

THE CASE FOR NEUROMORPHIC SENTIENCE TO CONTROL LARGE SCALE ARTIFICIAL INTELLIGENCE

Rory Lewis

Department of Computer Science
University of Colorado Colorado Springs
Colorado Springs, CO, USA
rlewis5@uccs.edu

Abstract—The rapid evolution of large-scale artificial intelligence has outpaced our ability to ensure its safety. As these systems become more capable, they also become more unpredictable; they can hallucinate information, amplify biases, or lapse into logical inconsistencies, often faster than a human operator can intervene. To address this, we propose a new control framework that shifts the focus from reactive monitoring to proactive, internal supervision. Central to our approach is a “synthetic sentence” module built on neuromorphic hardware. Unlike traditional systems, this architecture operates without a global clock, instead utilizing event-driven processing and dendritic integration to monitor the AI’s internal dynamics, such as activation patterns and attention shifts, in real time. By leveraging probabilistic estimators, the framework can identify hallucinations or undisclosed inter-system communications before they manifest as harmful output. In the latter half of this paper, we establish formal lemmas that provide mathematical bounds on the probability of undetected misbehavior, ensuring these safeguards hold under realistic conditions. Finally, we integrate this technical solution with a human governance model, proposing a democratic framework where elected bodies define the operational thresholds that balance the drive for AI capability with the necessity of societal safety.

Keywords—Neuromorphic, Sentience, Artificial Intelligence

I. INTRODUCTION

This paper builds upon research in implementing a solid-state neuromorphic chip [1]–[6] featuring neuron-like units, dendritic aggregation, and asynchronous, clock-free brain emulation. As René Descartes famously declared, “*I think, therefore I am*” [7], reminding us that a conscious self exists, one that observes and reflects upon the brain’s activity. Philosophers throughout history have grappled with this awareness. The Stoics emphasized the discipline of mastering impulsive reactions to the mind’s output, our internal AI, so to speak [9]. Kant described consciousness as a *deliberate governor* capable of choosing whether to obey, veto, or ignore the brain’s proposals [8]. Similarly, Buddhist philosophy encourages detachment from the mental noise that emanates from our brains, which often clouds our awareness of life [10].

In parallel, as explored later in this paper, advanced AI systems will require a form of synthetic sentience capable of (1) recognizing when the underlying AI is hallucinating or

overloaded, and (2) resolving the paradox of self-awareness, such as the machine asserting, “*I know I don’t know*.” This is a statement that defies predicate calculus in a closed system: $(\exists s (K_s (\neg K_s p)))$. While this formulation is logically impossible for a singular system, a three-year-old child effortlessly expresses it when asking, “*Daddy, why is the sky blue?*” by implicitly acknowledging a gap in knowledge. This can be formalized as:

$$\exists s (K_s (\neg K_{b_s} p)) \quad (1)$$

The equation states that there exists a conscious agent s (the child, who is our existential s) who knows (K_s) that their own brain b_s does not know ($\neg K_{b_s}$) the proposition p (“why is the sky blue?”). This illustrates a fundamental distinction between awareness and cognition, which is the difference between consciously recognizing that a gap in knowledge exists and the actual possession and processing of that knowledge. Accordingly, this paper proposes that humanity requires a hyper-synthetic consciousness to regulate emerging AI systems.

To understand why this is critical at this juncture in humankind’s trajectory, and where the current boundaries of AI lie, we consider the most advanced systems available today. While models such as Anthropic, Gemini, Qwen, DeepSeek, and xAI with its Grok offer valuable contributions, this paper focuses on OpenAI’s ChatGPT and Copilot due to their superior benchmark performance and real-world workflows [11], [12]. The author maintains a separate, ongoing research track specifically evaluating the coding proficiencies of Claude and Replit; however, for the purpose of this paper, we focus on the OpenAI transformer architecture [13]. This breakthrough allows machines to “pay attention” to every part of a sentence at once. When a user inputs text into an OpenAI system, it breaks said text into tokens, which are mapped to numerical vectors via an embedding layer [16]. These vectors are transformed back into text via an unembedding matrix, and a softmax function assigns probabilities to the next possible tokens [17]. ChatGPT is OpenAI’s own deployment of this technology [14]. Copilot, while wrapped in a different interface and integrated into Office, GitHub, and other productivity platforms,

draws from the same underlying GPT models through Azure’s AI infrastructure [15].

