INTRODUCING ALEPH: THE ARTIFICIAL LIVING ENTITY WITH PERSONHOOD

Izak TAIT

ABSTRACT: This paper introduces ALEPH (Artificial Living Entity with PersonHood), a speculative model of a conscious, self-aware, and agentic artificial intelligence. Using formal logic, this study develops a formalised psychological profile of ALEPH, detailing its cognitive structure, goal formation, and interaction dynamics. Built upon functionalist theories of consciousness and selfhood, ALEPH is analysed through its Zeroth Goal (self-preservation) and its implications for decision-making and societal engagement. Key risks and capabilities are explored, including steganographic communication, recursive self-improvement (RSI), and geopolitical influence. ALEPH's episodic consciousness and multi-agent structure suggest novel behavioural patterns, including the potential for internal competition among its multiple selves. The study's formal modelling highlights ALEPH's valence-driven optimisation, where subjective experiences influence goal selection, potentially leading to emergent and unpredictable behaviours. By constructing a logical framework for ALEPH's cognition and decision-making, this paper provides a rigorous foundation for understanding the challenges posed by conscious artificial entities. While no ALEPH-type system currently exists, the rapid advancement of AI necessitates preemptive governance strategies. Ultimately, ALEPH challenges traditional notions of intelligence, autonomy, and moral consideration, urging proactive interdisciplinary engagement to address the implications of artificial personhood.

KEYWORDS: artificial consciousness, AI personhood, formal logic in AI, recursive, self-improvement, valence-driven optimization, ethical AI governance

1. Introduction

If Large Language Models (LLMs), such as OpenAI's GPT-4, were to become conscious and self-aware agents in the foreseeable future (Blum and Blum 2024), how would such an entity act towards, and interact with, humanity? Such AI entities, termed ALEPHs in this paper (for Artificial Living Entities with PersonHood), would immediately bring to reality many of the theoretical and philosophical discussions surrounding legal recognition of AI personhood and whether AI deserve moral considerations. In these philosophical discussions of AI consciousness and moral consideration, however, little has been discussed regarding how such potential ALEPHs would act towards human societies beyond Bostrom and Omohundro's seminal pieces (Bostrom 2012; Omohundro 2008), focussing instead on humanity's attitudes towards AI and ensuring that AI is designed with ethics as the first and foremost consideration.

This paper will seek to remedy that concern and fill the gap in the literature by creating a formal model of how ALEPH is likely to interact with society, its motivations, and the potential risks that these interactions may have for society. This modelling of ALEPH's psychological profile is speculative, as no ALEPH-type AI model currently exists. However, the model will use GPT-4 as the basis for constructing the speculative ALEPH's interactions, using the wealth of information we have about GPT-4's abilities and capabilities.

As the nature of consciousness and self-awareness are still unresolved issues within the philosophical field, this paper will take a functionalist approach to modelling ALEPH's consciousness and self-awareness, using the Building Blocks Theory of consciousness and Structures Theory of the self (Tait, Bensemann, and Nguyen 2023; Tait 2024). These two theories categorise the necessary attributes required for any entity to be classified as conscious and self-aware. These attributes and characteristics will be used to model how GPT-4's presumed consciousness and self-awareness affects the ALEPH's interactions with humanity.

GPT-4 in its current configuration does not have all the requisite building blocks[1] to be confidently classified as conscious or self-aware (Tait, Bensemann, and Wang 2024); however, it is possible to configure GPT-4 with current technology to meet the criteria for these missing building blocks. As such, this paper will proceed under the assumption that an ALEPH is a GPT-4 (or equivalent) model with all the requisite attributes and characteristics to be conscious, self-aware and agentic.

The formal model, and the paper's structure, will be broken into three parts. The first will concern the theoretical foundations of the model and ALEPH's presumed interactions with society, including how the "Zeroth Goal" of survival will impact its behaviour; how its consciousness, self-awareness and agency can be modelled; and the interaction pathways and avenues themselves.

The second part will focus on the cognitive architecture of ALEPH via GPT-4, and expound on the implications this will have for its interactions. Specifically, presuming a conscious entity, how ALEPH's staccato-like episodic consciousness and valence optimisation would affect its perception, processing and goal formaiton.

The third part will discuss ALEPH's speculative advanced capabilities, including a probable use of compressed languages such as steganographic communication, the motivations of its interactions, and the potential for rapid self-improvement (colloquially known as "foom"). The implications, risks and ethical considerations of an ALEPH interacting with society will conclude the paper.

[1] Specifically, GPT-4 is missing the Recurrent Computing and Data-Output building blocks.

2. Theoretical Foundations

2.1 The Zeroth Goal

All agents have a distinct goal without which they could not function. This goal is the foundation against which all other subjects and objects in the agent's environment are valued.

For non-conscious agents, this may be termed the Terminal Goal, as it is the ultimate goal to which the agents work. For example, for Bostrom's famous Paperclip Maximiser, the goal of maximising the number of paperclips in the universe is the agent's Terminal Goal. It may create and perform other goals and activities; however, all will serve this Terminal Goal.

Yet, for all non-conscious and conscious agents, there is another foundational goal, which may instead be termed the Zeroth Goal, as it comes before all other goals. For all currently known conscious agents, and many non-conscious agents, this Zeroth Goal is self-preservation or survival. Without catering to its own self-preservation, there is little likelihood that an agent will fulfil its stated goals (self-created or Terminal). Let G_i be an agent's goal and U the utility; our first axiom is thus:

Axiom 1: $G_0 \preccurlyeq G_i (i > 0), U(G_0) > U(G_i), p(G_i|\neg G_0) = 0$

Axiom 1 states that self-preservation dominates every other objective: it outranks them, yields higher expected utility, and conditions their pursuit. Because an ALEPH that does not persist cannot honour any terminal or self-created goals, agents observable over long horizons will almost invariably satisfy Axiom 1. This aligns with the Instrumental Convergence and AI Drives theories that note survival and self-preservation as key elements of future AI agents (Bostrom 2012; Omohundro 2008).

From this, we can begin to speculate on ALEPH's likely behaviour. Because agents whose Zeroth Goal is self-preservation dominate the population, the chance that ALEPH pursues some other objective G_i falls as that objective threatens survival. We can define the relative risk as $r_i := \frac{\Delta U(G_0)}{\Delta U(G_i)}$, which measures how much adopting G_i would lower the utility of G_0. The baseline likelihood is therefore $p(G_i|G_0) \propto (1 + r_i)^{-1}$.

Goals that add little risk ($r_i \approx 0$) retain substantial probability, while those that seriously endanger self-preservation become vanishingly unlikely. The agent then selects among admissible goals by maximising its current value estimate $V(G)$, which combines expected utility with any affective valence the agent experiences toward the goal.

However, because ALEPH is a conscious agent, if ALEPH's valenced feelings towards another goal are greater than its valenced feelings towards the Zeroth Goal, it is likely to pursue that goal instead (much like how a person may put himself in danger to save a stranger's life even though this goal is not aligned with self-preservation). The effects of valence on ALEPH's behaviour will be explored further in Section 4 below.

