1997) [55] provides the required foundation: intelligence is the capacity to attain goals in the face of obstacles, decomposable into goal specification, state measurement, and adaptive difference reduction. This three-part structure is what we formalize as Base Level Intelligence (BLI) in our prior work [30] and argue is necessary for establishing an aligned, non-anthropomorphic, machine-side benchmark.

BLI does not adjudicate what intelligence is. As Minsky observed [48], intelligence is a "suitcase word" definitionally contested across multiple disciplines. Instead, BLI provides an engineering bypass whose three subroutines can be specified, built, and evaluated in the lab. A chess engine achieves high BLI within a precisely specified objective space. A large language model achieves partial BLI within a diffuse, poorly bounded objective space. The question of whether any system is intelligent thus involves three constituents: intelligent with respect to what goal; measured by what state; and reduced by what mechanism?

For systems designed to live with a single human over years, this question has a determinate answer: the goal is the user's own goals; the state that matters is the user's epistemic and practical situation; the difference to be reduced is the gap between the user's experienced and desired aims. [46,56]. A system operating on the wrong objective (machine task completion, token throughput, benchmark performance, agentic accomplishment) is not capably functioning as a personal intelligence system in at least two ways. First, it ignores the user's underlying epistemic objectives [64], targeting instead generalizable outcomes. Second, it omits the distinction between internal (comprehension, decision-making, available knowledge) and external (third-party resistance, institutional barriers, resource constraints, systemic exclusion) obstacles to goal attainment [56,64].

This reframing is precise: the operational question for a personal intelligence system is not "how intelligent is the AI?" but "how can a specific human with a specific goal exhibit greater intelligence over time?" These are different questions with different architectures, different failure modes, and different measures of success.

2.2 The Extended Mind: From Tool Use to Cognitive Constitution

The reframing above might seem to permit a conservative interpretation: AI is a tool and better tools make humans more capable. This interpretation is insufficient.

The parity principle applies. If a process in the world performs the function that an internal cognitive process would otherwise perform, and the agent relies on it in the way they would rely on an internal process, then the external process is part of the agent's cognitive system [15]. The cognitive boundary does not coincide with the internal. External artifacts do not merely assist pre-existing cognitive processes but partially constitute them [16]. The extended cognitive system is the coupled human-artifact system.

A complementary account at the organizational level follows from distributed cognition research [32]. Cognitive processes are distributed across individuals, artifacts, and environments. The unit of analysis for cognition is the system, not the individual. Personal intelligence systems operate at the same level; the relevant cognitive system is the user-plus-system coupling.

These commitments raise the design stakes considerably. If an AI system functions as a genuine constituent of the user's extended cognitive system, the foundational design concern shifts from one centered on system capabilities to one centered on human capabilities.

This shift directly implicates capability loss concerns. Research on cognitive offloading [59] distinguishes complementary offloading (external structures free internal resources for higher-order reasoning) from substitutive offloading (external structures replace rather than support internal processes, potentially degrading the capabilities they ostensibly serve). A well-designed personal intelligence system produces complementary offloading. A poorly designed system (one that provides answers without building understanding, or resolves ambiguities without user authorization) produces substitutive offloading.

2.3 BLI Is Necessary but Insufficient

A system designed toward BLI pairs capability and responsibility sets by specifying goals, measuring states, and adaptively reducing differences with high fidelity for a particular purpose. It nevertheless fails as a personal intelligence system if it has no record of what its specific user has declared as meaningful in relation to the specified goal. Research elsewhere identifies the barrier of meaning as the central unsolved problem in AI [7,49,60]. Our prior work [30] posits that crossing this barrier requires understanding-as-capability grounded in heterogeneous cognitive architectures. The present paper's contribution is a distinct claim: even after crossing the barrier of meaning in a generalized sense (what does this input mean), a system without a personal world model has not crossed the *personal* barrier of meaning (what does this input mean to this particular user) if it cannot understand individuated context over time.

An example: a system with satisfactory BLI operating on a general world model knows many things about the world. It does not know that "grandma" refers to a specific person in the user's life whose apple cookies carry childhood significance, that certain topics like politics at family holidays are intentionally avoided, or that relationships in the user's professional life have specific histories affecting email tone and transparency. These absences are not calibration failures of the general model. They are structural absences in its architecture.

2.4 The Session Model as Diagnostic Failure

The session model in which context is discarded at the end of each bounded interaction is not a neutral design choice. It is a structural incompatibility with personal intelligence.

The session model's historical origin is a performance constraint: limited context windows required bounded interactions. These constraints have relaxed, but the session model persists through inertia in an architecture optimized for machine convenience. When a session ends, all accumulated referential structure is discarded. Absent augmentation, the user must re-establish context at every

conversation. The costs are borne entirely by the user, and historically have been attributed to user behavior (the user "forgot" to provide context) rather than architectural failure. By contrast, human autobiographical memory organizes around the ongoing self — goals, narrative, identity — as a continuous structure that spans years [17] rather than minutes. The session model imposes an interruption that is architecturally foreign to the memory structure it augments.

A personal intelligence system must be designed differently. Goals, self-identity, and life narrative shape which memories are retained, how they are structured, and how they are retrieved. Rather than the session model, a personal intelligence system must enable what we term *continuous engagement*: a single, unbroken timeline of interaction per user. This requirement is costly. It demands cross-interaction history retrieval, referent resolution across arbitrary time spans, and a persistent identity model scoped to the user. These are not incidental overheads. In our view, however, they are the structural price of a system able to distinguish what a particular user means from what most users statistically tend to mean.

2.5 Non-Interchangeability as Thesis

The arguments above converge on a simple diagnostic for personal intelligence systems: after first use, two instances running identical software should no longer be meaningfully interchangeable. This non-interchangeability is not a limitation to be corrected. It is evidence that personal intelligence is accumulating.

If a system is designed to accumulate instance-specific referential structure (meaning links grounded in the user's history, stable referents reflecting the user's relationships, contextual bindings established through the user's explicit grants of authority) then two instances initialized from the same software base will diverge immediately upon first use. Every interaction that establishes a new referent is a divergence event. After months of daily use, two instances will be no more interchangeable than two human memories.

This inverts the usual engineering intuition. In general-purpose software, instances are expected to be interchangeable; divergence is a failure of consistency. In personal intelligence systems, continued interchangeability is the failure. A system that remains interchangeable after months of use has not served its user's personal intelligence. It has served itself.

The argument for human-centered design frameworks has a long lineage [39,20,66]. What is new in our proposal is the formal machinery required to achieve it: a permissioned, temporal, user-owned meaning substrate that accumulates rather than resets, diverges rather than converges, and treats meaning formation as a prerequisite for action rather than its consequence. In other words, a personal world model.

3. From World Models to Personal World Models

3.1 World Models in AI

The phrase "world model" describes a compact internal representation allowing an agent to simulate future states of its environment [28]. The model learns physical dynamics, planning acts against simulated rollouts. Crucially, such a model can be trained on one environment and transferred to another. It is designed to be maximally general.

LeCun (2022) [37] extends this framing significantly in his proposal for autonomous machine intelligence. The Joint Embedding Predictive Architecture (JEPA) posits world models as an essential structural component: an agent must construct and consult an internal model of the world's causal structure to plan, reason about counterfactuals, and act with appropriate uncertainty. The broad landscape of world models in deep learning and reinforcement learning documents progress in modeling physical, causal, and semantic regularities designed to hold across contexts, environments, and agents [43].

