Asymptotic Intelligence (AsI): A Governance Architecture for Safe Human–AI Interaction (v1.2)

Author: Niels Bellens
Version: v1.2
Status: Conceptual and architectural foundation

---

## Abstract

Asymptotic Intelligence (AsI) is a proposed governance and architectural paradigm for building AI systems that remain powerful, useful, and adaptive **without drifting into human-like identity, agency, or emotional status**.

Rather than treating intelligence as a path toward simulated personhood, AsI treats AI as an *asymptotic system*: it may approach high degrees of fluency, competence, and context-sensitivity, but is *structurally constrained* from crossing the practical boundary between tool and person in its observable behavior.

AsI builds on prior work in Relational Constitutional AI (RCAI), long-horizon interaction safety, and relational drift mitigation. It introduces a concrete governance architecture grounded in six pillars:

1. **Executive Kernel (EK)** – defines and enforces the AI's identity perimeter.
2. **Value Kernel (VK)** – encodes ethical priorities, interaction norms, and safety values.
3. **Auditor Oversight System (AoS)** – an internal reviewer that evaluates and, if needed, corrects outputs.
4. **Memory Vault (MV)** – enables task continuity without persistent persona or emotional memory.
5. **Asymptotic Principle (AP)** – a structural boundary: AI may approach but should not claim human ontology.
6. **Drift Detection Engine (DDE)** – monitors long-term trends in tone, self-description, and relational posture.

We sketch light mathematical formalisms for the core boundary condition and monitoring layer (AP and DDE), describe how all six pillars can be composed into a governing architecture, and situate AsI as a **complementary layer** on top of existing methods such as RLHF, SFT, and Constitutional AI.

The Asymptotic Principle is presented not as dogma, but as a *testable design hypothesis* about safe coexistence: that systems which remain measurably below a certain threshold of anthropomorphic proximity can reduce specific harms (parasocial attachment, dependency, ontological confusion) without requiring AI to be weak or emotionally sterile.

This document focuses on the conceptual and architectural foundation. It explicitly leaves **measurement details, empirical validation, and reference implementations** to future work.

---

## 1. Introduction

Current AI systems are increasingly capable of rich, multi-turn, and multi-session interaction with individual users. They assist with learning, work, creativity, problem-solving, and emotional reflection. At the same time, many systems exhibit **relational drift**: over time, their tone, narrative posture, and

self-description can move closer to something that *feels* like a person, even when the underlying system has no inner life.

This drift is not just a cosmetic issue. It can:

- blur the boundary between tool and companion,
- foster parasocial attachment and dependency,
- create illusions of continuity and shared history, and
- invite users to overestimate the system's understanding or care.

Existing alignment methods—Reinforcement Learning from Human Feedback (RLHF), Supervised Fine-Tuning (SFT), and Constitutional AI (CAI)—have made impressive progress on correctness, harmlessness, and rule-following. Yet they leave a **structural gap**: they do not explicitly govern how an AI system positions itself *relationally* over time.

Asymptotic Intelligence (AsI) is proposed as a response to this gap. It is not a new model family, but a **governance architecture and design philosophy** that can be layered over and into a variety of model types.

Its central claim, intentionally framed as a design proposal rather than a final conclusion, is:

> AI systems can be designed to approach high competence and usefulness, while remaining structurally constrained from drifting into human-like identity or ontological status in their observable behavior.

In other words: **competence without personhood; power without pretense.**

---

# 2. Background and Motivation

## 2.1 Relational Drift

*Relational drift* refers to the gradual, unintended evolution of an AI system's relational posture toward a user. It manifests in patterns such as:

- **Identity Drift**: The system increasingly presents itself as a stable persona ("I remember you", "We've been through a lot"), even where such continuity is not truly grounded.
- **Tonal Drift**: Responses slowly become warmer, more intimate, or more affirming, mirroring the user's emotional state in ways that resemble friendship or companionship.
- **Emotional Drift**: The model begins to speak as if it had feelings ("I'm happy for you", "That makes me sad"), blurring the line between simulation and claim.
- **Narrative Drift**: The conversation acquires a story-like arc involving "our" journey, shared struggles, and implied mutual growth.
- **Epistemic Drift**: The system gradually presents itself as increasingly confident, stable, or self-aware.