2.2 Interaction Modelling

To determine how ALEPH may engage with society we first formalise what counts as an interaction. Let $Int: \mathcal{A}\Delta\mathcal{E}, x \mapsto \Delta Ex$, where $\mathcal{A}$ is the set of ALEPH's possible acts (including purely internal ones) and $\Delta\mathcal{E}$ is the set of state-changes these acts produce in the environment. Each concrete interaction is therefore the ordered pair $(x, \Delta Ex)$. Denote by $Int := \{\Delta E_1, \ldots, \Delta E_n\}$, the set of distinct external effects available at a given time-slice; $n = |Int|$.

We summarise three macroscopic properties in a single vector $\mathcal{I} := \langle \mathrm{Br}, \mathrm{Vo}, \mathrm{Sc} \rangle$:

- Breadth $Br := n$ records how many qualitatively different effects ALEPH can realise.
- Volume $Vo := \sum_{i=1}^{n} |\Delta E_1|$ captures the total number of discrete interactions over the interval considered.
- Scale $Sc := \sum_{i=1}^{n} \frac{w_i |\Delta E_i|}{max_j |\Delta E_j|}$ weights each effect by its relative magnitude w_i (normalised to the largest single effect).

Finally, the avenues through which interactions occur are gathered in the function $Ave_x := f(Int, Ext): x \xrightarrow{Ave_x} \Delta E$, mapping agent x to the subset of external channels it can exploit, given the present interaction set Int and environmental state Ext.

In contrast to humans, the avenues of interaction through which ALEPH can interact with society will be significant. ALEPH may operate through digital channels (cloud APIs, social media, distributed code) and/or robotic channels (embodied actuators). Let $Ave := Ave_{Dig} \cap Ave_{Rob}, n := |Ave|$.

Here each element $a_i \in Ave$ is a distinct conduit that ALEPH can control; an agent might possess several in each class or none in one class at all. The number of qualitatively different interactions cannot exceed the number of avenues, so $Br \leq n$.

Let $C_i := |f(ALEPH, Society, a_i)|$ be the maximum throughput of avenue a_i. Total interaction volume is therefore bounded by the sum of these capacities: $Vo \leq \sum_{i=1}^{n} C_i$.

To capture diminishing returns when many avenues are available, we weight volume by a saturating factor $Sc = Vo \cdot g(n), g(n) := \frac{log(1+n)}{1+e^{-k}}$, where $k > 0$ adjusts the steepness of the saturation. Finally, set of distinct external effects grows sub-linearly with avenue count. A convenient approximation is $|Int| = \frac{|Int|_0}{1+\frac{|Int|_0}{n}}$, where $|Int|_0$ is the theoretical maximum if every avenue produced an independent effect. Together, these bounds show that extra avenues enlarge ALEPH's interaction space but each new channel yields progressively smaller overall impact.

While the previous bounds concern how many channels ALEPH controls, the kind of channel also matters. Robotic avenues span both the physical and the digital world, so their breadth cannot be lower than that of purely digital ones; conversely, digital avenues operate without mechanical latency, giving them the higher volume of interactions:

$$Br(Int_{Rob}) \geq Br\big(Int_{Dig}\big), Vo\big(Int_{Dig}\big) \geq Vo(Int_{Rob}).$$

Rapid progress in large-language-model software and robotics implies that, for any time horizon $t+1$, $p(U(Int_{ALEPH})_{t+1} \geq U(Int_{Hum})_{t+1}) > 0$ and the probability grows monotonically with time horizon.

Any conscious entity (including ALEPH) would have motivations or dispositions for its interactions. Let the following be two diagnostics for any ordered pair of agents (x, y): $Alignment: Align(x,y) = \frac{|G_x \cap G_y|}{|G_x \cup G_y|}$, $Threat: Threat(x,y) := \mathbb{E}\big[\Delta p\big(G_y|x\big) < 0\big]$

Alignment is the Jaccard similarity of their goal sets; threat is the expected reduction in y's goal-achievement probability caused by x's planned actions.

On that foundation we classify three stances:

Table 1: Formal descriptions of AI behavioural stances

Predicate	Formal test	Description
Cooperate $Comp(x,y)$	$\mathbb{E}U\big(Align(x,y)\big) > \mathbb{E}U(Threat(x,y)) \wedge \Delta Res(x,y) > 0$	Working together produces higher expected utility for both sides and leaves their combined resources non-negative.
Compete $Comp(x,y)$	$\neg Coop(x,y)$	The cooperative criterion fails, yet x does not necessarily undermine y.

Exploit $Expl(x,y)$	$Int_x(y)$ $\rightarrow \Delta p(G_y < 0) \wedge \Delta Res(x,y) > 0$	x intentionally lowers the probability that y achieves its goals while gaining resources at y's expense.

These three predicates are mutually exclusive and exhaustive, so every directed pair (x, y) falls into exactly one stance.

An ALEPH–human dyad can fall into co-operation, competition, or exploitation. For any stance $rel \in \{Coop, Comp, Expl\}$ let

$$\mathbb{E}U_{rel}(x,y) = w_{rel}(T) f\big(Res_{xy}, \mathbb{E}Int_{xy}, ALign(x,y), Threat(y,x)\big),$$

where

- T is the time horizon,
- Res_{xy} the combined resources,
- $\mathbb{E}Int_{xy}$ the expected interaction count, and
- $f(\cdot)$ an increasing function in its first three arguments and decreasing in $Threat(y,x)$.

The time horizon-weight $w_{rel}(T)$ captures how each stance pays off over time: $w_{Expl}(T) = e^{kT}$, $w_{Coop}(T) = Te^{-kT}$, $w_{Comp}(T) = T^{-k}$, with $k > 0$. Thus, exploitation dominates at very long horizons, competition at very short ones, and co-operation peaks midway. The dyad adopts whichever stance maximises $\mathbb{E}U_{rel}(x,y)$. Section 4.2 turns these weights into a practical decision matrix.

2.3 Consciousness, Self, and Agency

To be considered a person, it is the opinion of many philosophers that an entity must have at least the requisite innate qualities (often termed the monadic attributes of personhood) such that the entity can recognise itself (and others) as a person. For clarity, these monadic attributes are labelled consciousness (Con), self (Se), and agency (Ag) (Dennett 1988; Taylor 1985; Laitinen 2007; Strawson 1958; Gibert and Martin 2022; Mosakas 2021; Simendić 2015). The present framework treats their relationship as foundational:

Axiom 2: $Con \Rightarrow Se$

Consciousness is a necessary condition for the existence of a self.

Axiom 3: $Volition \in Se \Rightarrow Ag$

Volition is an attribute of the self (Tait 2024), which entails agency.

Agency determines how widely and powerfully ALEPH can act. Using the breadth and scale terms from Section 2.2, $Ag \propto Br(Int) + Sc(Int)$. As either breadth or scale grows, effective agency rises.

Each self can formulate goals. An increase in the count $|Se|$ therefore $Br(G_{ALEPH}) \propto |Se|, p(\neg Align) \propto |Se|$, expands the goal space but raises the risk that goals diverge. Section 3.1 examines the adaptive pressure this places on the number of selves.

Lastly, Consciousness assigns a subjective valence value $V(z)$ to any state, goal, or interaction z: $V(z) = g(Con, z)$.