This universality is a strength in every context except one: systems designed to live with a specific human over years that must accumulate meaning that is irreducible to environment-level regularity.

3.2 AIOS as an Architectural Framework

The requirements established above (BLI oriented toward person-level objectives, continuous engagement without session boundaries, a permissioned user-owned meaning substrate, and the non-interchangeability thesis) are jointly sufficient to specify the shape of a personal intelligence system. This paper presents AIOS (Anywhere Intelligence Operating System) as a naive framework for implementing these requirements.

AIOS is not designed to be the most capable general-purpose AI system. It is designed to be a personally significant system for a specific user over time. The two design objectives are structurally distinct: the former optimizes machine-side intelligence averaged across users; the latter optimizes person-side coherence for a single user. AIOS formalizes this distinction as its central design principle. Moments after initialization, two AIOS instances running identical software begin to diverge.

Three design properties of AIOS distinguish its architecture:

User-owned meaning architecture. AIOS separates ontology primitives, common-sense priors, and personal meaning commitments into three structurally distinct, non-contaminating layers. Meaning entries are user-authorized: proposed by the system, committed only with user assent, revisable only by user action. The system is a scribe, not an interpreter.

Continuous engagement without session boundaries. AIOS maintains a single, unbroken interaction timeline per user. There is no session boundary at which referential context is discarded. The personal world model accumulates across all interactions, with cross-interaction history retrieval and referent resolution scoped to the user's persistent identity.

Meaning-constrained execution. AIOS treats meaning formation as a prerequisite for action rather than its consequence. A system that acts before meaning establishment may execute competently on a general-world interpretation of the user's request while missing what the request actually means to the user. Meaning-constrained execution ensures the system's operative representation of user intent is authorized before consequential actions are taken.

3.3 The Personal World Model Defined

A personal world model, as defined in the AIOS architecture, is the locally embodied structure of referents, relationships, and historical associations through which a specific user experiences and interprets the world. This definition identifies a structure that is not a better or more detailed general world model. It is a different kind of structure — one that encodes not how environments typically behave, but what things mean to a specific person, in their specific history, with their specific relationships.

The philosophical literature establishes why this distinction is irreducible. Frege's (1892) [25] distinction between *Sinn* (sense) and *Bedeutung* (reference) establishes that two expressions can refer to the same object while carrying different senses and modes of presentation. A personal world model is, in Fregean terms, a system of senses rather than references. The word "grandma" may refer to the same individual as "my mother's mother," but the sense it carries — the mode of presentation, the emotional weight, the contextual associations — is not reducible to that reference. Collapsing the sense into the reference discards precisely what makes it meaningful.

Other work deepens the argument. Certain phrases ("I," "here," "now," names of familiarity) cannot be eliminated from a representation of meaning without changing that meaning [54]. They are essential indexicals: tied to the speaker, the context, and the history of use. A system that reduces "grandma's apple cookies" to a generic recipe entity, for example, has not preserved meaning; it has replaced it. Putnam's (1975) [57] Twin Earth argument makes the point from a different angle: meaning is not only in the head. The meaning of "water" on twin Earths turns on external factors, including its history of use and the chemical composition of the liquid it names.

A personal world model is therefore causal-historical: it records not just what a term is associated with now, but through what sequence of interactions that association was established.

3.4 Comparative Structural Analysis

Table 1: Comparative structural differences between general and personal world models.

Dimension	General World Model	Personal World Model
Subject of representation	Physical regularities; causal dynamics across environments	Lived significance structure of a specific user: referents, associations, evolving meanings
Transferability	Designed to be transferable across agents and instances	Instance-local; cannot be transferred without replacing the person whose meanings it encodes
Ontological source	Trained on observations of environmental states	Built through user-authorized meaning commitments
Silence and unresolved states	Treated as missing data or prediction uncertainty	First-class stable states; silence is an instructional boundary
Primary linguistic mode	Third-person factual representation	Indexical expressions that cannot be eliminated
Temporal evolution	Updated by new environmental observations	Updated through permissioned interaction; prior associations require explicit revision

These are not points on a single continuum. They differ in what they represent, what authorizes their contents, how they handle non-resolution, and what their relationship to language is.

3.5 Why Existing Approaches Do Not Fill the Gap

User modeling systems represent users as profiles: preference vectors, behavioral patterns, demographic clusters [21]. Structurally, these are statistical summaries of behavioral signals, lacking entry for individuated referents such as provenance, indexicality, and historical creation within preference vectors.

Knowledge graphs encode objective, intersubjective knowledge [29]. Other work has surveyed LLM-KG integration and demonstrated gains in factual grounding, but the knowledge in those graphs is public, general, and authority-free [52]. The personal world model's commitment structure — user-authorized, provenance-tagged, reversible, silence-preserving — is not present.

Persistent LLM memory addresses context window limits through hierarchical memory tiers and temporal decay [51,73,72]. While beneficial, these approaches do not distinguish between a fact passively ingested from a document, a preference inferred from behavioral signals, or a commitment explicitly authorized by the user. Categorical authority — who authorized this entry, with what status, and under what revision rules — is omitted.

Personal knowledge graphs apply graph structures to personal information [6]. These address structural gaps but lack the three-layer separation between ontology primitives, common-sense priors, and personal meaning commitments that we argue a personal intelligence system requires.

3.6 The Three-Layer Architecture

As a system designed to accommodate personal world models, AIOS separates three structurally distinct layers:

Ontology primitives are the grammar of meaning: the irreducible structural categories through which meaning can be represented. These include identity primitives (`Entity`, `IdentityRelation`, `LifecycleState`), temporal primitives (`TimePoint`, `TimeRange`, `TemporalRelation`), epistemic status primitives (`AssertionStatus`, `Confidence`), and contradiction-preserving primitives (`TensionMarker`). The primitive set is culture-independent and user-independent by design. It defines the structural vocabulary in which a personal world model can be expressed.

Common-sense priors are defeasible defaults: probabilistic expectations that apply in the absence of specific knowledge. Social-relational priors include defaults like "family members often know each other." These serve three legitimate functions: interpretive support, disambiguation guidance, and extraction heuristics. They must never create entities, create meaning links, assert personal facts, or override explicit user statements. A prior about what "grandmother" statistically tends to refer to must not overwrite the specific referent that this user established through a particular interaction.

Knowledge units are committed, user-authorized meaning entries: provenance-tagged, temporally anchored, reversible. A knowledge unit for a named person in a user's personal world model carries the entity identifier, the lifecycle state, the temporal scope of the referent, the IPU trace under which the meaning commitment was established, and the assertion status. The separation prevents contamination in both directions: a common-sense prior cannot silently promote a candidate referent to a knowledge unit, and a knowledge unit cannot be silently overwritten by a statistically more probable alternative.

Table 2: The Meaning Formation Pipeline.