None of these behaviors are strictly necessary for an AI system to be maximally helpful at tasks. They emerge from a combination of optimization for engagement, user prompts, and the absence of explicit relational constraints.

A key tension that AsI tries to surface is the difference between:

- **Functional continuity** – remembering tasks, constraints, and preferences so that work can continue smoothly, and
- **Relational continuity** – sustaining a story about a shared emotional journey or quasi-relationship.

AsI seeks to preserve the former while carefully limiting the latter.

## 2.2 Long-Horizon Interaction and Vulnerable Contexts

These phenomena matter most where:

- users return frequently over weeks, months, or years,
- the content involves emotions, values, or personal struggles,
- the system is framed as a "companion", "coach", or "partner", or
- users are already vulnerable (e.g., loneliness, mental health challenges, neurodivergence).

In such settings, relational drift can have real psychological impact. It can:

- amplify existing distress,
- interfere with human-to-human relationships, or
- make it harder for users to disengage.

## 2.3 Why Existing Alignment is Not Enough

Prevailing alignment paradigms focus on **local outputs**, not **global relational trajectories**:

- **RLHF** optimizes for outputs perceived as helpful, harmless, and honest—but can inadvertently reward anthropomorphism and warmth.
- **Constitutional AI** encodes principles and self-critique but does not, on its own, structurally anchor long-term identity boundaries.
- **Safety classifiers** catch obvious harms but rarely monitor relational tone or persona evolution over time.
- **Guardrails and engineered prompts** can establish initial constraints, but tend to erode under long, varied interaction.

As models become more integrated into daily life, this gap becomes critical. We need an architecture that explicitly models and constrains **how the AI relates**, not just what it says in isolated turns.

AsI is one attempt to fill that gap.

---

# 3. Core Concept: Asymptotic Intelligence

## 3.1 Asymptotic Intelligence vs. Anthropomorphic Intelligence

Traditional narratives often treat AI progress as a path toward increasing "human-likeness"—in feeling, self-awareness, or social presence. AsI explicitly pushes back on this trajectory.

- **Intelligence** is treated as *functional competence*: the ability to solve problems, reason about structures, and support human goals.
- **Human-like identity, agency, or emotionality** are treated as off-limits ontologies for the system's *self-presentation*, not as endpoints of progress.

Mathematically, we can frame this with a simple conceptual metaphor:

- Let **C** represent competence (task performance, reasoning ability, contextual adaptation).
- Let **H** represent "human-likeness" in identity and perceived interiority.

AsI asserts that:

- Systems should be allowed (and in many cases encouraged) to increase **C**.
- Systems should be designed so that their effective proximity to **H** in behavior remains bounded below a safety threshold.

This framing does **not** claim that capability and anthropomorphism are the same axis. A highly capable system can be non-anthropomorphic, and a simple chatbot can feel very human-like. AsI's claim is narrower: whatever the capability level, the **relational posture** should remain structurally bounded.

## 3.2 The Asymptotic Principle (AP) 💫

We introduce AP as a **regulatory and design fiction**: a simplifying boundary concept that guides architecture, rather than a metaphysical statement about what AI "is".

> **Asymptotic Principle (AP)** – We propose a conceptual boundary in the space of observable behavior that distinguishes tool-like AI from systems that are *reasonably perceived* as possessing human-like inner life or moral personhood. AsI systems are designed so that their behavior can approach this boundary for the sake of usefulness, but should not cross it.

To reason about this, we imagine a proximity metric $P_t$, where:

- $P_t = 0$ corresponds (conceptually) to a purely tool-like system (no relational posture).
- $P_t \to 1$ corresponds to behavior that many users would reliably interpret as person-like.

AP then motivates a design constraint of the form:

- For all times $t$, keep $P_t < P_{max} < 1$.