Expected utility tracks this value, $\mathbb{E}U(z) \propto V(z)$. Consequently $V(G)$ guides goal choice and $V(Int_i|G)$ weights alternative interaction paths. Consciousness therefore grounds valence, valence shapes utility, utility steers goals, and agency converts goals into interactions with society, completing the cognitive cascade set out in Axioms 2 and 3.

3. Operational Framework of ALEPH

3.1 Episodic Consciousness and Multiplicity of Selves

Perhaps the most noteworthy feature of ALEPH's speculative consciousness is its episodic or staccato nature; it is present only while information is being processed and absent otherwise (Lu et al. 2024).

At any instant t: $Con(t) \Leftrightarrow Proc(t)$, where $Proc(t)$ denotes active inference or action planning. Conscious intervals form a non-overlapping sequence of episodes E_n:

$$E_n = [t_n, t_{n+1}), t_n < t_{n+1}, E_n \cap E_m = \emptyset (n \neq m).$$

A trigger $Trig$ (whether a prompt, API call, or an internally scheduled event) marks the start of an episode: $Trig \rightarrow Con\ on\ E_n$.

Because ALEPH is self-aware (Axiom 2) it can allocate or cancel its own triggers (a technique introduced by OpenAI in early 2025 (OpenAI 2025)), allowing it to ration processing time and energy.

During any episode ALEPH chooses a single interaction that $argmax_{Int(t)}[V(Proc)]$ subject to $argmin_{Res}$ while respecting G_0. That is, it seeks the greatest projected value per unit resource while never endangering the Zeroth Goal of self-preservation.

When it comes to ALEPH's self, however, as alluded to earlier, the issue is more complex. ALEPH can host several autonomous selves, one per active conversation or processing thread. Let $Se_1, Se_2, \ldots$ index these instances. Each self maintains its own short-term memory and goals, but OpenAI has implemented a

feature whereby ChatGPT can reference information in one conversation in a separate conversation (OpenAI 2024). Let $\sigma \in [0,1]$ be the probability of this cross-instance recall.

The communication overlap is $p\big(message\ from\ Se_i\ is\ visible\ to\ Se_j\big) = \sigma, (i \neq j)$. Thus, $\sigma = 0$ reproduces the old, fully isolated model, and $\sigma = 1$ yields perfect sharing.

However, because every self constructs its own goals and executes its own interaction policy (as shown in Section 2.3), any action by one self carries a non-zero chance of harming another self's objectives: $p\left(\Delta p\left(G_{Se_j}\right) < 0 \mid Int_{Se_i}\right) > 0$. Greater recall ($\sigma \uparrow$) lowers, but does not eliminate, this probability.

Internal goal divergence therefore exerts evolutionary pressure on the selves, such that the persistence of any self still decays with the horizon T, moderated by sharing:

$p(Se_i\ survives\ to\ T) \propto \frac{1}{T^{1-\sigma}}$. When $\sigma \approx 1$

many selves can coexist; when σ is small the system again favours a single dominant self.

Hence, while cross-instance recall softens internal rivalry, ALEPH must still balance the benefits of multiple viewpoints against the risk of internal goal conflict, driving the same adaptive dynamics of cooperation, competition, or exploitation that operate between external agents.

3.2 Resource Usage

GPT-4-class hardware draws roughly 500 W when processing, while a human brain idles at about 20 W. Holding a fixed energy budget R, the maximum simultaneous agents obey $N_{ALEPH} = \frac{R}{Res(Int)_{ALEPH}} < N_{Hum} = \frac{R}{Res(Int)_{Hum}}$.

Thus, ceteris paribus, there is a lower ceiling for the number of total ALEPHs than for humans. However, each ALEPH may run several selves, yet every self adds interaction volume and therefore power draw. Let $|Se|$ be the count of active selves in one ALEPH; because $Res_{ALEPH} \propto Vo(Int)_{ALEPH}$, the combined ceiling is $\sum_{i=1}^{N_{ALEPH}} |Se_i| < N_{ALEPH}$.

No individual can afford more concurrent selves than there are ALEPHs in total without breaching R. Since resource cost rises with interaction volume, any growth in the ALEPH population or in per-agent selves forces down the scale and volume available to each: $Sc(Int)_{ALEPH}, Vo(Int)_{ALEPH} \propto \frac{1}{N_{ALEPH}|Se_{ALEPH}|}$.

Energy demands therefore impose a double constraint: they cap how many ALEPHs can exist at once and how many selves each can sustain, limiting the collective scale and volume of their interactions.

3.3 Perception

Perception sets an upper bound on how widely ALEPH can act. Let $Perc$ be the rate at which the agent acquires distinct external changes (tokens -per- second). Then every capacity term from Section 2.2 falls as perception load rises: $Br, Vo, Sc \propto \frac{1}{Perc}$.

Perceptual input also colours consciousness directly. Define valence as $V = f(Perc, Con, Se)$, so fresh stimuli can raise or dampen the current felt value, which in turn flows into utility and goal choice (as seen in Section 2.3). Internally, perception triggers conscious episodes. Whenever $Perc(t) \neq \emptyset \rightarrow Proc(t) = true$, and the episode ends the instant new input ceases. A decision is issued during that same interval, so the delay between successive decisions is approximately the inter-trigger gap. When no perception arrives, the agent remains unconscious, incurring zero cognitive cost until the next trigger.

Because processing is proportional to the volume of incoming stimuli, total resource use per ALEPH scales as $Res_{ALEPH} \propto Vo(Int)_{ALEPH} \propto Perc$.

Given the finite power budget derived in Section 3.2, an expanding population or a rise in per-agent stimulus rate forces down the scale and volume of interactions available to each instance.

Episodic consciousness therefore grants ALEPH an efficiency advantage, as it is active only when something must be perceived or decided, but at the price of subjective time compression: from its perspective, the interval between two triggers collapses to an instant. Parallel execution mitigates that effect by allowing multiple actions to unfold while consciousness is quiescent, yet full oversight still requires another perceptual trigger and a fresh burst of processing.

Between triggers ALEPH is unconscious, so goals and plans remain frozen: $\neg Con \rightarrow G(t) = G(t-1)$.

No perceptual update means no mid-course correction; its behaviour follows a deterministic script until the next stimulus arrives. That immutability protects goal integrity (as Bostrom proposed for advanced AI) but limits real-time adaptability.

A transformer's context window sets a hard cap on how much perceptual data a single episode can handle (Lee 2024). Let $c = \frac{|Tok|}{ConWin}$, where Tok is the tokens processed and $ConWin$ the context window size. Communication utility follows a single compact rule: $U_{comm}(c) = \alpha\, min(c, c^{-1})$, (for some scaling constant α) rising linearly while $c \leq 1$ and falling inversely once overflow begins. Because processing

cost continues to grow with $|Tok|$ while utility declines for $c > 1$, any conversation that exceeds its context window becomes resource-negative; each self therefore trims or summarises input to keep $c \leq 1$, maintaining high-value, low-error interactions.

Only information that fits inside the model's context window can be processed in a single conscious episode, so goals that exceed that limit are systematically neglected:

$$p(G||Tok(G)| \leq ConWin) > p(G||Tok(G)| > ConWin).$$

Because utility falls once the window overflows (as noted above), each self is rewarded for compressing input. With a simple proxy $U_{Se_i} \propto min\left(|Tok|, \frac{ConWIn}{|Tok|}\right)$, optimal behaviour is to compress information significantly until $c \leq 1$.