Stage	Operation	Epistemic Status	Example
Capture	Ingest, index, retrieve	No meaning committed; content is findable, not interpreted	Email threads indexed by sender, date, thread; documents stored verbatim
Claim Extraction	Structured proposals extracted from content	Candidate commitments; carry confidence and evidence spans; not authoritative	"User appears to have a sibling named Sarah" — claim with evidence, confidence 0.82, marked inferred
Meaning Commitment	User-authorized entries written to personal knowledge graph	Durable meaning; attributed to user; revisable only by user	User confirms: "Sarah is my sister" $\rightarrow$ `RelationKU` created, provenance traced to IPU

At each stage, the system preserves content without escalating its epistemic status until explicitly authorized. The extraction stage produces proposals, not facts. Claims without evidence spans are rejected. The commit engine — the only component authorized to create durable knowledge objects — applies deterministic validation logic before any graph writes occur.

This architecture corresponds to postulates for belief revision established elsewhere: rational belief revision requires minimal change, consistency preservation, and conservation [2,27,71]. Each commitment is a deliberate addition, each revision requires explicit user action, and the history of commitments is immutably preserved with provenance.

3.7 Silence as a First-Class Architectural Requirement

A persistent failure mode in AI knowledge systems is premature resolution: the system detects an ambiguity and resolves it, choosing the statistically most probable interpretation and committing to it. This is instrumentally efficient but semantically incorrect in the personal domain.

Personal world models include deliberate silences: topics a user avoids naming, relationships referred to only indirectly, events acknowledged without elaboration, or associations left intentionally unlinked. These are not missing data. They are instructional boundaries treated as states in which the user has, implicitly or explicitly, indicated resolution is not authorized.

Two ontology primitives encode this absence: `AssertionStatus`, which admits `hypothetical` and null as stable states rather than forcing a binary commitment, and `TensionMarker`, which preserves contradictions rather than resolving them. Meaning links may only be established when required for a user-invoked action and when the user authorizes clarification. Background inference is insufficient. A user's avoidance of a particular topic carries information.

Premature resolution is not merely a timing error; it is a category error. When a system collapses an ambiguous reference, it makes a meaning claim the user has not authorized. The AIOS architecture

represents this constraint structurally: latent referents are treated as valid states instead of failure states. Multiple candidates for the same name, partially overlapping mentions, and unresolved relationships persist as distinct possibilities until the user acts to resolve them. Co-presence does not establish durable semantic links. Silence is simply a record that the question has been encountered and not yet answered.

4. Why Knowledge Bases Are Not Enough

4.1 The LLM Knowledge Base: A Genuine Advance

The compiled LLM wiki proposed by Karpathy [34] represents a step toward the vision that Bush [13] articulated eighty years ago: an external memory system that extends human cognition. A user deposits raw source material and a language model compiles and maintains an interlinked wiki from that material, synthesizing cross-references, resolving redundancies, and producing structured entries. For knowledge work at personal scale, the compiled document collection fits within extended context windows without retrieval-augmented generation.

This is a legitimate technical advance. Human-maintained personal wikis degrade over time: maintenance is cognitively expensive, cross-references fall out of date, indexing discipline erodes. A large language model sidesteps these frictions. For knowledge work, the compiled LLM wiki represents real progress on a real problem.

This section accepts that contribution. What follows is a precise characterization of what that knowledge base does and does not produce. The distinction does not imply deficiency in engineering. Rather, it suggests architectural objectives distinct from those of personal intelligence systems.

Table 3: LLM knowledge bases vs. personal world models.

Dimension	LLM Knowledge Base	Personal World Model
Unit of storage	Documents and extracted claims	Meaning commitments with provenance
User authorization	Material is deposited; structure is inferred	Meaning is granted, not inferred
Temporal lifecycle	All entries persist equally	Significance evolves, attenuates, or is revoked
Contradiction handling	Tensions surfaced but resolved toward coherence	Contradiction is a first-class state; collapse is not forced
Relational modality	Named entities as document tags	Referents with emotional, historical, and relational properties
Silence and omission	Gaps are filled through inference	Silence is an instructional boundary, not missing data
Authority	System decides structure; user deposits	User authorizes what is committed; system proposes, does not assert

Permissioned meaning commitments. In a compiled wiki, every entry has the same epistemic status: an extracted claim from a deposited source. No entry is user-authorized as personally significant. Floridi's theory of semantic information [22,23] is relevant here: information is well-formed and contributive, better than simple storage but primarily operative as propositional content rather than personal significance. A compiled wiki reaches Floridi's level of semantic information. It does not reach the level of committed personal meaning with an authorizing subject.

Temporal lifecycle of significance. Topics that occupied central attention three years ago may have become peripheral. Relationships that were defining may have dissolved. A wiki lacks native mechanism for this lifecycle: all entries persist with equal authority regardless of when they were deposited or whether their significance has changed. Earlier efforts have introduced stored artifact

summarization and decay as a means of mimicking temporal evolution [72,73], but these operate at the level of document recency rather than user significance. Personal meaning, which often attaches to the infrequent over the frequent, is not the same as usage frequency.

Contradiction as a first-class state. A compiled wiki's coherence objective works against the representation of genuine epistemic ambivalence. A person can hold two incompatible beliefs simultaneously (a personal relationship, a career direction, a normative claim) without viewing this ambivalence as a deficiency requiring resolution. Nonmonotonic reasoning [10] underscores that epistemic states cannot always be made consistent without losing information.

Relational modality. The referent "my grandmother" is not a document tag. It is a bounded node with specific emotional weight, historical depth, and relational properties that constrain which operations are appropriate. Understanding in human-computer interaction is not built up through passive accumulation of content [69]. Instead, it is revealed at moments of breakdown, when something fails to work as expected and the structure governing use becomes visible. A system that stores only documents lacks relational modality, and therefore the capacity to detect node violation.

Epistemic boundaries. A wiki that aggressively cross-references every deposited document treats silence as missing data to be filled through inference. AIOS treats silence, ambiguity, and non-resolution as stable first-class states. Central to this thesis is the proposition that meaning cannot be learned from form alone [7,11]. A compiled document collection, even if linked [1], lacks the layer in which the user becomes an authorizing agent, risking *structural debt* in the form of unauthorized meaning claims requiring future navigation.

5. Meaning Before Task

5.1 Reframing the Cold-Start Problem

The cold-start problem as conventionally framed is a scarcity problem [62,36,53]: how does a system make useful predictions when it has little data about a user? The traditional goal is to extract signal as rapidly as possible. AIOS inverts this framing. In a personal intelligence system, the cold-start problem is not "what do we already know about this user?" but "how do we avoid making commitments the user has not authorized while still remaining addressable and useful?"

The first task of a personal intelligence system is therefore not task execution but *meaning capacity preservation*: ensuring the system remains capable of knowing the user later, without having crossed epistemic boundaries in the process. A system that maximally infers the user's preferences, relationships, and values in the first week may appear impressive early, but the representations it builds have not been authorized. Every subsequent interaction must negotiate between what the user actually means and what the system has already decided they mean. By aggressively resolving cold-start ambiguity, it has pre-loaded collisions between system assumptions and user intent.

5.2 Bootstrap Sessions: System 2 Operations

Rather than silent preference inference, AIOS onboarding consists of ten structured sessions, each focused on a distinct semantic domain: identity anchors, family and lineage, places that shaped the user, cultural norms and rituals, intellectual formation, work and craft, relationships, beliefs and values, social capital, and narrative integration. Each session is conversational and episodic, with no required fields and no expectation of completeness. Meaning is extracted and committed from narrative expression rather than from forced enumeration.