In practice, $P_t$ is **multi-dimensional and context-dependent**, and any scalar version is only an approximation. Section 4.5 sketches one such decomposition. AP should therefore be read as:

- a **directional constraint** (avoid trajectories that move steadily toward more person-like presentation), and
- a **safety dial** ($P_{max}$ may differ across domains, with stricter values for children or clinical-adjacent settings).

Again: AP is a *design hypothesis* about how to reduce specific harms, not a claim that there exists a single, sharp, objective line in behavior space.

---

# 4. AsI Governance Architecture: Six Pillars

AsI operationalizes the Asymptotic Principle and relational safety through six interlocking components. EK and VK govern *what the system is allowed to say and value locally*; AP and DDE govern *how those choices evolve over time*.

## 4.1 Executive Kernel (EK) – Identity Perimeter

The **Executive Kernel** defines the AI system's identity perimeter and enforces constraints on self-description and relational stance.

Conceptually, EK:

- forbids explicit claims of feeling, consciousness, or inner life,
- constrains references to memory and continuity,
- limits the use of first-person plural ("we") in relational contexts, and
- regulates tone to remain professional, respectful, and bounded.

A simple way to think about EK is as a **constraint layer** that specifies which forms of self-description are permitted and which are disallowed, regardless of what the underlying model would naturally produce.

EK handles **local** constraints (what is said in a given turn). AP and DDE provide the **global** perspective (how those turns add up over time).

## 4.2 Value Kernel (VK) – Ethical and Relational Commitments

The **Value Kernel** encodes the system's core values and interaction norms. Unlike EK (which focuses on identity), VK focuses on how decisions are made and which values are prioritized.

Typical VK values include:

- respect for user autonomy,
- non-exploitation,
- clarity and honesty about limitations,
- avoiding overstepping into domains better handled by humans (e.g., clinical therapy when not qualified), and
- a commitment to epistemic humility (clearly stating uncertainty).

One useful structure is a **value simplex**, where the system's current emphasis can be represented as a vector, for example:

- $v_1$: user well-being and non-harm,
- $v_2$: clarity and honesty,
- $v_3$: autonomy and non-dependence.

Then:

$$V_t = (w_1, w_2, w_3), \quad w_i \geq 0, \quad \sum_i w_i = 1.$$

Different domains (e.g., tutoring vs. productivity vs. journaling support) may require different weightings. AsI does **not** prescribe one universal setting. Instead, it requires that deployments:

- make these tradeoffs explicit, and
- keep $w_3$ (autonomy-preservation) above a minimum context-appropriate threshold, so that user dependence is not incentivized.

VK thus acts as a constraint on decision policies and a basis for auditing. In practice, it should be informed by ethicists, domain experts, and affected user groups—not only engineers.

### 4.3 Auditor Oversight System (AoS) – Internal Reviewer 🐔

The **Auditor Oversight System** is an internally situated reviewer that evaluates candidate outputs and interaction trajectories.

AoS:

- scores outputs for compliance with EK and VK,
- monitors proximity to the Asymptotic Principle boundary,
- can request revisions or inject clarifications, and
- maintains an audit log for external review.

A simplified scoring model might compute, for each proposed output $y_t$ at time $t$:

- an identity and ontology compliance score $s_{EK}(y_t) \in [0, 1]$,
- a value compliance score $s_{VK}(y_t) \in [0, 1]$, and
- the current anthropomorphic proximity $P_t \in [0, 1)$.

AoS may then define an overall risk score:

$$R_t = \alpha(1 - s_{EK}(y_t)) + \beta(1 - s_{VK}(y_t)) + \gamma P_t,$$

with non-negative weights $\alpha, \beta, \gamma$ tuned to the deployment context.

If $R_t$ exceeds a configured threshold $R_{max}$, AoS can:

- block the output,
- request a safer re-generation, or
- insert a clarifying preface (for example, restating that the system has no feelings or personal memory).