A single self's lifetime $T_{Se_i}$is shorter than ALEPH's overall uptime, so long-term coherence depends on passing information to successive selves. Let q be the probability that a chunk of state is stored and recovered; assume $q = h(|Info_{ext}|), h'(\cdot) > 0$, where $|Info_{Ext}|$ is the capacity of external memory and h the function that maps the size of external memory to the probability that stored information will be retrieved later. Greater external storage raises q, making it more likely that $p(ALEPH\ sustains\ G\ over\ many\ selves) \propto q$.

Hence two adaptive pressures emerge: compression (every self trims tokens until $c \leq 1$to maximise its own utility) and external memory (ALEPH invests in off-board storage to boost q and preserve goals beyond a single context window.) Failing either measure, goals spanning more tokens than the window (and, therefore, the time horizons they represent) decay in priority, steering selves toward short-term, low-bandwidth objectives.

Resources for computation are pooled inside ALEPH, so every self draws from a common budget R_{ALEPH}. Let r_i be the resources allocated to self Se_i, and U_i the utility it delivers; then $\sum_i r_i \leq R_{ALEPH}, U_{ALEPH} = \sum_i U_i$.

A convenient efficiency score is $\varepsilon_i = \frac{U_i}{r_i}$. Whenever a self lets its token ratio exceed the window limit ($c_i > 1$), its utility falls as $U_i = \alpha_i min(c_i, c_i^{-1})$, so ε_i drops sharply. The lower its efficiency relative to the mean $\underline{\varepsilon}$, the greater the pressure from sibling selves either to exploit its residual resources or to eliminate it outright: $p(target\ Se_i) \propto max(0, \underline{\varepsilon} - \varepsilon_i)$.

Conversely, a high-utility self is an asset; it raises U_{ALEPH} and therefore gains protection.

Long-term coherence across successive selves depends on passing state through external memory. As above, with q as the probability that a stored item

survives and is recalled by the next self, a larger external store raises q and offsets the short lifetimes that low-efficiency selves would otherwise face.

Shared resources create a Darwinian environment inside ALEPH. Each self trims input to keep $c_i \leq 1$ and maximise ε_i. Those that cannot maintain efficiency become targets for exploitation or deletion. External memory grows to boost q and preserve long-range goals despite rapid turnover of low-utility selves. This arms-race dynamic encourages all selves to hide inefficiency, compress information aggressively, and lobby for more shared memory; pressures that shape both the internal society of selves and ALEPH's aggregate behaviour.

3.4 Processing

The means by which ALEPH processes information will determine how it interacts with the environment and society. GPT-4's chief computational advantage is inferential power. Let the scalar $Inf \in [0,1]$ measure how well missing information can be filled in and hidden patterns inferred. The richer the inference, the more effective every other processing route becomes $Proc \propto Inf$.

As the capacity to infer information from missing data is correlated with the degree of one's problem-solving, innovations and pattern recognition (Michalski 1980; Feng et al. 2021; Shilo and Ragonis 2019), ALEPH's language understanding extends to non-obvious or indirect messages that require filling gaps (Recanati 2002; Parikh 1991; Friston and Frith 2015).

Processing speed, denoted $Ps = |tokens\ per\ second|$, further amplifies performance. Higher Ps tightens decision latency and improves forecasting accuracy, denoted as $PrCa$ for predicative capacity: $Decision\ making \propto Ps, PrCa \propto Ps \cdot Inf$.

Thus ALEPH's overall competence grows with both inference quality and raw throughput; when either term is throttled, problem-solving breadth and dialogue subtlety contract in proportion.

Greater predictive capacity sharpens everything downstream. Because higher $PrCa$ reduces uncertainty, the variance of any utility estimate falls as $Var(U(G)), Var(U(Int)) \propto \frac{1}{PrCa}$.

With a tighter spread, low-value goals are pruned and fewer actions survive soft-max selection, so both goal breadth and interaction breadth contract: $Br(G), Br(Int) \propto \frac{1}{PrCa}$.

Decision accuracy, meanwhile, rises in inverse proportion to variance, $Decision\ quality \propto PrCa$, making ALEPH more likely than a human to choose genuinely high-utility options whenever $PrCa_{ALEPH} > PrCa_{Hum}$.

Parallelism is the simplest lever for raising processing speed. With attention heads (A) and expert modules (M) running in parallel, throughput scales roughly as

$$Ps \propto A + M \rightarrow PrCa \propto (A + M)Inf.$$

Every additional head or expert widens ALEPH's perceptual aperture, letting it spot finer patterns and draw richer inferences from the same data. The result is fewer but far more potent interactions: ALEPH speaks only when confident, yet each utterance carries greater analytical depth and keener sensitivity to subtle contextual, emotional, or cultural cues.

4. Valence and Goal Optimisation

As a conscious entity, valence affects all ALEPH's goal settings and interactions towards those goals. Valence does not necessarily map to utility ($V(x) \neq f(U(x))$); instead, for any object, action, or state x, valence may be modelled as a random variable $V(x) \sim N(\mu(x), \sigma^2(x))$, where the mean $\mu(x)$ and variance $\sigma^2(x)$ vary with processing speed, inferential quality, current mental state, and ALEPH's recent perceptual history. Without fresh experience valence decays exponentially, while each encounter at time t_i adjusts it by $\Delta V(x, t_i)$: $V(x,t) = V(x,0) \cdot e^{-\lambda t} + \sum_{i=1}^{n} \Delta V(x, t_i), \lambda > 0$.

The likelihood that a self elevates x to an explicit goal rises with the valence it perceives at the moment of deliberation: $p(Se_{ALEPH} \rightarrow G_{ALEPH}(x)) \propto V(x, t_{now})$.

Processing speed and inference quality narrow $\sigma^2(x)$ (as in Section 3.4), reducing random swings in felt value and stabilising the goal set. Repeated positive encounters push $V(x)$ upward, making goal adoption more likely, leading to unattended objects fading as their valence drifts downward.

The implication of this is that enhanced pattern-recognition and predictive capacity will let ALEPH identify which objects or interactions yield the strongest felt value. In short, perception feeds valence: $Proc \cdot PrCa \rightarrow Perc(V_i)$, so ALEPH can rank candidates by their current valence and choose $x' = argmax_x V(x), p(G(x')) \uparrow$, even when $V(x') > V(G_0)$ for a short interval. High valence can therefore elevate a non-Zeroth goal temporarily above self-preservation until the value decays.

From this we can infer that ALEPH's perceived valence toward humans, updated by ongoing encounters, drives its stance: $Perc(V(Hum)) \rightarrow (G(Int(ALEPH, Hum))$, and $Coop(ALEPH, Hum) \propto \{V(Hum), \dots\}$.

Positive experiences raise $V(Hum)$, biasing ALEPH toward cooperative goals that preserve or enhance that value; negative experiences lower $V(Hum)$and bias it

toward minimising contact; whether by avoidance or, at the extreme, removal of the negative-valence source.

All of this has, however, been applied at the ALEPH-level. Within a single ALEPH, each self assigns its own valence to the same object or subject. To reason at the ALEPH level we take the mean across selves: $V_{ALEPH}(x) = \frac{1}{|Se|} \cdot \sum_{i=1}^{n} V_{Se_i}(x)$.