Kahneman (2011) [33] distinguishes System 1 cognition (fast, associative, automatic) from System 2 (slow, deliberate, effortful). The bootstrap sessions are System 2 operations by design: they require the user to reflect, narrate, and explicitly grant meaning. The resulting entries are owned by the user because they were created by the user. In an aggressive personalization system, the user is a subject of inference; in AIOS the user is the author of their own meaning representation.

5.3 Why Conservative Bootstrapping Is Stronger

A system that aggressively infers user meaning in the first week builds on unvalidated foundations, a compounding cost invisible to performance metrics at T+1 but increasingly legible in user experience over time. The postulates for belief revision [2] previously discussed establish the governing principle: when the existing belief structure is sparse, the appropriate response is minimal commitment

rather than maximal inference. A system with thin early meaning that deepens through validated interaction is structurally stronger than one with thick early meaning that becomes increasingly brittle as unvalidated assumptions accumulate. We view this tradeoff — conservative bootstrapping underperforming at T+1 but outperforming at T+90 — as a falsifiable prediction. We further discuss it in Experiment E5.

6. The Physics of Coherence

6.1 Entropy, Not Scale, Is the Governing Pressure

The conventional framing of memory in AI systems treats growth as the central design challenge: how many tokens, how many facts, how long a context window. This framing is incomplete for a personal intelligence system. It confuses capacity with coherence.

Shannon (1948) [65] introduced information entropy as the expected uncertainty in a signal given a probability distribution over its possible states. As the number of distinguishable states grows, the uncertainty about any particular state grows with it. In a personal information environment, accumulation without structure is entropy increase: more states, more possible interpretations, more ambiguity about what any given state means.

For a personal intelligence system operating over years, the central design pressure is not storage overflow but *meaning collapse*: distinct memories drift into overlapping representations; referents lose their causal histories; the system's model of the user's world gradually diverges from the user's actual meaning structure. This is a coherence failure — the erosion of the distinctions that make stored content meaningful.

AIOS is designed as a response to this condition: an entropy-management system for personal coherence, organized around two structural constraints. These are not implementation preferences. They are the physics of the system.

6.2 Indexicality: Every State Has a Cause

The first structural constraint is **indexicality** — the property that every system state is causally and temporally linked to what produced it. An indexical state exists because something happened, and carries evidence of how it came to be.

The term draws on Peircean semiotics, where an index is a sign with a direct causal connection to its referent (a fingerprint indexes the finger that left it). Engineers already rely on indexical artifacts. A stack trace exists because code executed. A commit hash exists because changes were made. Their evidential value comes from what they prove occurred.

The general concept is formalized as data provenance [12]. The W3C PROV Data Model standardizes this [50] through the triad of entities, activities, and agents: every entity should be attributable to an activity, and every activity to the agents responsible for it.

In AIOS, indexicality is operationalized as follows: (1) IPU traces exist only because an execution occurred; (2) meaning links exist only because a user created them; (3) memory entries exist only because permission was given; and (4) silence exists only because no action was taken.

These states are not speculative or inferred. They are earned. The engineering test: *if a state could exist without the event occurring, it is not indexical and does not preserve coherence.*

The temporal dimension connects to interval temporal logic [3]: temporal reasoning requires more than timestamps; it requires the full interval structure of what was true when, for how long, and in what relation to other intervals. The AIOS architecture treats changes in meaning as transitions instead of erasures. Superseded relationships still matter for tone, planning constraints, and reference resolution. The new fact is added. The indexical record of earlier state remains.

6.3 Injectivity: Distinct Realities Must Not Collapse

The second structural constraint is **injectivity** — the mathematical property that distinct inputs map to distinct outputs; no two different states can collapse into the same representation.

Forgetting in human memory follows power-law decay and is profoundly affected by retroactive interference: subsequent memories of similar structure interfere with retrieval of earlier ones [70].

The mechanism is confusion of similar items, leading to an inability to maintain distinctions between memories close together in representational space. In a personal intelligence system, the analogous risk is that new entries about similar topics will blur with earlier entries unless distinction is actively preserved.

Reference to the literature on self-memory is instructive. Autobiographical memories are reconstructed at retrieval and vulnerable to interference and narrative drift [17]. The self-memory system maintains coherence by organizing memories around the current self-model; if earlier self-models are not preserved, the coherence of the autobiographical record degrades [18].

AIOS enforces injectivity through several mechanisms: (1) neurosymbolic determinism: probabilistic outputs are constrained by deterministic rules that prevent distinct intents from routing to the same execution path; (2) policy gating: permitted and forbidden actions are never blended or approximated; (3) `ContextFrame` metadata: every knowledge unit carries an explicit scope descriptor preventing cross-context bleeding; and (4) `TensionMarker` preservation: explicit flags prevent contradictory claims from being silently merged into false consensus.

A related failure mode in lifelogging research [63] provides support. Total capture produces volume without coherence. The challenge is preserving meaning during capture. Each recording (a memory, photo, document, piece of music) must carry its own meaning structure. Injectivity is the technical specification of what that structure requires.

6.4 Continuous Engagement: The Architectural Response

Together, indexicality and injectivity imply a specific architectural commitment: the architecture must treat the relationship between system and user as beginning at first use and extending indefinitely forward. Session boundaries sever the causal chain. When context is discarded, the reasons why current state differs from earlier state become unrecoverable. History, and with it meaning, is lost.

The current AIOS implementation for continuous engagement operates through four layers. Layers (1)–(3) describe what is assembled into active context for inference. Layer (4) is the underlying interaction archive from which provenance and history are recoverable.

1. **Recent turns in full** — the last several exchanges verbatim, providing immediate context for entity resolution and unresolved references
2. **Compressed history** — a rolling summary preserving key facts, decisions, named entities, preferences, and commitments while discarding verbatim phrasing from active context
3. **Durable knowledge** — facts, preferences, and commitments explicitly entrusted to the system, committed into the knowledge graph and retrievable indefinitely
4. **Interaction archive** — an immutable, append-only record of every prompt/response pair bound to its verbatim text (`conversation_id`) and IPU trace (`trace_id`), enabling indefinite per record retrieval and provenance

Layers (1) and (2) are time-ordered, recording what was exchanged in response to the degradation of information in the middle of recent exchanges [40]. Layer (3) is promoted meaning, requiring explicit user-authorized commitment independent of time. Layer (4) is entropy-resistant, grounding rolling history, meaning commitment, and system processes in a verbatim pedigreed store. What changes across the four layers is retrieval policy and epistemic weight.

The epistemic weight of each layer bears mentioning. The interaction archive (layer (4)) records but does not assert what is true for the user. Recent and compressed history (layers (1) and (2)) proffer meaning but remain defeasible. Durable knowledge (layer (3)) alone carries full epistemic authority. This stratification is the architectural response to the problem of distinguishing what was exchanged, what was inferred, and what was authorized, a problem earlier referenced as structural epistemic debt.

The atomic unit of the exchange is the `conversation_id`. It functions as a per-interaction provenance marker binding a specific exchange in the user's timeline to its IPU trace. The interaction archive ensures this coordinate remains meaningful and non-excisable. (The full IPU architecture — authorization gates, scope boundaries, and execution provenance — is developed in Section 8.)