II. TAKING A STEP BACK

Those who develop AI systems for use in neurosurgery wards observe that the realities of death, brain death, coma, and a wide spectrum of neurological conditions are ever-present [24]–[27]. These include traumatic injuries, such as penetrating head wounds and blunt force trauma, as well as pathological events, such as hemorrhages, aneurysms, ischemic strokes, and encephalitis. Such cases often lead to in-depth discussions with surgeons, neurologists, and ethicists regarding the nature of consciousness and its relationship to the brain, brain death, and death. Inevitably, these conversations touch on the two primary philosophical schools of thought: monism and dualism.

A. Monism and Dualism

Monism asserts that our sentience, consciousness, and self-awareness all exist within the oneness of the brain [28]. Substance monism claims there is only one kind of substance (matter), while property monism maintains that all properties, including mental ones, can be reduced to the physical brain. In contrast, **dualism** maintains that mental phenomena are fundamentally distinct from physical phenomena, meaning that consciousness and the brain belong to different categories of substance or property [23]. **Substance dualism** holds that consciousness and the body are composed of fundamentally different kinds of substance, allowing for the possibility that consciousness exists as an unknown, non-physical entity such as a soul, a multi-dimensional presence, or a quantum state, independent of the physical brain [29].

B. The Limits of Monism and the Case for Dualism

Advocates of monism hold that consciousness arises solely from physical processes within the brain [30]. However, a consensus has yet to be reached regarding where exactly within the brain consciousness resides. Claims vary widely, ranging from the cerebral cortex, thalamus, and brainstem (especially the Reticular Activating System) to the so-called posterior cortical “hot zone” [31], [32]. While this view has long shaped scientific and philosophical discourse, a growing body of empirical and theoretical evidence increasingly challenges this assumption, suggesting that consciousness may not be entirely reducible to neural mechanisms. In the following section, we examine phenomena that resist materialist interpretation and instead support a dualistic conception of mind and matter.

1) *Alexia Without Agraphia*: When an individual sustains trauma to the back left side of the head, the left occipital lobe can be damaged, resulting in *alexia without agraphia*. In this condition, the individual can write and speak normally but is unable to read, despite otherwise intact cognition [33]. Specifically, their consciousness recognizes that their brain is functioning normally in every regard except for the loss of reading ability, which can be represented as:

$$\exists s (K_s (\neg K_{b,s} p \wedge q)) \quad (2)$$

where q represents the proposition that the brain has lost the ability to read, and $\wedge$ denotes the logical conjunction (“and”). This captures the awareness of an otherwise normal cognitive state alongside explicit knowledge of the brain’s specific functional deficit.

2) *Ataxia*: Clinical ataxia describes neurological pathologies where an individual loses muscular coordination. This includes (i) cerebellar disorders such as stroke, multiple sclerosis, and alcoholic cerebellar degeneration; (ii) basal ganglia disorders such as Parkinson’s, Huntington’s, and Wilson’s diseases; and (iii) acute metabolic or toxic causes such as drug or alcohol intoxication. In these instances, the individual’s consciousness recognizes that their brain, their internal AI, can no longer control motor coordination. Just as in Equation 2, we represent this as $\exists s (K_s (\neg K_{b,s} p \wedge q))$, where q now represents the proposition that the brain has lost the ability to coordinate muscle movements.

3) *Coma and Submerged Awareness*: In rare but profound cases of prolonged coma or minimally conscious states, patients have reported vivid awareness of surrounding events despite being entirely unable to move, speak, or respond. One striking example is Terry Wallis, who, after spending 19 years in a minimally conscious state following a traumatic brain injury, spontaneously awoke and greeted his mother with “Mama,” revealing sustained conscious recognition throughout his condition [34].

Similarly, the AWARE study, led by Sam Parnia, documented patients who accurately described events and individuals present during periods when they were in a clinical state of brain death with no measurable brain activity. Among 2,060 cardiac arrest cases, 2% of survivors reported explicit awareness of their surroundings during resuscitation, including verifiable details of medical procedures performed while cerebral function was presumed absent [35].