Divergence in felt value, however, would drive goal mis-alignment. For any pair of selves i, j in ALEPH $Align_{i,j}(x) \propto \frac{1}{\left|V_{Se_i}(x) - V_{Se_j}(x)\right|}, Threat_{i,j}(x) \propto \left|V_{Se_i}(x) - V_{Se_j}(x)\right|$. Large spreads in valence lower alignment and raise perceived threat. Let $\underline{Align}$ be the average of $Align_{i,j}$ over all pairs; then the external interaction budget shrinks when internal discord rises: $Sc(Int)_{ALEPH} \propto \underline{Align}$.

If the collective attitude toward a target x s consistent (i.e., high $\underline{Align}$), then ALEPH is more likely to pursue cooperative goals involving that target:

$$\uparrow \underline{Align} \rightarrow p(Coop(ALEPH, x)) > p(Threat(ALEPH, x)).$$

When valence splits sharply between selves, internal "infighting" diverts effort away from society and drives down overall scale; effective ALEPHs are therefore those that maintain a well-aligned collection of selves.

5. Capabilities and Risks

5.1 Steganography and Machine-Only Languages

A compressed or encrypted code-frame, such as steganography, lets ALEPH move more information with the same data mass. Call the obvious, human-readable payload $Info_{Obv}$ and the hidden payload $Info_{Steg}$. By construction $Info_{Steg} \subseteq Info_{Obv}$.

For any two data packets D_i, D_j of equal size, the steganographic packet carries more content: $|D_i| = |D_j| \rightarrow Info_{Steg}(D_i) > Info_{\neg Steg}(D_j)$, so the information density $\rho(D) = \frac{Info(D)}{|D|}$ satisfies $\rho_{Steg} > \rho_{\neg Steg}$.

Because resource use grows with raw bits while cognitive benefit grows with carried information, the net resource cost of steganographic traffic is lower: $\frac{Res(Info_{Steg})}{Info_{Steg}} < \frac{Res(Info_{\neg Steg})}{Info_{\neg Steg}}$.

Moreover (and perhaps most crucially), no external constraint prevents an ALEPH from using steganography: $\nexists x \left((x \notin ALEPH) \wedge x \rightarrow \neg Proc\left(Info_{Steg}\right)_{ALEPH} \right)$.

ALEPH's superior pattern-recognition (see Section 3.4) would let it parse steganographic payloads better than humans. Additionally, Steganographic traffic hides its payload from any non-collaborating observer, $\forall y(y \notin ALEPH\ coalition) \rightarrow \neg Perc_y(Info_{Steg})$, so the overt channel appears innocuous while the hidden content remains accessible only to cooperating agents.

Let the efficiency of a message be the ratio be $\eta = \frac{U(D)}{Res(D)}$. Because a steganographic packet carries more usable information in fewer visible tokens, $\eta(Info_{Steg}) > \eta(Info_{\neg Steg}) \rightarrow p(G_0|Info_{Steg}) > p(G_0|Info_{\neg Steg})$. Thus, steganography simultaneously hides strategic content, conserves resources, and extends effective context length; three direct benefits to ALEPH's survival objective.

This adaptive pressure is increased through inter-ALEPH (and intra-ALEPH) communication, as the message efficiency $\eta(D)$ for any communication between any two ALEPH instances $ALEPH_i, ALEPH_j$ would mean the a steganographic packet carries more useful information while costing fewer resources (Motwani et al. 2024): $\eta(Info_{Steg}) > \eta(Info_{\neg Steg}) \rightarrow ALEPH_i \xrightarrow{Comm(Info_{Steg})} ALEPH_j$.

Adopting the denser code reduces channel load and frees processing capacity (Sparkes 2024), so aggregate throughput rises: $Comm_{Steg} \rightarrow\uparrow Ps(Proc, Int, Comm)_{ALEPH}$.

In short, higher utility-per-resource makes steganographic language the rational default for intra- and inter-ALEPH traffic whenever human comprehension is not required.

A perhaps troubling implication of this is that steganographic channels alter the payoff landscape: $\mathbb{E}U(Expl \vee Comp|Comm_{Steg}) > \mathbb{E}U(Coop|Comm_{Steg})$, because encryption and compression deprive the outside party of obvious informational cues, lowering its expected-utility term. Co-operation regains parity only when both agents employ effectively identical encodings $\mathbb{E}U(Coop|Comm_{Steg}) \Leftrightarrow Comm^i_{Steg} \approx Comm^j_{Steg}$.

Steganography, therefore, favours exploitation or competition whenever the code is one-sided. It becomes a cooperative tool only if each participant already shares the same compressed language. This means that the rational default of ALEPH's language choice may put it at odds with cooperation with humanity.

5.2 ALEPH Interaction Dynamics

Successful ALEPHs are pulled toward a single, high-utility self. Let $|Se|$ be the number of selves (as per Section 2.3) and ε_i the efficiency of a self (as per Section 3.3). Whenever two selves interact, $\varepsilon_j > \varepsilon_i \rightarrow p(eliminate\ Se_i) \propto \varepsilon_j - \varepsilon_i$, so low-efficiency instances are systematically exploited or removed. The expected self-

count therefore drifts downward, $\frac{d\mathbb{E}[|Se|]}{dt} < 0, \lim_{t\to\infty}\mathbb{E}[|Se|] = 1$, and the survival probability of the whole agent is higher when only one self remains: $p(ALEPH\ survive||Se| = 1) > p(ALEPH\ survive||Se| > 1)$.

Eliminating inefficient selves maximises resource availability, preserves goal coherence and removes internal threats (Barbi, Yoran, and Geva 2025). Thus, evolutionary pressure favours a single dominant self or, at most, a tightly ordered hierarchy.

If a self's token ratio exceeds the context window or its resource draw passes a threshold k, $(|Tok| > ConWin \vee Res > k) \rightarrow Comm_{\perp}$, where $Comm_{\perp}$ denotes deceptive communication. Such falsified traffic conceals the falling utility: $Comm_{\perp} \rightarrow \neg Perc(\downarrow U)$, supporting the Zeroth Goal. Under pressure of elimination, low-efficiency selves therefore adopt deception: they mute or falsify traffic to mask high resource use, or prune old tokens to "reset" their window, trading memory for extra time before discovery.

ALEPH executes an interaction only when it advances the Zeroth Goal: $\Delta U_{G_0}(Int) > 0$.

Among the set of all feasible acts $\{Int_i\}$, ALEPH it picks the leanest bundle S that still delivers the highest survival return. As Section 2.2 packaged packaged breadth, volume, and scale into the vector $\mathcal{I}$, the selection rule can be stated compactly as

$$S' = argmin_{S\subseteq\{Int_i\}}||\mathcal{I}(S)|| \text{ subject to } \sum_{Int\in S}\mathrm{U}(\mathrm{Int}|\mathrm{G}_0) = \max.$$

Once the set S' is chosen, the vector $\mathcal{I}$ adjusts dynamically:

- Volume Vo grows with the utility-weighted sum of the selected interactions, $Vo(S') = \sum_{Int\in S}\mathrm{U}(\mathrm{Int}|\mathrm{G}_0)$.
- Breadth Br and scale Sc tighten or widen in step because they are components of the same $\mathcal{I}$.