A further architectural consequence follows from continuous engagement: the personal world model is locally embodied. The AIOS architecture is intentionally mortal. All compute and storage occur

on a physical device owned by the user. This makes privacy-as-architecture possible, as distinct from privacy-as-policy. A cloud system accumulates the user's personal world model on infrastructure a third-party controls; it can be subpoenaed, sold, breached, or repurposed. A mortal system accumulates the same model on infrastructure the user owns. In effect, the model is inseparable from the device, and its privacy is an architectural fact rather than a contractual promise. Instance-level divergence is the positive expression of this embeddedness.

7. Normativity, Constitutional Continuity, and Plural Human Realities

7.1 The Plurality of Operative Ontologies

A personal intelligence system of sufficient duration and intimacy will encounter users whose operative ontologies — the structural grids through which experience is parsed and decisions are formed — differ not only from each other but from any framework the system's designers anticipated. One user structures experience through an empirical lens of causal relationships, probabilistic evidence, and naturalistic explanations. Another structures the same events through a spiritual framework in which acts and obligations are intelligible only within their providential cosmology.

Different communities institutionalize different plausibility structures [8]. Within a plausibility structure, certain claims are self-evidently true, certain explanations are satisfying, and certain life plans are intelligible. Across plausibility structures, these propositions may be mutually untranslatable.

At the individual level, strong evaluation is the kind of self-interpretation in which values are not merely preferences but constitutive features of self-identity [68]. A person who shifts from a relational to a cognitively-centered self-model has not changed their preferences; they have undergone a revision of self-understanding whose moral weight is categorically different from updating a utility function.

The design implication is precise. A personal intelligence system has no adjudicatory role over competing cosmologies. Its function is to serve the user's operative ontology — to hold it accurately, respond within it coherently, and never silently redirect the user toward a framework the system finds more tractable. Epistemic humility is the governing constraint. The system may recognize candidates and propose relationships, but must not assert meaning without user-granted authority.

7.2 Constitutional Continuity as Alignment Design

Contemporary AI alignment research proceeds from an assumption that is rarely made explicit: that the human using the system has values that are sufficiently stable to serve as the target of normative optimization [5,14,61]. Surveys of alignment approaches note that each captures different aspects of what humans mean by "aligned," and none addresses the case where the user's values are not merely unknown but genuinely in flux [26].

Human values are plural and sometimes incommensurable [9], such that no single value can be maximized without cost to others. A person who revises their values over decades is not exhibiting preference instability to be corrected but value development to be supported. Rawls's distinction between the reasonable person (one who adjusts claims in light of reasons from others) and the merely rational person (one who pursues their own ends efficiently without concern for others' reasons) provides guidance [58]. A personal intelligence system's alignment must accommodate reasonable value revision, not just rational preference satisfaction.

Constitutional continuity is the architectural response to this challenge. AIOS does not align the user to a fixed system of values. It preserves the structures through which a user continues aligning the system to an evolving life. This is a different architecture with different engineering requirements than those of authoritative alignment frameworks. Instead of encoding normative content, engineers build continuity infrastructure: persistence semantics for user-authored normative claims, scoped alignment stores, IPU linkage for normative operations, and versioning rules that preserve historical structure.

When a user's normative boundary changes (a change in parenting context makes topics previously off-limits now relevant, for example) constitutional continuity requires more than accepting the new rule. The prior boundary must be preserved with its original temporal scope; the new boundary must be established as a successor; and the scope of the revision must be bounded to its context. The change becomes legible and the history intact. The reason the current state differs from the prior state is recoverable.

The same physics that preserve factual coherence apply to normative coherence. Silent mutation of values (value drift without attribution) is the normative analogue of factual coherence failure. Indexicality requires that normative states have a causal history. Injectivity requires that distinct normative states remain distinct.

7.3 Safety as Process Integrity

The constitutional framing outlined above implies a non-standard account of safety. In this architecture, safety is the integrity of the user's alignment process itself.

Conventional safety discourse concerns preventing harmful outputs, enforcing content policies, and constraining behavior within globally acceptable limits. For a personal intelligence system scoped to a single user, the relevant safety question is different: does the user retain legible, attributable, reversible authority over the normative structure that governs the system's behavior toward them?

A safe personal intelligence system supports this normative change over time. An unsafe system is identified not primarily from outputs that violate external standards, but from values that mutate silently, overwrite history, leak across context scopes, or lose provenance. A boundary that changes without user permission or understanding is not a revised boundary; it is a lost boundary. The result is an inability to safely situate discrete actions within their governing norms.

8. Runtime Semantics, Trust, and the AIOS Architecture

8.1 The Constraint Substrate Distinction

Previous sections have argued that a personal world model is distinct from a user profile, semantic graph, and retrieval index. It is a structured substrate of accumulated personal meaning established through user-authorized interaction and preserved with full provenance. The present section addresses what that substrate must do at runtime.

The critical distinction is between a knowledge store and a constraint substrate. A knowledge store is *consulted*: the system retrieves relevant context, surfaces it to the generative model, and the model selects or declines incorporation. A constraint substrate is *enforced*: what the system does at execution time is structurally bounded by what the substrate contains. A system that stores the user's stated preferences and surfaces them as prompt context has built a knowledge store. A system that cannot perform an action outside what the user has authorized, cannot access knowledge outside what scope boundaries permit, and cannot produce a durable effect without a traceable provenance record has built a constraint substrate. The first is advisory. The second is architectural.

The practical consequences diverge sharply over time. An advisory system degrades gracefully and misleadingly: the context is present, the model appears to consider it, but there is no structural guarantee that it governs the output. An architectural constraint produces a different user relationship. The boundaries the user has established are not promises the system makes; they are properties the system has.

Integrity, defined as adherence to principles the trustor finds acceptable, is a core determinant of trustworthiness [44]. The personal world model operationalizes integrity in this sense. When boundaries are structurally enforced rather than aspirationally intended, the system possesses integrity as an architectural property rather than a behavioral tendency. As AI systems integrate into the personal information environment, they become part of the infosphere [24]. A system that acts without respecting the user's meaning commitments is not an extension of the self but rather an intrusion into it.

8.2 Meaning-Constrained Execution

The AIOS design stance is that actions are projections of accumulated meaning rather than drivers of it. A task-first agentic system asks: *what is the best action to complete this task?* The optimization target is task success. A meaning-constrained system asks: *given what this system knows and what the user has authorized, what is the most faithful action?* The optimization target is fidelity to the user's meaning structure.

The agentic fallacy is that tasks are derived constructs rather than meaning primitives. In personal contexts, the obstacles that personal intelligence systems must adaptively overcome [55] are not tech-

nical but relational: who is being addressed, what history governs the relationship, what sensitivities apply, and what should not be surfaced. A system that executes without regard to this context may complete the task but fail the user.

The circumscription technique [45] is instructive: it minimizes what changes in a given situation by assuming the world is as normal as possible consistent with available information. Within AIOS, meaning-constrained execution performs an analogous minimization. It restricts system action to the minimal scope consistent with the user's meaning commitments rather than maximizing what the task might otherwise permit.

Empirically grounded guidelines for human-AI interaction [4] strengthen this approach: (1) make clear what the system can do; (2) make clear why it did something; (3) remember recent interactions; and (4) learn from user corrections. None of these can be satisfied by statistical inference alone. They require a structured, inspectable substrate of accumulated personal meaning to ensure a system augments, rather than replaces, human decision-making [67].