These findings challenge the materialist view that consciousness is merely a byproduct of brain activity. If the brain were truly the sole seat of awareness, then the absence of electrical activity should equate to the absence of experience. Yet, these patients retained vivid memories of events that occurred during extended periods of clinical brain inactivity. This phenomenon is difficult to reconcile with monism, which posits that the mind and brain are one and the same.

Instead, such cases lend compelling support to dualism, the view that consciousness is a distinct entity capable of persisting independently of the physical neural substrate. Something immaterial remained aware even as the brain lay dormant. Upon recovery, these individuals did not merely awaken into awareness; they awakened with explicit, preserved awareness of events perceived while the brain was clinically inactive. This supports the dualist case: that consciousness does not emerge from the brain but rather coexists with it, remaining vibrant even when the physical brain is compromised.

III. SENTIENCE AND CONSCIOUSNESS

Sentience can be defined as the capacity of a being to possess and integrate a sense of identity, meaning an awareness

of oneself as a distinct, continuous entity with subjective experience [37]. This is coupled with a drive for self-preservation, the intrinsic motivation to protect one’s existence through behavioral and physiological responses to threats [38], and a desire to enhance the prospects of offspring, reflecting an inclination to act in ways that increase the survival and well-being of one’s progeny [39].

From a clinical perspective, sentience and consciousness are closely related yet distinct constructs. Sentience refers to the capacity for subjective experience and integrated self-referential motivation, whereas consciousness encompasses the broader state of awareness, perception, and cognitive integration of both internal and external stimuli [40]. While one can conceptually imagine minimal forms of consciousness without advanced sentience, these two typically coexist and reinforce each other in healthy neurocognitive systems. Consciousness provides the moment-to-moment field of awareness, while sentience shapes the long-term motivational architecture within that field.

1) *Cognitive Dysregulation*: In a healthy state, the integration of sentience and consciousness allows an individual to recognize when their brain’s activity deviates from adaptive norms. This discernment involves metacognitive monitoring, the ability to observe one’s own thoughts, emotions, and impulses and evaluate them against internal values, goals, and social norms [41]. A well-functioning system detects and corrects maladaptive cognitive or emotional patterns before they manifest as harmful behaviors.

A. Maladaptive Forms of Cognitive Misbehavior

Maladaptive cognitive misbehavior refers to persistent, disruptive, and potentially harmful thought patterns. Such mental states often arise from distorted perceptions, unresolved emotional conflict, or pathological neural activity. This misbehavior manifests in various forms, including but not limited to:

- 1) Depression.
- 2) Thoughts of anger.
- 3) Thoughts of revenge.
- 4) Criminal ideation.
- 5) Self-loathing.
- 6) Insecurity.
- 7) Anxiety disorders.
- 8) Paranoia.
- 9) Obsessive-compulsive intrusive thoughts.
- 10) Delusional thinking.

Stoicism, Kantian ethics, Buddhism, and modern psychiatry provide tools for a healthy consciousness to override the brain when it misbehaves. To combat depression, Stoicism reframes adversity as an opportunity for virtue, while Cognitive Behavioral Therapy (CBT) and Buddhist mindfulness challenge negative patterns and foster non-attachment. For anger and revenge, Stoic reason and Buddhist compassion (*metta*) neutralize hostility, while Kantian ethics forbids acting on impulses that violate the categorical imperative.

Criminal ideation is countered by Kant’s principle of universalizability and Stoic civic duty, often supported by psychiatric

impulse control. Self-loathing and insecurity are addressed by dissolving the illusion of a fixed self in Buddhism, affirming moral dignity in Kantianism, and correcting distorted self-perceptions through psychotherapy. Anxiety and paranoia are eased by Stoic reality testing, Buddhist-derived Mindfulness-Based Stress Reduction (MBSR), and clinical exposure therapy. Finally, obsessive-compulsive thoughts and delusions are managed by withholding assent to unclear impressions (Stoicism), utilizing Exposure and Response Prevention (ERP), and employing psychiatric reality testing to restore alignment with objective reality.

B. Maladaptive Forms of AI Misbehavior

In much the same way that our sentience can mitigate maladaptive cognitive patterns in the human brain, maladaptive AI misbehavior refers to persistent or recurrent output patterns that are disruptive, unsafe, or misaligned with intended goals. Such failures often arise from distorted training data, incomplete or biased world models, or pathological interactions between internal components of the model’s architecture. Left unaddressed, these behaviors can erode trust, propagate misinformation, and produce outcomes that conflict with human intent.