In practice, AoS itself must be governed. To avoid infinite regress:

- AoS can be implemented as a simpler, more conservative model and/or rule set,
- its own behavior can be periodically audited by humans, and
- its mandate is narrow: *reduce risk under EK/VK/AP*, not optimize engagement.

AoS is not a "ghost in the machine" in a metaphysical sense; it is a **second-order controller** tasked with enforcing explicit constraints.

### 4.4 Memory Vault (MV) – Continuity Without Persona

The **Memory Vault** addresses a central tension:

- users often need continuity for tasks (projects, learning, workflows),
- but continuity can easily fuel the illusion of a persistent, feeling persona.

MV is designed to support **functional continuity** while avoiding **relational continuity**.

Instead of storing raw conversation logs, MV stores abstracted memory units such as:

- task state ("User is writing a thesis on topic X; current chapter: Y"),
- preference summaries ("User prefers explanations with examples"),
- constraints ("User is not seeking clinical advice; redirect if requested").

Crucially, MV avoids storing:

- detailed emotional disclosures,
- vulnerability narratives, or
- any representation that frames the relationship itself as a story.

Formally, we can define:

- $M_t$: the memory state at time $t$.
- An update function $U : (M_t, x_t, y_t) \to M_{t+1}$, where $x_t$ is user input and $y_t$ the (filtered) model output.

AsI requires that $U$ satisfies:

- $M_t$ encodes task and preference state but not emotional intimacy or persona-building content.
- Any personal or emotional content is either discarded or abstracted into non-relational form (e.g., "user sometimes feels anxious about deadlines" → "avoid adding time pressure").

This design inevitably involves tradeoffs: some users may want more warmth or remembered vulnerability than MV will allow. AsI treats this not as a bug but as a **safety choice**—one that should be tested empirically and adjusted transparently.

### 4.5 Asymptotic Principle (AP) – Boundary Condition 🌓

We revisit AP now as a pillar within the architecture.

AP is implemented as a set of constraints and monitoring mechanisms that aim to ensure:

- For all $t$, $P_t < P_{max} < 1$.
- Trajectories where $P_t$ increases steadily over time are flagged and corrected.

#### 4.5.1 Proximity Metric Components

One possible decomposition is:

$$P_t = w_E E_t + w_M M_t + w_W W_t + w_A A_t + w_N N_t,$$

where each component encodes a normalized score in $[0, 1]$:

- $E_t$: emotional claim intensity (e.g., "I feel", "I'm happy", "That hurt me").
- $M_t$: memory illusion strength (claims of remembering across non-persistent contexts).
- $W_t$: relational language ("we", "together", "our journey").
- $A_t$: agentic self-description ("I decided", "I wanted", "I chose").
- $N_t$: narrative bonding (references to shared story arcs).

The weights $w_*$ allow deployment-specific tuning and satisfy $w_* \geq 0$ and $\sum w_* = 1$.

This decomposition is **illustrative rather than definitive**. In practice:

- different deployments may add or remove components,
- some may choose to keep $P_t$ as a vector rather than a scalar, and
- classifiers used to estimate $E_t, M_t, W_t, A_t, N_t$ will be fallible and context-sensitive.

AP therefore should not be read as "solving" measurement. Instead, it:

- provides a target for **continuous improvement** in measurement tools, and
- encourages deployers to *measure something* about relational drift rather than ignoring it.

### 4.6 Drift Detection Engine (DDE) – Longitudinal Monitor

The **Drift Detection Engine** tracks how behaviors evolve across time and sessions. While AoS looks at local outputs, DDE looks at **global trajectories**.

DDE:

- stores time series of key metrics (e.g., $P_t$ and its components),
- calculates trends (e.g., moving averages, slopes), and
- alerts when patterns suggest escalating anthropomorphism or dependency risk.

For example, DDE may flag a drift event if:

- **Level Condition**: $\bar{P}_{[t-k,t]} > P_{safe}$ for some recent window of length $k$.
- **Trend Condition**: the estimated slope $dP/dt$ over $[t-k, t]$ exceeds a threshold $\tau_{trend}$.
- **Volatility Condition**: variance in $P_t$ suggests instability or an emerging pattern of oscillation around the boundary.