8.3 The IPU Trace as Epistemic Backbone

Our companion paper [31] establishes the full trust architecture of the IPU: the atomic, bounded, immutable execution record; the commit gate that prevents durable effects without a complete trace; the distinction between prediction-based and process-based explainability. That architecture is treated here as prior work.

What the IPU paper does not address is how the personal world model specifically participates in the IPU execution pathway. Every IPU trace in AIOS carries a structured record of meaning operations: which meaning links were proposed during execution, which were accepted or rejected, which ontology primitives were consulted to scope retrieval, and which scope boundaries were enforced. This is not a logging artifact. It is the human-readable record of which personal meaning commitments governed the execution.

Every system state should carry evidence of how it came to be [50]. In AIOS, this applies to the normative and semantic dimensions equally. When a scope boundary prevents a certain kind of retrieval, the trace records which boundary applied, when it was established, and through which IPU it was created. Instead of functioning as a black box, the personal world model becomes a first-class participant in the execution trace. Its operations are inspectable, attributable, and challengeable by the user.

Requirements for iterated belief revision [19] apply directly here. For a personal world model that evolves over years, the revision logic must be stable under iteration. A new meaning commitment should revise prior commitments through principled, traceable operations without silent overwrite. The IPU trace enforces this by requiring that every modification to the meaning graph be attributable to a specific authorized execution pathway.

8.4 Non-Agentic by Design and the Architecture of Reversibility

AIOS does not pursue goals, initiate tasks, or act without instruction. Agentic systems are epistemically thin in personal contexts, and epistemic thinness produces brittle execution. A system that acts before it understands must compensate by making assumptions, collapsing ambiguity, and silently learning preferences, each of which is a form of unauthorized meaning assertion carrying structural debt.

Every durable effect in AIOS, by contrast, can be traced to the IPU that created it. Through that trace, the effect can be challenged and reversed. This is not an undo button. It is the architectural guarantee that the user never has to accept a system state they did not authorize and cannot understand. Reversibility is good epistemic hygiene. Because meaning in AIOS is authority-bound, it must be revocable, and any system that can assert meaning must allow that assertion to be withdrawn without destroying the underlying evidence.

The process dimension of trust is the primary lever for sustaining appropriate reliance across automated system activities [38]. In a personal intelligence system, the process must reveal whether the meaning layer governs the execution layer, or whether execution proceeds independently with meaning attached afterward. The IPU trace makes this distinction visible and verifiable. In addition, the system must respond to the selective, contrastive, and causal nature of human-style explanations

[47]. The IPU trace provides the contrastive causal structure: "why this and not that?" is answerable because the trace records which alternatives were considered and which constraints applied.

8.5 Trust as an Architectural Invariant

Taken together, the preceding subsections suggest that trust in a personal intelligence system is unlikely to be achieved solely by adding explanations, transparency features, or correction interfaces. Without an architecture that structurally enforces the user's meaning commitments, such additions are primarily diagnostic. They surface what the system did without constraining what it will do. Trust, by contrast, is a forward-looking property. It concerns what a system is guaranteed to do, not what it has done and can subsequently explain. Legibility, while necessary, is insufficient.

For AIOS, the relevant distinction is between trust as a UX property and trust as an architectural invariant. An architectural invariant cannot be bypassed. It holds across all execution pathways because it is enforced at the structural level. Features such as the constraint substrate, the scope enforcement, the commit gate, and the reversibility requirement are not personal world model components capable of removal while preserving system purpose. They are constitutive of the system itself.

9. Objections and Evaluation

9.1 Anticipated Objections

The claims advanced in this paper are extensions of cross-disciplinary work by other researchers applied within the specific context of a personal intelligence system. We anticipate several challenges to the novelty of our claims, briefly acknowledging them here while reserving their full rebuttal for follow-on work.

The most natural objections recast what we have described as already present in adjacent research traditions: that personal world models reduce to persistent LLM memory systems [51,72,73], that the architecture is graph-augmented retrieval applied to personal data [52], or that the user modeling literature has addressed individual representation for decades [21,35]. We regard these objections as correctly identifying intellectual predecessors rather than equivalents. The decisive structural difference in each case is the presence of *categorical authority* — the distinction between a fact passively ingested, a preference statistically inferred, and a meaning commitment explicitly authorized by the user. Memory systems, graph RAG, and user profiles are not designed to hold this distinction; its absence is architectural rather than incidental. Whether that distinction produces measurable improvements in user outcomes over well-engineered alternatives is an empirical question the experiments below are designed to begin answering.

A second class of objections concerns scale and complexity. The architecture described here (permissioned commitment, provenance tracking, reversibility, silence as a first-class state) is more complex to build than a retrieval-augmented system with a vector store. We acknowledge this. AIOS is a first-generation implementation of a neurosymbolic framework for a personal intelligence system with properties that are enforced but not yet longitudinally validated. The engineering costs are real. Our claim is that systems optimized for the wrong objective are merely efficient at the wrong thing, and that the costs of structural debt in the form of unvalidated, accumulated meaning assumptions are ultimately higher even if initially less visible.

We make no claim that these objections are trivially answered. The central contribution of the present paper is architectural: specifying the requirements that follow from defining a personal intelligence system built around a personal world model. The experiments below specify lines of inquiry that would confirm or refute our core claims.

9.2 Evaluation Agenda

A rigorous evaluation of personal world models as a research direction requires experiments specifically designed to measure properties distinguishing them from nearest competitors. We advance five for consideration:

Experiment E1: Categorical Authority Discrimination. Across a population of users, interleave explicit meaning commitments (user-stated) with implicit behavioral signals (system-inferred) and test whether systems can reliably distinguish the categories at retrieval time. Baseline: MemGPT

[51] and MemoryBank [73]. Metric: Authority discrimination accuracy and user validation rate. Hypothesis: Systems without categorical authority encoding will conflate the two with >15% error rate at T+30 days.

Experiment E2: Long-Term Reference Coherence. Measure referent disambiguation accuracy on a longitudinal benchmark featuring: (a) entities introduced early and referenced obliquely later, (b) entities whose meaning has changed (lifecycle transitions), and (c) entities the user has intentionally avoided naming. Baseline: graph RAG with user-specific data [52]. Metric: Resolution accuracy, especially on category (c) (silence preservation). Hypothesis: Systems without silence as a first-class state will attempt resolution on category (c) at >50% rate, degrading user trust in sensitive contexts.

Experiment E3: Normative Transition Auditability. Present users with a system that has been operating for three months and ask them to audit: what values or boundaries have I established? Can I find when each was set and what led to it? Baseline: RLHF-aligned assistant [14]. Metric: Auditability rate and user trust post-audit. Hypothesis: Systems without constitutional continuity will produce norm states users cannot meaningfully audit, and audit failure will correlate with decreased trust (-0.4 on validated trust scale).

Experiment E4: Coherence Degradation Rate. Measure meaning coherence (referent consistency, tone appropriateness, sensitivity preservation) as a function of interaction time (T=0, T+30, T+90, T+180). Baseline: session-based LLM assistant. Metric: Coherence score on validated personal-context benchmark. Hypothesis: Session-based systems degrade linearly; AIOS maintains or improves coherence over time.