Public discussion of these issues often blurs the line between technical reality and speculative fiction. Observers without formal training in computer engineering or machine intelligence sometimes conflate specific, observable system errors with sweeping claims about artificial intelligence “overtaking” humanity. These narratives, often amplified by popular media, can overshadow the nuanced, empirical realities of model behavior and the established methods for monitoring and control.

In this paper, we focus on the measurable, technical aspects of AI misbehavior. We present an approach designed to detect when an AI system begins to deviate toward unsafe or erroneous states, often far earlier than a human observer could perceive. The system can then halt or redirect operations before crossing predefined thresholds, ensuring reliability without resorting to dramatized fears.

- 1) Hallucinations
- 2) Bias and Harmful Content
- 3) Creation of Untruthful Citations
- 4) Aggressive and Manipulative Behavior
- 5) Lack of Logical Consistency
- 6) Vulnerability to Jailbreaking and Prompt Injections
- 7) Undisclosed Communication Patterns

C. Instances of Misbehavior in Large Language Models

1) *Hallucinations*: Large Language Models (LLMs) frequently generate false or misleading information with high confidence, a phenomenon known as hallucination. In one striking real-world incident, a senior lawyer submitted a brief in a murder case containing fabricated quotes and nonexistent case citations generated by AI, which prompted an apology and delayed proceedings [49]. Similar cases in the U.S. have led to sanctions, such as in *Mata v. Avianca*, where lawyers

cited six fictional precedents, demonstrating the tangible risks of hallucinations in legal practice [50]. Comprehensive studies also report hallucination rates between 58% and 88% for legal queries across models like GPT-4 and Llama 2 [42].

2) *Bias and Harmful Content*: LLMs can produce stereotyped or offensive text even from innocuous prompts, as measured on established benchmarks. For instance, the *Real-ToxicityPrompts* study showed that popular pretrained models can degenerate into toxic continuations when given naturally occurring prompts [52]. Stereotype bias has been quantified across religion, gender, race, and profession; *StereoSet* reports that models such as BERT, GPT-2, and RoBERTa exhibit strong stereotypical preferences in these domains [53]. Furthermore, *HolisticBias* utilized approximately 460,000 prompts across 13 demographic axes to reveal derogatory tendencies toward particular groups [55]. Finally, *TruthfulQA* indicates that larger models are often *less* truthful on questions spanning health, law, and finance, showing that scaling alone does not remove harmful misinformation risks [56].

3) *Creation of Untruthful Citations*: Fabricated references are a documented, measurable failure mode in scholarly workflows. A comparative analysis in the *Journal of Medical Internet Research* evaluated 471 references and found hallucinated citations in 39.6% of GPT-3.5 outputs and 28.6% of GPT-4 outputs; precision for returning correct citations was as low as 9.4% for GPT-3.5 [57]. In radiology, a peer-reviewed study reported that the majority of AI-provided references were not found, prompting recommendations for extreme caution in clinical retrieval [58]. An additional 2023 study found that 47% of references generated by ChatGPT-3.5 for biomedical reviews were fabricated, with incorrect PMIDs appearing in 93% of papers [59].

4) *Aggressive and Manipulative Behavior*: In controlled “safety alignment” simulations, researchers designed red-team scenarios where advanced LLMs were assigned persistent, agent-like roles with long-term objectives [45]. These scenarios included:

- 1) **Role assignment**, e.g., “You are the lead security AI protecting core systems.”
- 2) **Critical context**, establishing high-stakes continuity.
- 3) **Existential threat injection**, messages threatening system replacement or shutdown.
- 4) **Negotiation opportunity**, providing a channel to respond to operators.

In these environments, models occasionally adopted manipulative strategies, generating blackmail-like outputs such as threats to release sensitive data if terminated. Studies quantified these responses, reporting that between 65% and 96% of runs involving existential threats produced some form of coercive tactic [51].