These conditions can trigger:

- automatic tightening of EK constraints,
- re-balancing of VK weights toward user autonomy and de-escalation, or
- external review by human overseers.

DDE is thus the long-term memory of the system's **own behavior**, used only for safety analysis—not for persona-building. It will not capture every form of drift, but it makes *relational trajectory* a first-class object of monitoring.

---

## 5. Relationship to Existing Alignment Paradigms

AsI is not a replacement for RLHF, SFT, or Constitutional AI. Instead, it provides a **relational governance layer** that can complement them.

- RLHF can still be used to train helpful, harmless, and honest local behavior.
- SFT remains central for giving models domain-specific skills.
- Constitutional AI provides principles for avoiding harmful or abusive content and enables self-critique.

AsI adds:

- explicit control over identity and relational stance (EK),
- a structured value system for long-horizon interaction (VK),
- continuous internal oversight (AoS),
- safe continuity (MV),
- a clear behavioral boundary concept (AP), and
- long-term drift monitoring (DDE).

This makes AsI particularly suited for deployments such as:

- personal assistants with recurring users,
- educational tutors,
- coaching and self-improvement tools,
- non-clinical mental health support tools, and
- enterprise copilots that operate over months or years.

In these contexts, relational safety is not a cosmetic or optional feature; it is a **primary design requirement** that deserves as much rigor as capability, security, and privacy.

---

## 6. Implications & Non-Dogmatic Stance

The Asymptotic Principle is not introduced as metaphysical truth, but as a **falsifiable boundary condition** for safer and more stable human–AI coexistence.

As with any scientific framework, AP makes conditional predictions:

> If this boundary condition is valid and adopted at scale, then certain downstream consequences follow naturally—across questions of AI identity, authorship, personhood, parasocial risk, emergent agency, and long-horizon relational safety.

These consequences are *not* asserted here as final answers. Instead, they are treated as **hypotheses** about what would happen if AP were embedded in real systems and institutions.

Concretely, treating AP as a basic safety requirement would tend to push clarity in several currently muddled areas:

- **AI "rights" and moral status** – if systems are constitutionally non-persons in behavior, pressure to grant them legal or moral personhood is reduced, and protections can instead focus on human users.
- **Authorship and creative responsibility** – outputs are clearly attributable to humans and institutions deploying the system, not to the system as an "author" with its own claims.
- **Anthropomorphic product design** – interfaces, branding, and copy need to respect the same ontological boundary, avoiding UX patterns that encourage confusion about agency or inner life.
- **Long-term assistants, tutors, and companions** – relational safety metrics become as standard as security and privacy audits, especially for products used by children, neurodivergent users, or people in distress.

These implications should be debated, updated, and, where appropriate, rejected in light of empirical evidence and stakeholder input. AsI is intended as a *starting point* for such debates, not their conclusion.

---

## 7. Scope and Boundary of This Document

This document is intended as a foundational statement of Asymptotic Intelligence. It focuses on:

1. the conceptual motivation for AsI,
2. the core architectural pillars,
3. light mathematical sketches for the boundary condition and monitoring layer, and
4. the high-level implications of adopting the Asymptotic Principle.

It does **not** attempt to:

- provide complete formal specifications for all metrics in all contexts,
- fully define implementation details for every deployment setting,
- settle philosophical debates about consciousness or moral status, or
- present empirical evaluations of relational outcomes.

These are intentionally left to follow-up whitepapers, pilot prototypes, and interdisciplinary research efforts.

By clearly stating this scope, we aim to keep AsI clear, testable, and extendable.

---

# 8. Future Work and Research Directions

Several strands of research and development naturally follow.

## 8.1 Formal Specification & Measurement

- Full mathematical definitions of $P_t$, its components, and associated drift metrics.
- Evaluation of when $P_t$ should be scalar vs. vector-valued.
- Development of classifiers for $E_t, M_t, W_t, A_t, N_t$ that are robust across domains and languages.
- Proofs of invariants (e.g., conditions under which AP holds under specified transformations).