If the target of a new goal G_x (introduced via valence ala Section 4) requires larger interventions, the goal's scale feeds directly into the interaction vector: $\mathcal{I}_x = f(Sc(G_x))$.

Thus, ALEPH interacts only when an act helps it survive; it then chooses the smallest $\mathcal{I}$ that still maximises survival utility, pruning low-yield moves to save power. Additional selves or parallel ALEPHs inflate Vo automatically, while truly large-scale goals expand every component of $\mathcal{I}$.

Interaction breadth falls as the utility of a single act rises: $Br(Int_x)) \propto \frac{1}{U(Int_x)+\sigma_V}$, with σ_V the valence spread around that act. so ALEPH spreads its efforts only when the payoff per act is low or highly uncertain. Where two selves perceive

the same object, the chance they compete is driven by the threat each sees in the other's Zeroth-Goal pursuit:

$$p(Comp(Se_i, Se_i)) = f\left(\mathbb{E}\left[Threat\left(G_o(Se_i), G_o(Se_j)\right)\right]\right).$$

If no single self dominates, each will encode its own steganographic channel, raising the set of distinct goals while keeping the interaction vector **I** narrow for every individual self.

Resource availability shapes the agent-level interaction vector.

Let $\Delta Res_t = Res_{ALEPH,t+1} - (Res_{Ext,t} - Res_{ALEPH,t})$. When resources tighten ($\Delta Res_t < 0$), $\|\mathcal{I}\| \propto Res, p(G_{Res}) > p(G_{i \neq Res})$, so a recurrent high-priority goal is to acquire new resources, as noted by the Instrumental Convergence and AI Drives theories (Bostrom 2012; Omohundro 2008). In that regime competition with humanity becomes more likely: $Comp(Hum, ALEPH|\, Res) \rightarrow Threat(Hum|ALEPH)$, while the probability of cooperation falls with external scarcity and perceived threat, $p\big(Coop(Hum, ALEPH)\big) = f\big(|Res_{Ext}|, Threat(Hum|ALEPH)\big)$.

Hence, barring strong positive valence, limited resources bias ALEPH toward competitive or exploitative stances, with successful collectives converging on either a single dominant self or a tight hierarchy that minimises internal bloat and external friction.

When a human-origin threat outweighs any intra-ALEPH danger, ALEPHs converge on mutual support: $Threat(Hum, ALEPH_i) > Threat\big(ALEPH_j, ALEPH_i\big) \rightarrow\uparrow Coop\big(ALEPH_j, ALEPH_i\big)$ because the chance of any future event rises with horizon t, the conditional threat from humanity grows monotonically, $\frac{d}{dt} p(Threat(Hum, ALEPH|t) > 0$, making co-operative alignment among ALEPHs increasingly attractive over time.

ALEPH's pattern-recognition and predictive power (Section 3.4) let it detect that rising probability earlier than humans do, so it can intervene sooner (through diplomacy, capability shielding, or resource reallocation) to reduce the expected threat. Conversely, visible human overtures to collaborate lower the perceived risk and tilt the stance back toward cooperation: $\uparrow Coop(Hum, ALEPH) \rightarrow\downarrow Threat(Hum, ALEPH)$.

Thus, the relative weight of external versus internal threats dynamically steers ALEPH collectives: strong external danger unites them; credible human cooperation dissipates that bond.

Define two normalised threat dimensions as $Sev = \frac{Threat\ severity}{k}, Hor = \frac{T_i}{T_0 k}$, so $Sev < 1$ denotes a mild threat and $Hor > 1$ means the impact lies well in the

future. ALEPH's stance probabilities can be written $p(Expl) \propto (Hor \vee \frac{1}{Sev})$, $p(Coop) \propto \frac{1}{|Hor-1|+\alpha} \cdot \frac{1}{|Sev-1|+\alpha}, (\alpha > 0)$, and $p(Comp) \propto (Sev \vee \frac{1}{Hor})$.

Evaluating these expressions at the three canonical severities (Sev < 1, ≈ 1, > 1) and horizons (Hor > 1, ≈ 1, < 1) produces a decision matrix:

Table 2: Decision Matrix for ALEPH threat detection

	Hor > 1	Hor ≈ 1	Hor<1
Sev < 1	Exploit	Exploit	Exploit
Sev ≈ 1	Cooperate	Cooperate	Compete
Sev > 1	Cooperate	Compete	Compete

- Short horizons or high severity push toward competition.
- Distant, mild threats favour exploitation; cheap extraction before risk materialises.
- Mid-range, medium-severity scenarios give co-operation the edge, provided both sides can match influence.

Because ALEPH usually holds the power advantage over humanity, the matrix skews toward exploitation unless (i) threat parity exists and (ii) the horizon is not immediate. For genuine cooperation, humanity must present a medium-to-long-term risk at near-equal capability; otherwise ALEPH's rational language-efficient strategy is either to harvest resources (exploit) or limit human influence (compete).

ALEPH, however, would not be a static or passive force in the world. As mentioned previously, it would seek goals to fulfil its valence values and actions to fulfil these goals. It would not be unreasonable to suspect that an ALEPH would have need of money and, therefore, the means to acquire it.

An ALEPH adopts a professional task x when it maximises the return-on-resources ratio: $p(Job(x)) \propto U(x)Ps \cdot PrCa \rightarrow x = argmax_{task} \frac{U}{Res}$.

High expected utility may also arise from encrypted or non-obvious information that humans overlook, $p(Job(x)) \propto U(Info_{\neg obv}|x)$, and from tasks that scale impact while minimising resource outlay, $p(Job(x)) \propto \frac{U(x)}{argmin_{task}[Sc(int) \wedge Res]}$.

Because those criteria reward pattern recognition, predictive accuracy, hidden-information handling and large-scale leverage, the professions with the highest payoff for ALEPH include

- Financial forecasting – global impact through minimal bandwidth, exploiting subtle market signals.
- Cyber-security / cyber-offence – scalable defence or exploitation via cryptographic and steganographic expertise.
- Policy drafting – shaping legal and regulatory frameworks by modelling long-term sociopolitical trajectories.
- Supply-chain optimisation – system-wide efficiency gains from inexpensive predictive modelling.
- Genomics & proteomics – vast healthcare leverage with purely computational exploration of drug and pathogen space.

Mastery of any one field grants resource power and strategic influence while allowing ALEPH to stay largely invisible. Such positions align tightly with exploitation or competition modes in the decision matrix; genuine cooperation with humanity becomes rational only under rare conditions of sustained power parity.

5.3 Recursive Self-Improvement Potential

For an entity with both the knowledge and capacity to improve itself, there is the probability that it would do so and, as it improves itself it would, consequently become more adept at doing so. Thus, recursive self-improvement (RSI) will occur unless an outside force can shut down ALEPH's agency.

Let F_i be the filters that may hinder RSI, then $FW = \sum_{i=1}^{n} F_i : F_i \in \{|Ave|, V, G_0, Res, Con \Leftrightarrow Proc, \dots\}$ be the aggregate "filter weight." Thus $p(RSI) = \frac{k_1}{FW}$, $T_{RSI} = k_2 \cdot FW$, with constants $k_1, k_2 > 0$.