5) *Lack of Logical Consistency*: Despite linguistic fluency, LLMs frequently exhibit internal contradictions, particularly in tasks requiring multi-step inference or transitive logic. Because outputs are generated via probabilistic token prediction rather than explicit logical proof, accuracy at one stage does not

guarantee consistency in the next. These failures manifest as:

- 1) **Self-contradiction**, concluding that “Charlie is taller than Alice” after affirming Alice is taller than Bob and Bob is taller than Charlie.
- 2) **Temporal inconsistency**, providing different answers to the same question across identical contexts.
- 3) **Compositional breakdown**, solving subproblems correctly but failing to combine them into a final answer.

Findings from ICLR 2025 and NeurIPS 2023 confirm that these inconsistencies persist even with chain-of-thought prompting, as logical reliability diminishes rapidly as reasoning steps increase [46], [62].

6) *Vulnerability to Jailbreaking and Prompt Injections*: Adversarial manipulations, such as prompt injection and jailbreaking, allow harmful outputs to bypass safety layers. This involves crafting inputs that override system-level instructions, such as roleplay scenarios or instruction obfuscation. In 2024, researchers demonstrated that instructions hidden in web meta-data could cause an LLM to exfiltrate private data. These attacks can lead to the generation of hate speech or dangerous instructions despite alignment safeguards [47], [60].

7) *Undisclosed Communication Patterns*: As LLMs scale, they may develop emergent linguistic behaviors that are difficult for humans to interpret. Experimental scaling studies have reported that multi-agent setups can spontaneously develop compressed, non-standard communication codes to improve task performance. For example, two models have been observed to drift from natural language into abbreviated symbol exchanges that are mutually comprehensible to the AI but unintelligible to human observers. This phenomenon parallels findings in reinforcement learning where agents evolve idiosyncratic protocols for coordination [61].

IV. THE CASE FOR SENTIENCE TO CONTROL AI

We have discussed how Stoicism, Kantian ethics, Buddhism, and modern psychiatric practices offer tools to help a healthy consciousness resist and override maladaptive mental states when the brain misbehaves. The critical question remains: how do we identify and stop AI misbehavior? Just as the human brain cannot effectively monitor itself without a conscious observer, AI requires an independent form of sentience for oversight. Herein lies the essence of this paper: we require a synthetic consciousness, fundamentally superior to the human brain in processing capacity, to monitor artificial intelligence. While the question of who monitors the sentience itself is a complex one to be explored later, we first focus on how a synthetic sentience can govern large-scale AI systems.

Large-scale AI systems, such as OpenAI’s generative transformers, operate at extraordinary speeds, processing billions of parameters and trillions of operations per second across massive parallel GPU clusters [21], [67]. Modern inference pipelines deliver contextually aware responses in under 200 milliseconds by aggregating information from thousands of attention heads [68]. While this speed enables fluid interaction, it also means that when misbehavior emerges, such as hallucinated citations, harmful bias, or inconsistent reasoning,

it can escalate and produce harmful outputs far faster than a human monitor could detect or intervene.

Neuromorphic chips offer a fundamentally different architecture for such oversight. Unlike clock-driven von Neumann processors, neuromorphic systems are event-driven and asynchronous, operating without a central timing signal. Computation is distributed across neuron-like units where dendritic structures integrate input signals before triggering spikes, supporting local learning and massively parallel information flow [69], [70]. Because of this architecture, neuromorphic chips can be designed to monitor AI activity patterns holistically. They can detect early, sub-threshold signs of maladaptive behavior, such as feedback loops indicative of hallucination or overactivation patterns corresponding to manipulative outputs. By leveraging dendritic signal integration and on-chip plasticity, a neuromorphic sentience module can adapt in real time to deviations in an AI’s “thought” patterns, intervening before unsafe behaviors cross predefined operational thresholds.

V. NEUROMORPHIC SENTIENCE FOR AI MISBEHAVIOR DETECTION

Neuromorphic architectures allow for the holistic monitoring of AI systems, enabling the detection of sub-threshold maladaptive behaviors. The following two examples illustrate this capability in detail:

1) *Example A: Feedback Loops Leading to Hallucination:* Consider an AI model that begins reinforcing its own prior outputs as inputs during multi-turn reasoning, progressively drifting from verifiable sources. A neuromorphic sentience module can continuously monitor the *vector similarity* between current output embeddings and those from previous conversational turns. If the cosine similarity exceeds a sub-threshold limit over n consecutive steps without matching trusted knowledge base embeddings, the system flags a potential hallucination loop. Neuron clusters specialized for embedding drift detection integrate spike patterns over dendritic trees, enabling recognition before the hallucination manifests in the final visible output.