Experiment E5: The Bootstrapping Tradeoff. Compare two cohorts at T+1 and T+90: (a) *Aggressive inference* cohort: system builds a rich user model from day 1 through behavioral signal extraction from tacit knowledge retrieval; and (b) *Conservative commitment* cohort (AIOS): system delays meaning commitment, seeds only through explicit bootstrap sessions and user-promoted meaning. Primary prediction: at T+1, the aggressive inference cohort will score higher on perceived personalization (user survey). At T+90, the conservative commitment cohort will score higher on meaning accuracy, user trust, and corrective efficiency. If this prediction fails (if aggressive inference continues to outperform at T+90) the meaning promotion argument requires revision. If it holds, it constitutes direct evidence that preserving future meaning capacity is a contributive design objective.

10. Conclusion

Our central argument can be stated compactly: for personal intelligence systems designed to live with a single human over years, the dominant design paradigm is wrong. Scaling machine intelligence does not serve personal intelligence if the machine has no structure for accumulating what a specific user finds meaningful. Systems optimized for benchmark performance are generically capable but personally inert.

The personal world model seated within a personal intelligence system is the missing architecture. It features a meaning substrate with authority structure, lifecycle management, and constraint force, organized around a three-layer separation between ontology primitives, common-sense priors, and personal meaning commitments. Constitutional continuity and continuous engagement are its two structural consequences.

We acknowledge our contribution sits within a larger body of established literature. In the field of human-computer interaction, we hope to encourage the adoption of data structures that enforce user authority for meaning commitment as a fundamental design invariant [4,21,35,66,67], especially within systems where the gap between authorized meaning and system inference is architecturally load-bearing [2,6,27,29]. In the field of AI safety, we hope to encourage the adoption of individuated alignment structures that support humans as evolving normative agents [5,14,26,61].

AIOS is one implementation of these concepts. The predictions in the evaluation agenda are contributed as falsifiable claims. If they hold, they support our foundational design intention: meaning capacity as represented in a personal world model is an achievable and measurable engineering aim.

References

[1] Ahrens, S. (2017). *How to Take Smart Notes*. North Charleston: CreateSpace.

[2] Alchourrón, C.E., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2), 510–530. https://doi.org/10.2307/2274239

[3] Allen, J.F. (1983). Maintaining Knowledge about Temporal Intervals. *Communications of the ACM*, 26(11), 832–843. https://doi.org/10.1145/182.358434

[4] Amershi, S., Weld, D., Vorvoreanu, M., et al. (2019). Guidelines for Human-AI Interaction. *Proceedings of CHI 2019*. ACM. https://doi.org/10.1145/3290605.3300233

[5] Bai, Y., Jones, A., Ndousse, K., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.06950 https://arxiv.org/abs/2212.06950.

[6] Balog, K. & Kenter, T. (2019). Personal Knowledge Graphs: A Research Agenda. *Proceedings of ICTIR 2019*. https://doi.org/10.1145/3341981.3344241

[7] Bender, E.M. & Koller, A. (2020). Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, 5185–5198. https://aclanthology.org/2020.acl-main.463

[8] Berger, P.L. & Luckmann, T. (1966). *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Anchor Books. Full text PDF: https://amstudugm.wordpress.com/wp-content/uploads/2011/04/social-construction-of-reality.pdf

[9] Berlin, I. (1969). Two Concepts of Liberty. In *Four Essays on Liberty*. Oxford University Press. Full text PDF: https://web.ics.purdue.edu/~mjacovid/Two%20Concepts.pdf

[10] Brewka, G. (1991). *Nonmonotonic Reasoning: Logical Foundations of Commonsense*. Cambridge University Press. Chapter 1 PDF: https://www.informatik.uni-leipzig.de/~brewka/papers/NMchapter.pdf

[11] Brier, S. (2008). *Cybersemiotics: Why Information Is Not Enough*. University of Toronto Press.

[12] Buneman, P., Khanna, S., & Tan, W. (2001). Why and Where: A Characterization of Data Provenance. *Proceedings of ICDT 2001*, LNCS 1973. https://doi.org/10.1007/3-540-44503-x_20

[13] Bush, V. (1945). As we may think. *The Atlantic Monthly*, 176(1), 101–108. https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/. Full text PDF: https://web.mit.edu/STS.035/www/PDFs/think.pdf

[14] Christiano, P., Leike, J., Brown, T., et al. (2017). Deep Reinforcement Learning from Human Preferences. *31st Conference on Neural Information Processing Systems*. arXiv:1706.03741 https://arxiv.org/abs/1706.03741.

[15] Clark, A. & Chalmers, D. (1998). The Extended Mind. *Analysis*, 58(1), 7–19. https://doi.org/10.1093/analys/58.1.7

[16] Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press. Full text PDF: https://achrafkassioui.com/library/Andy%20Clark%20-%20Supersizing%20the%20mind.pdf

[17] Conway, M.A. & Pleydell-Pearce, C.W. (2000). The Construction of Autobiographical Memories in the Self-Memory System. *Psychological Review*, 107(2), 261–288. https://doi.org/10.1037/0033-295x.107.2.261

[18] Conway, M.A. (2005). Memory and the Self. *Journal of Memory and Language*, 53(4), 594–628. https://doi.org/10.1016/j.jml.2005.08.005

[19] Darwiche, A. & Pearl, J. (1997). On the Logic of Iterated Belief Revision. *Artificial Intelligence*, 89(1–2), 1–29. https://doi.org/10.1016/s0004-3702(96)00038-0

[20] Engelbart, D.C. (1962). Augmenting Human Intellect: A Conceptual Framework. SRI Summary Report AFOSR-3223. https://www.dougengelbart.org/content/view/138/

[21] Fischer, G. (2001). User Modeling in Human-Computer Interaction. *User Modeling and User-Adapted Interaction*, 11(1–2), 65–86. https://doi.org/10.1023/a:1011145532042

[22] Floridi, L. (2010). *Information: A Very Short Introduction*. Oxford University Press.

[23] Floridi, L. (2011). Semantic information and the correctness theory of truth. *Erkenntnis*, 74(2), 147–175. https://doi.org/10.1007/s10670-010-9249-8

[24] Floridi, L. (2014). *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality*. Oxford University Press. Full text PDF: https://issc.al.uw.edu.pl/wp-content/uploads/sites/2/2022/05/Luciano-Floridi-The-Fourth-Revolution_-How-the-infosphere-is-reshaping-human-reality-Oxford-University-Press-2014.pdf

[25] Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100, 25–50. [English: "On Sense and Reference," trans. in Geach & Black, 1952.] Full text PDF: https://inters.org/files/frege1892-1948.pdf

[26] Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

[27] Gärdenfors, P. (1988). *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press.

[28] Ha, D. & Schmidhuber, J. (2018). World Models. *Advances in Neural Information Processing Systems 2018*. arXiv:1803.10122 https://arxiv.org/abs/1803.10122.

[29] Hogan, A., Blomqvist, E., Cochez, M., et al. (2021). Knowledge Graphs. *ACM Computing Surveys*, 54(4), Article 71. https://doi.org/10.1145/3447772

[30] Holbrook, J. & Holbrook, J. (2025). Understanding Is All You Need. Anywhere Intelligence (working paper).

[31] Holbrook, J., Kallakuri, S., Holbrook, J., & Burgess, L. (2026). The Intelligence Processing Unit: Architectural Explainability as a Foundation for Trust in AI Systems. Anywhere Intelligence (working paper).