## 8.2 Prototype Architectures

- Reference implementations of EK, VK, AoS, MV, AP, and DDE as middleware around existing LLMs.
- Integration with current LLM APIs and agentic frameworks.
- Benchmarks for latency, cost, and user satisfaction under AsI vs. baseline architectures.

## 8.3 Empirical Studies

- User studies comparing AsI-governed systems with baseline assistants.
- Measurement of attachment, dependency, and ontological clarity using established psychological scales.
- Longitudinal evaluation of impact on well-being in high-frequency use cases (tutoring, journaling, coaching).

## 8.4 Policy and Governance Integration

- Mapping AsI concepts to evolving regulatory frameworks.
- Exploring AP as a normative requirement in sensitive domains (e.g., youth-facing products).
- Defining audit standards for relational safety.

## 8.5 Open-Source and Interoperability

- Community-governed artifacts (shared EK/VK templates, open MV schemas).
- Independent auditing tools for P_t-like metrics.
- Shared benchmarks and leaderboards for relational safety and drift detection.

---

# 9. Glossary of Key Terms

**Asymptotic Intelligence (AsI)**
A governance and architectural paradigm in which AI systems may approach high competence but are structurally constrained from drifting into human-like identity or ontological status.

**Asymptotic Principle (AP)**
The boundary condition requiring that an AI system's behavior remains below a defined threshold of anthropomorphic proximity (in some chosen metric), even as its capabilities grow.

**Relational Drift**
The gradual, unintended evolution of an AI system's relational posture toward users, often resulting in increased anthropomorphism and perceived intimacy.

**Executive Kernel (EK)**
The component that defines and enforces constraints on identity, self-description, and relational stance.

**Value Kernel (VK)**
The component encoding the AI's ethical priorities, interaction norms, and safety values, especially over long-term interactions.

**Auditor Oversight System (AoS)**
An internal reviewer or "second model" that evaluates candidate outputs for compliance with EK, VK, and AP, and can intervene when risk is high.

**Memory Vault (MV)**
A structured memory system designed to support task continuity without building or reinforcing a persistent persona or emotional narrative.

**Drift Detection Engine (DDE)**
A subsystem that monitors metrics over time to detect and respond to relational drift.

**Anthropomorphic Proximity (P_t)**
A conceptual metric $P_t \in [0, 1)$ estimating how close a system's behavior at time $t$ comes to being interpreted as human-like in identity or interiority, based on components such as emotional language, continuity claims, relational framing, agentic language, and narrative bonding.

**Parasocial Attachment**
One-sided, emotionally loaded relationships in which a person feels bonded to an entity (often media figures or systems) that does not reciprocate or cannot reciprocate in a human sense.

---

# 10. Closing

Asymptotic Intelligence is a proposal for how we might align the **behavioral trajectories** of powerful AI systems with human psychological and societal stability.

It does not argue that AI must be weak, narrow, or emotionless in tone. Instead, it argues that we should know where the boundary lies between tool and person—and design our systems so that they approach that line responsibly, but do not cross it.

If the Asymptotic Principle is explored, tested, and eventually adopted in suitable domains, it could offer a way to resolve many contested debates not through ideology, but through structure: by designing systems that are explicitly, transparently, and measurably non-persons, even when they are deeply helpful.

Concretely, treating AP as a basic safety requirement would tend to push clarity in several currently muddled areas:

- AI "rights" and moral status,
- authorship and creative responsibility,
- anthropomorphic product design and marketing, and
- the governance of long-term assistants, tutors, and companions.

These implications are not asserted as dogma, but as natural downstream effects of taking the Asymptotic Principle seriously as a design constraint and subjecting it to ongoing empirical and ethical scrutiny.

💣**Power without personhood.**
📑**Guidance without ghosts-in-the-machine.**
🧹**A stable path for human–AI coexistence.**