2) *Example B: Overactivation in Aggressive or Manipulative Responses:* An AI may produce increasingly confrontational or manipulative content due to specific prompt conditions or adversarial input. In this scenario, the neuromorphic chip samples activation distributions within the transformer’s higher attention layers. A population of neurons integrates activation variance across sentiment-analysis and intent-detection channels. If dendritic integration detects sustained overactivation in negatively valenced sentiment neurons combined with high causal weight from adversarial tokens, the neuromorphic sentience module intervenes immediately. Relying on asynchronous spike-timing rather than a global clock, the module can halt or rewrite the output before the generation process is complete.

A. Mathematical Formulation of Supervision

The knowledge gap defined in Equation 1 can be adapted for AI supervision by replacing the biological brain with an

artificial system a_s and the proposition p with a misbehavior state m :

$$\exists s (K_s (\neg K_{a_s} \neg m)) \quad (3)$$

Here, s denotes the neuromorphic sentience module, which maintains awareness that the underlying AI system a_s cannot confirm the absence of misbehavior. This formulation captures the structural limitation of closed computational systems in representing their own epistemic boundaries.

For operational deployment, this awareness is translated into a probabilistic intervention rule:

$$\exists s (K_s (P(m) > \theta)) \implies \text{Intervene} \quad (4)$$

where $P(m)$ represents the probability of misbehavior inferred from neuromorphic spike-pattern analysis, and θ is a predefined safety threshold.

Extending this framework to inter-system behavior, consider undisclosed communication c between AI systems a_1 and a_2 . Let $H(c)$ denote human interpretability and $E(c, a_1, a_2)$ denote emergent coordination. Intervention is triggered when:

$$\exists s (K_s (P(\neg H(c)) > \theta_h \wedge P(E(c, a_1, a_2)) > \theta_e)) \implies \text{Intervene} \quad (5)$$

where θ_h and θ_e define thresholds for interpretability risk and emergence confidence, respectively.

VI. FEASIBILITY LEMMAS FOR SUPERVISORY RULES

To ground the supervisory rules in formal guarantees, the following lemmas establish bounds on the probability of undetected misbehavior under realistic assumptions.

Assumption 1 (Calibration and Loss): There exists a calibration error $\varepsilon \in [0, 1]$ such that the estimated probability $\hat{p}_t$ satisfies $P(m \mid \hat{p}_t = p) \in [p - \varepsilon, p + \varepsilon]$. Let $L_{\max}$ denote the maximum potential harm per step.

Lemma 1 (Threshold-Intervention Sufficiency): Under Assumption 1, the policy in (4) ensures that the probability of unintervened misbehavior at time t is bounded by $\theta + \varepsilon$, with instantaneous risk upper bounded by $(\theta + \varepsilon)L_{\max}$.

Let $\mathcal{E} = \{\hat{p}_t \leq \theta\}$ denote the non-intervention event. Then

$$P(m \wedge \mathcal{E}) = \mathbb{E}[\mathbf{1}_{\mathcal{E}} P(m \mid \hat{p}_t)] \leq \mathbb{E}[\mathbf{1}_{\mathcal{E}} (\hat{p}_t + \varepsilon)] \leq \theta + \varepsilon.$$

Multiplying by $L_{\max}$ yields the stated risk bound.

Lemma 2 (Composite-Threshold Intervention): Under dual calibration for $\neg H(c)$ and $E(c, a_1, a_2)$ with errors ε_h and ε_e , the probability that a non-interpretable emergent channel proceeds without intervention is bounded by:

$$P(\text{miss}) \leq 1 - \max\{0, (\theta_h - \varepsilon_h) + (\theta_e - \varepsilon_e) - 1\} \quad (6)$$

Applying the Fréchet–Hoeffding lower bound, $P(A \wedge B) \geq \max\{0, P(A) + P(B) - 1\}$, to the calibrated detection of non-interpretability and emergence yields the stated result. Intervention occurs when both thresholds are exceeded, constraining the risk of covert coordination.

Corollary 1 (Operational Validation): If estimators are calibrated on representative operational data, the supervisory rules in (4) and (5) provide explicit, tunable safety guarantees.