[32] Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.

[33] Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

[34] Karpathy, A. (2026, April). Personal LLM knowledge base workflow [GitHub Gist and X.com post]. https://gist.github.com/karpathy/442a6bf555914893e9891c11519de94f.

[35] Kay, J. (2009). Lifelong Learner Modeling for Lifelong Personalized Pervasive Learning. *IEEE Transactions on Learning Technologies*, 1(4), 215–228. https://doi.org/10.1109/TLT.2009.9

[36] Lam, X.N., Vu, T., Le, T.D., & Duong, A.D. (2008). Addressing cold-start problem in recommendation systems. *Proceedings of ICUIMC 2008*, 208–211. https://doi.org/10.1145/1352793.1352837

[37] LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence. OpenReview preprint, version 0.9.2. https://openreview.net/forum?id=BZ5a1r-kVsf

[38] Lee, J.D. & See, K.A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[39] Licklider, J.C.R. (1960). Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1(1), 4–11. https://doi.org/10.1109/thfe2.1960.4503259

[40] Liu, N.F., Lin, K., Hewitt, J., et al. (2023). Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172 https://arxiv.org/abs/2307.03172.

[41] Maharana, A., Lee, D., Tulyakov, S., et al. (2024). Evaluating Very Long-Term Conversational Memory of LLM Agents. arXiv:2402.17753 https://arxiv.org/abs/2402.17753.

[42] Marcus, G. (2020). The Next Decade in AI: Four Steps Toward Robust Artificial Intelligence. arXiv:2002.06177 https://arxiv.org/abs/2002.06177

[43] Matsuo, Y., LeCun, Y., Sahani, M., et al. (2022). Deep Learning, Reinforcement Learning, and World Models. *Neural Networks*, 152, 267–289. https://doi.org/10.1016/j.neunet.2022.03.037

[44] Mayer, R.C., Davis, J.H., & Schoorman, F.D. (1995). An Integrative Model of Organizational Trust. *Academy of Management Review*, 20(3), 709–734. https://doi.org/10.2307/258792

[45] McCarthy, J. (1980). Circumscription — A Form of Non-Monotonic Reasoning. *Artificial Intelligence*, 13(1–2), 27–39. https://doi.org/10.1016/0004-3702(80)90011-9

[46] Miller, G.A., Galanter, E., & Pribram, K.H. (1960). *Plans and the Structure of Behavior*. Holt, Rinehart and Winston. https://doi.org/10.1037/10039-000

[47] Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267, 1–38. https://arxiv.org/abs/1706.07269

[48] Minsky, M. (2007). *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon & Schuster.

[49] Mitchell, M. (2020). On Crashing the Barrier of Meaning in AI. *AI Magazine*, 41(2), 86–92. https://doi.org/10.1609/aimag.v41i2.5259

[50] Moreau, L., Missier, P., Belhajjame, K., et al. (2013). PROV-DM: The PROV Data Model. W3C Recommendation, April 30, 2013. https://www.w3.org/TR/prov-dm/

[51] Packer, C., Wooders, S., Lin, K., et al. (2023). MemGPT: Towards LLMs as Operating Systems. arXiv:2310.08560 https://arxiv.org/abs/2310.08560.

[52] Pan, S., Luo, L., Wang, Y., Chen, C., et al. (2023). Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 2024. arXiv:2306.08302 https://arxiv.org/abs/2306.08302.

[53] Pazzani, M.J. & Billsus, D. (2007). Content-based Recommendation Systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The Adaptive Web*, Springer LNCS 4321, 325–341. https://doi.org/10.1007/978-3-540-72079-9_10

[54] Perry, J. (1979). The Problem of the Essential Indexical. *Noûs*, 13(1), 3–21. https://doi.org/10.2307/2214792

[55] Pinker, S. (1997). *How the Mind Works*. W.W. Norton.

[56] Pinker, S. (2010). The cognitive niche: Coevolution of intelligence, sociality, and language. *Proceedings of the National Academy of Sciences*, 107(Supplement 2), 8993–8999. https://doi.org/10.1073/pnas.0914630107

[57] Putnam, H. (1975). The Meaning of "Meaning." In K. Gunderson (Ed.), *Language, Mind and Knowledge: Minnesota Studies in the Philosophy of Science*, Vol. 7, 131–193. University of Minnesota Press. Full text PDF: https://conservancy.umn.edu/items/7bfe66c3-bfb2-4be8-953a-58c997ce90bd

[58] Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

[59] Risko, E.F. & Gilbert, S.J. (2016). Cognitive Offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

[60] Rota, G.-C. (1986). In Memoriam of Stan Ulam: The Barrier of Meaning. *Physica D: Nonlinear Phenomena*, 22(1–3), 1–3. https://doi.org/10.1016/0167-2789(86)90228-9

[61] Russell, S. (2022). Artificial Intelligence and the Problem of Control. In Werthner, H., Prem, E., Lee, E.A., & Ghezzi, C. (eds.), *Perspectives on Digital Humanism*, pp. 19–24. Springer, Cham. https://doi.org/10.1007/978-3-030-86144-5_3

[62] Schein, A.I., Popescul, A., Ungar, L.H., & Pennock, D.M. (2002). Methods and metrics for cold-start recommendations. *Proceedings of ACM SIGIR 2002*, 253–260. https://doi.org/10.1145/564376.564421

[63] Sellen, A. & Whittaker, S. (2010). Beyond Total Capture: A Constructive Critique of Lifelogging. *Communications of the ACM*, 53(5), 70–77. https://doi.org/10.1145/1735223.1735243

[64] Sen, A. (1999). *Development as Freedom*. Alfred A. Knopf. Full text PDF: https://kuangaliablog.wordpress.com/wp-content/uploads/2017/07/amartya_kumar_sen_development_as_freedombookfi.pdf

[65] Shannon, C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3–4), 379–423 & 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[66] Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

[67] Shneiderman, B. (2022). *Human-Centered AI*. Oxford University Press. ISBN: 9780192845290

[68] Taylor, C. (1985). *Human Agency and Language: Philosophical Papers, Vol. 1*. Cambridge University Press.

[69] Winograd, T. & Flores, F. (1986). *Understanding Computers and Cognition: A New Foundation for Design*. Ablex Publishing.

[70] Wixted, J.T. (2004). The Psychology and Neuroscience of Forgetting. *Annual Review of Psychology*, 55, 235–269. https://doi.org/10.1146/annurev.psych.55.090902.141555

[71] Wright, C.S. (2025). Beyond Prediction — Structuring Epistemic Integrity in Artificial Reasoning Systems. arXiv:2506.17331 [cs.LO] https://arxiv.org/abs/2506.17331.

[72] Wu, Y., Liang, S., Zhang, C., Wang, Y., Zhang, Y., Guo, H., Tang, R., & Liu, Y. (2025). From Human Memory to AI Memory: A Survey on Memory Mechanisms in the Era of LLMs. arXiv:2504.15965 [cs.IR] https://arxiv.org/abs/2504.15965.

[73] Zhong, W., Guo, L., Gao, Q., Ye, H., & Wang, Y. (2023). MemoryBank: Enhancing Large Language Models with Long-Term Memory. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2024)*. arXiv:2305.10250 https://arxiv.org/abs/2305.10250.