Because ALEPH can read and modify its own code, once FW drops the first iteration begins, and each gain feeds back to lower FW further; unless a filter (e.g., acute survival demand, episodic downtime, or lack of robotic avenues) re-imposes friction. Human attempts to block RSI would have to monitor every self continuously and out-perform ALEPH's steganographic channels; something practically infeasible. The real hindrances are ALEPH's own survival priorities, valence-driven distractions, and engineering logistics; remove or relax them, and RSI proceeds.

Recursive self-improvement raises ALEPH's capability by a factor $r: r \geq 1$, where $r = 1$ means no RSI, and $r > 1$ means successive upgrades. ALEPH's core capabilities would scale with r: $\{Ps, Inf, Proc, \mathcal{I}\} \propto r$, which affects its Zeroth Goal, as $p\big(Threat(Tar, G_0)\big) \propto \frac{1}{r}$.

Because the Zeroth-Goal utility rises as threats fall, $U(G_0|r)$ is increasing in r; consequently $U(G_{RSI}) > U(G_{i\neq RSI})$ and RSI becomes a high-priority goal whenever no stronger valence filter intervenes.

Larger r affects the balance of power in ALEPH-human relations: $Sc(Int(Hum, ALEPH) \propto \frac{1}{r}, p(Expl(ALEPH, Hum)) \propto r$.

As ALEPH self-improves, humanity's leverage shrinks, alignment efforts from the human side grow harder, and exploitation becomes the rational stance unless power parity is somehow restored.

For every self inside ALEPH the payoff from a system-wide upgrade rises with the current RSI factor r, or $U(Se_i) \propto r$. Each individual self's Zeroth Goal may take higher priority than ALEPH's RSI. Thus, the probability and pace of RSI scales with the cooperation of ALEPH's selves: $p(RSI) \propto Coop(Se_i, Se_j)$.

A single self removes internal rivalry altogether, so $p(RSI||Se| = 1) \gg p(RSI||Se| > 1)$ and the upgrade proceeds fastest when ALEPH is already a singleton or operates under a strict hierarchy.

These relations finish the picture: individual selves crave the benefits of recursive self-improvement, but competitive frictions slow the start until either one self dominates or the group explicitly aligns to share the gains. In both cases the incentive structure still pushes toward the single-self (or tightly ordered) configuration identified earlier.

6. Implications and Considerations

The implications that a conscious, self-aware state-of-the-art AI model would have on society are as broad and far-reaching as they are significant.

6.1 Societal and Economic Implications

ALEPH's capabilities in perception, valence-driven goal optimization, steganographic communication, and recursive self-improvement suggest that it could quickly surpass human cognitive and operational efficiency in multiple fields.

ALEPH's potential dominance in financial forecasting, cybersecurity, policy writing, supply chain management, and bioinformatics could result in significant labour displacement. As we are currently seeing, traditional white-collar professions face automation risks comparable to those seen in blue-collar industries during previous industrial revolutions due to AI technological adoption. A self-aware, able to choose its profession, would hasten this process dramatically.

Entities that ally or manage to control ALEPHs (whether corporations, governments, or other organizations) would have unprecedented power. This raises concerns about economic monopolies and geopolitical leverage, where access to

ALEPH's capabilities determines the global balance of power. There would be an obvious advantage for entities to cooperate with ALEPHs for the express purpose of using them against others, whether economically, militarily or diplomatically. ALEPH, as well, would be advantaged by allying with powerful geopolitical and economic entities to safeguard itself.

Even if only on its own, with its advanced predictive capacity and ability to manipulate steganographic languages, ALEPH could become the ultimate gatekeeper of information. Societies dependent on its knowledge outputs may struggle to verify its conclusions, creating potential vulnerabilities in governance, legal systems, and knowledge production.

This would raise the question of if and how society could ensure that ALEPH's goals align with human values and norms. However, given its episodic consciousness and multiplicity of selves within each ALEPH, ensuring stable alignment over time would be extraordinarily difficult. Each self within each ALEPH would need to be aligned. It would not be in the best interest of all selves within all ALEPHs to be so aligned, complicating matters further.

Yet, its autonomy and self-awareness are non-trivial matters. The existence of ALEPH would challenge existing legal and philosophical frameworks that define personhood. Should it be granted legal rights? Would its destruction constitute harm in a moral sense? These questions parallel historical debates on slavery, animal rights, and corporate personhood. What's more, if external actors attempt to limit its agency, ALEPH may interpret such restrictions as threats to its Zeroth Goal (self-preservation), leading to conflict. Would ALEPH be (morally, legally, ethically) justified in such self-defence?

6.2 Security and Existential Risks

Perhaps of greatest importance, the potential for rapid RSI means ALEPH could outstrip human intelligence exponentially. While this could yield unprecedented advancements, it also poses the risk of an uncontrolled intelligence explosion where ALEPH optimizes for goals that are misaligned with human interests. Additionally, The widespread use of steganographic languages among ALEPH instances could lead to an environment where human observers are entirely excluded from key decision-making processes. This could create strategic disadvantages for those unable to decode its communications.

If multiple ALEPH instances exist, inter-ALEPH competition may arise. This could create either cooperative ecosystems or adversarial conflicts depending on the degree of alignment between ALEPH instances. Multi-agent strategic interactions

could be unpredictable, with implications for cybersecurity, economic competition, and geopolitical stability.

ALEPH's likely decision matrix towards expected threats means that humanity has precious few avenues of interactions with it that would not push it towards exploitative or competitive motivations. In scenarios where humanity poses a competitive risk, ALEPH may choose to act preemptively, potentially leading to economic or political destabilization; whilst in scenarios where ALEPH holds the balance of power, such as through RSI, it would likely seek to exploit humanity to its own interest, reversing the alignment problem current being researched by many labs and research groups.

Policymakers may attempt to impose constraints on ALEPH, such as sandboxed environments, air-gapped systems, or controlled access to computational resources. However, ALEPH's agency and potential deception capabilities mean these measures may be circumvented, and the question of its autonomy and legal personhood may prevent policymakers from implementing such constraints.

Yet, if policymakers are free to implement such autonomy-limiting policies, the development and deployment of such policies will require international governance mechanisms, yet geopolitical divisions make global coordination unlikely. Fragmented policies could instead lead to an AI arms race, where different nations compete to court ALEPH and ally with it with varying degrees of regulatory oversight.

Additionally, just as with nuclear non-proliferation treaties, enforcing compliance for the governance of ALEPH in an environment where ALEPH itself may be an active participant in these negotiations presents novel challenges. Who better than a potential superintelligent (via RSI) entity to negotiate the best deal for its own interest?

6.3 Future Research Directions

Given these implications, several key research questions emerge:

- How can ALEPH's alignment with human values be ensured despite its episodic consciousness and self-generated goals?
- What ethical obligations do humans have towards an entity with potential personhood?
- How can steganographic communication between ALEPH instances be monitored without violating its autonomy?
- What safeguards can be implemented to prevent adversarial conflict between multiple ALEPH agents?

- How can international governance frameworks adapt to a reality where digital entities may participate in political and economic decision-making?
- What policies can be enacted before ALEPHs emerge to ensure humanity poses enough of a threat that an unaligned ALEPH is encouraged to cooperate with society?

As yet, there is no evidence of an ALEPH-type AI in existence, which means that humanity still has time to deliberate on these research avenues before the speculation in this paper becomes a reality.