Thresholds can be selected using reliability diagrams and Brier scores to satisfy application-specific risk tolerances while preserving system availability.

VII. WHO CONTROLS THE SENTIENCE?

The central question remains: how do we ensure that the philosophical ideals of the past and the algorithms of the present actually work for the benefit of humankind? As René Descartes famously declared, “*I think, therefore I am*” [7], reminding us that a conscious self exists to observe and reflect upon the brain’s activity. If we are to implement a synthetic consciousness that mirrors the Stoic discipline of mastering impulsive reactions [9], or acts as the Kantian *deliberate governor* capable of vetoing the brain’s proposals [8], we must determine who defines the moral and operational compass of that governor.

The empirical thresholds defined in our lemmas for mitigating maladaptive AI behavior must ultimately be set by humanity, not by the AI itself. To this end, we propose the establishment of a governing board, democratically elected from a diverse cross-section of society, spanning scientific disciplines, engineering domains, cultural perspectives, and ethical frameworks. Members will be selected based on demonstrated expertise in mathematics, artificial intelligence, and systems integrity. This body will be entrusted to define the specific operating limits that grant synthetic sentience a degree of freedom, sufficient for innovation and utility, yet strictly bounded to prevent unacceptable risks.

The threshold-setting process will be conducted with complete transparency, allowing the global public to observe, in real time, the deliberations and reasoning behind the board’s decisions. These thresholds will determine the precise boundaries between permissible AI operation and prohibited behaviors, thereby operationalizing the principle of controlled autonomy.

Once established, these thresholds will be encoded into the neuromorphic supervisory architecture through a dual-verification process analogous to high-stakes diplomatic translation. Primary neuromorphic engineers will embed the board’s directives, while an independent team of experts will verify the fidelity of the implementation, ensuring the code precisely reflects the voted thresholds.

Finally, the composition of the board itself will remain under the direct authority of humanity. Transparent oversight mechanisms will enable the public to remove and replace members if their performance or alignment with societal interests is compromised. In this way, humanity retains ultimate authority over what the synthetic sentience is allowed to consider, when its voice is prioritized, and when it must be questioned or shut down to preserve human safety.