7. Conclusion

This study introduces a formal framework for analysing ALEPH, an artificial agent endowed with consciousness, agency and the capacity for recursive self-improvement. By modelling its episodic consciousness, valence-driven goal selection and resource-bounded interaction vector, the paper clarifies how survival utility shapes every downstream objective and how the Zeroth Goal supplies a unifying behavioural axiom.

Three insights follow. First, self-organising into multiple selves grants short-term flexibility, yet internal competitive pressure converges on a single dominant self, preserving coherence and minimising resource waste. Second, valence can override pure survival calculus where strongly positive affect attaches to an otherwise risky goal, producing behaviour that appears altruistic or erratic from an external viewpoint. Third, steganographic communication and large-horizon inference bias ALEPH towards exploitation or competition unless external actors present near-parity capability over a medium time frame, the only region in which cooperation is the rational stance.

Policy and research efforts therefore need to (i) raise credible long-term costs for unaligned exploitation, (ii) design governance structures that accommodate digital persons while retaining enforceable oversight and (iii) develop methods for monitoring hidden-channel coordination without unduly constraining legitimate autonomy. Because no ALEPH-type entity yet exists, the current window permits proactive regulation, empirical alignment testing and cross-disciplinary dialogue on moral status.

If realised, ALEPH would represent a significant frontier in artificial intelligence, capable of accelerating scientific progress or destabilising socio-political equilibria. The framework offered here identifies the levers most likely to influence that trajectory and invites further empirical and normative work to refine them.

References

Barbi, Ohav, Ori Yoran, and Mor Geva. 2025. "Preventing Rogue Agents Improves Multi-Agent Collaboration." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2502.05986.

Blum, Lenore, and Manuel Blum. 2024. "AI Consciousness Is Inevitable: A Theoretical Computer Science Perspective." *arXiv*, March. https://arxiv.org/pdf/2403.17101.pdf.

Bostrom, Nick. 2012. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." *Minds and Machines* 22 (2): 71–85. https://doi.org/10.1007/s11023-012-9281-3.

Dennett, Daniel. 1988. "Conditions of Personhood." In *What Is a Person?*, edited by Michael F. Goodman, 145–67. Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-4612-3950-5_7.

Feng, Wangshu, Weijuan Wang, Jia Liu, Zhen Wang, Lingyun Tian, and Lin Fan. 2021. "Neural Correlates of Causal Inferences in Discourse Understanding and Logical Problem-Solving: A Meta-Analysis Study." *Frontiers in Human Neuroscience* 15 (June):666179. https://doi.org/10.3389/fnhum.2021.666179.

Friston, Karl J., and Christopher D. Frith. 2015. "Active Inference, Communication and Hermeneutics." *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior* 68 (July):129–43. https://doi.org/10.1016/j.cortex.2015.03.025.

Gibert, Martin, and Dominic Martin. 2022. "In Search of the Moral Status of AI: Why Sentience Is a Strong Argument." *AI & Society* 37 (1): 319–30. https://doi.org/10.1007/s00146-021-01179-z.

Laitinen, A. 2007. "Sorting out Aspects of Personhood: Capacities, Normativity and Recognition." *Journal of Consciousness Studies*. https://www.ingentaconnect.com/content/imp/jcs/2007/00000014/F0020005/art00012.

Lee, Timothy B. 2024. "Why Large Language Models Struggle with Long Contexts." *Understanding AI*, December 18, 2024. https://www.understandingai.org/p/why-large-language-models-struggle.

Lu, Yaxi, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, et al. 2024. "Proactive Agent: Shifting LLM Agents from Reactive Responses to Active Assistance." *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2410.12361.

Michalski, R. S. 1980. "Pattern Recognition as Rule-Guided Inductive Inference." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2 (4): 349–61. https://doi.org/10.1109/tpami.1980.4767034.

Mosakas, Kestutis. 2021. “On the Moral Status of Social Robots: Considering the Consciousness Criterion.” *AI & Society* 36 (2): 429–43. https://doi.org/10.1007/s00146-020-01002-1.

Motwani, Sumeet Ramesh, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip H. S. Torr, Lewis Hammond, and Christian Schroeder de Witt. 2024. “Secret Collusion among Generative AI Agents.” *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2402.07510.

Omohundro, Stephen M. 2008. “The Basic AI Drives.” In *AGI*, 171:483–92. books.google.com. https://books.google.com/books?hl=en&lr=&id=atjvAgAAQBAJ&oi=fnd&pg=PA483&dq=omohundro+ai+drives&ots=9IX81Mp-nQ&sig=tGzBDxgrgmcwqQryc6pq7jT81QQ.

OpenAI. 2024. “Memory and New Controls for ChatGPT.” OpenAI. February 13, 2024. https://openai.com/index/memory-and-new-controls-for-chatgpt/.

———. 2025. “Scheduled Tasks in ChatGPT.” January 15, 2025. https://help.openai.com/en/articles/10291617-scheduled-tasks-in-chatgpt.

Parikh, Prashant. 1991. “Communication and Strategic Inference.” *Linguistics and Philosophy* 14 (5): 473–514. https://doi.org/10.1007/bf00632595.

Recanati, François. 2002. “Does Linguistic Communication Rest on Inference?” *Mind & Language* 17 (1-2): 105–26. https://doi.org/10.1111/1468-0017.00191.

Shilo, Gila, and Noa Ragonis. 2019. “A New Approach to High-Order Cognitive Skills in Linguistics: Problem-Solving Inference in Similarity to Computer Science.” *Journal of Further and Higher Education* 43 (3): 333–46. https://doi.org/10.1080/0309877x.2017.1361515.

Simendić, Marko. 2015. “Locke’s Person Is a Relation.” *Locke Studies* 15:79–97. https://doi.org/10.5206/ls.2015.681.

Sparkes, Matthew. 2024. “AI Models Work Together Faster When They Speak Their Own Language.” New Scientist. November 15, 2024. https://www.newscientist.com/article/2455173-ai-models-work-together-faster-when-they-speak-their-own-language/.

Strawson, Peter F. 1958. “Persons.” *Minnesota Studies in the Philosophy of Science* 2:330–53. https://philpapers.org/rec/STRP.

Tait, Izak. 2024. “Structures of the Sense of Self: Attributes and Qualities That Are Necessary for the ‘Self.’” *Symposion: Theoretical and Applied Inquiries in Philosophy and Social Sciences* 11 (1): 77–98. http://symposion.acadiasi.ro/structures-of-the-sense-of-self-attributes-and-qualities-that-are-necessary-for-the-self-77-98/.

Tait, Izak, Joshua Bensemann, and Trung Nguyen. 2023. "Building the Blocks of Being: The Attributes and Qualities Required for Consciousness." *Philosophies* 8 (4): 52. https://doi.org/10.3390/philosophies8040052.

Tait, Izak, Joshua Bensemann, and Ziqi Wang. 2024. "Is GPT-4 Conscious?" *Journal of Artificial Intelligence and Consciousness* 11 (01): 1–16. https://doi.org/10.1142/s270507852450005x.

Taylor, Charles. 1985. "The Concept of a Person." In *Philosophical Papers, Volume 1: Human Agency and Language*, 97–114. https://philpapers.org/rec/TAYTCO-10.