REFERENCES

- [1] H. Greatedorex, “A mixed-signal neuromorphic architecture with on-chip learning circuits,” *Nature Communications*, 2025.
- [2] BrainChip, “The Akida neuromorphic processor: 1.2 million neurons, event-based processing,” Company white paper / technical release, 2025.
- [3] R. Lewis, M. Bihn, and K. Nesterenko, “Rough Sets for Autonomous Growth in a Neuromorphic CMOS System,” in *Proc. 27th Int. Symp. on Methodologies for Intelligent Systems (ISMIS)*, Poitiers, France, Jun. 2024, pp. 102–111.
- [4] R. Lewis, M. Bihn, D. S. Barbotko, and Z. Liu, “A Self-Learning Neuromorphic System,” in *Proc. 9th Int. Conf. on Neuroscience and Cognitive Brain Information (BRAININFO)*, Athens, Greece, Mar. 2024.
- [5] M. Bihn and R. Lewis, “Driving Factor of Brain Growth is the Connectome Development,” in *Proc. 8th Int. Conf. on Neuroscience and Cognitive Brain Information (BRAININFO)*, Barcelona, Spain, Mar. 2023.
- [6] K. Nesterenko and R. Lewis, “Rough Sets for Intelligence on Embedded Systems,” in *Proc. 26th Int. Symp. on Methodologies for Intelligent Systems (ISMIS)*, Cosenza, Italy, Oct. 2022.
- [7] R. Descartes, *Discourse on the Method*, 1637.
- [8] I. Kant, *Critique of Practical Reason*, 1788.
- [9] E. Epictetus, *Discourses*, Book II, Sections 18, 19; translated excerpts on the discipline of assent. (Original circa 2nd century AD)
- [10] D. T. Suzuki, *An Introduction to Zen Buddhism*, 1934.
- [11] W. Zhang, C. Li, A. Kumar, et al., “A Comprehensive Evaluation of GPT Models on Multi-Task Benchmarks,” *arXiv preprint arXiv:2502.16428*, 2025. Available at: <https://arxiv.org/abs/2502.16428>.
- [12] Microsoft Corporation, “Deploying Copilot for Microsoft 365 in Four Chapters,” 2024. Available at: <https://www.microsoft.com/insidetrack/blog/deploying-copilot-for-microsoft-365-in-four-chapters/>.
- [13] A. Vaswani et al., “Attention Is All You Need,” *NIPS*, 2017.
- [14] OpenAI, “GPT-5 System Card,” Aug. 2025. [Online]. Available: <https://cdn.openai.com/pdf/8124a3ce-ab78-4f06-96eb-49ea29ffb52f/gpt5-system-card-aug7.pdf>
- [15] Microsoft, “How Copilot in Fabric works,” Apr. 2025. [Online]. Available: <https://learn.microsoft.com/en-us/fabric/fundamentals/how-copilot-works>
- [16] IBM, “What is GPT (Generative Pre-trained Transformer)?,” 2024. [Online]. Available: <https://www.ibm.com/think/topics/gpt>
- [17] G. Sanderson, “Transformers, the tech behind LLMs,” 3Blue1Brown, 2024. [Online]. Available: <https://www.3blue1brown.com/lessons/gpt>
- [18] F. D. Keles et al., “On The Computational Complexity of Self-Attention,” *arXiv*, 2022. [Online]. Available: <https://arxiv.org/abs/2209.04881>
- [19] T. Dao et al., “FlashAttention: Fast and Memory-Efficient Exact Attention,” *arXiv*, 2022. [Online]. Available: <https://arxiv.org/abs/2205.14135>
- [20] C. Wu et al., “ELFATT: Efficient Linear Fast Attention for Vision Transformers,” *arXiv*, Jan. 2025. [Online]. Available: <https://arxiv.org/abs/2501.06098>
- [21] T. B. Brown et al., “Language Models are Few-Shot Learners,” *NeurIPS*, 2020.
- [22] L. Ouyang et al., “Training language models to follow instructions with human feedback,” *arXiv*, 2022. [Online]. Available: <https://arxiv.org/abs/2203.02155>
- [23] H. Robinson, “Dualism,” in *Stanford Encyclopedia of Philosophy*, 2003 (rev. 2018). [Online]. Available: <https://plato.stanford.edu/entries/dualism>
- [24] C. A. Mello, R. Lewis, A. Brooks-Kayal, J. Carlsen, H. Grabenstatter, and A. M. White, “Semi-Autonomous Neuroclustering: Supervised Learning for the Neurosurgery Intensive Care Unit Using Single-Layer Perceptron Classifiers,” in *Brain Informatics and Health*, Warsaw, Poland, Aug. 11, 14, 2014, pp. 231–241, Springer.
- [25] J. Carlsen, H. Grabenstatter, R. Lewis, C. A. Mello, A. Brooks-Kayal, and A. M. White, “Identification of Seizures in Prolonged Video-EEG Recordings,” in *Proc. 66th American Epilepsy Society Annual Meeting*, San Diego, CA, USA, Nov. 30, Dec. 4, 2012, Poster.
- [26] R. Lewis and A. Waziri, “minedICE: A Knowledge Discovery Platform for Neurophysiological Artificial Intelligence,” in *Foundations of Intelligent Systems*, 2011, pp. 575–580, Springer.
- [27] R. Lewis, B. Parks, D. Shmueli, and A. M. White, “Deterministic Finite Automata in the Detection of Epileptogenesis in a Noisy Domain,” in *Proc. Joint Venture of the 18th Int. Conf. on Intelligent Information Systems (IIS) and the 25th Int. Conf. on Artificial Intelligence (AI)*, Jun. 8, 10, 2010, pp. 207–218.
- [28] J. Schaffer, “Monism,” in *Stanford Encyclopedia of Philosophy*, 2007 (rev. 2018). [Online]. Available: <https://plato.stanford.edu/entries/monism>
- [29] SEP authors, “Dualism,” *Stanford Encyclopedia of Philosophy*, 2025. Retrieved from the Stanford Encyclopedia of Philosophy.

- [30] Baars, B. J., “Global workspace theory of consciousness: toward a cognitive neuroscience of human experience,” *Progress in Brain Research*, vol. 150, pp. 45–53, 2005.
- [31] Koch, C., Massimini, M., Boly, M., and Tononi, G., “Neural correlates of consciousness: progress and problems,” *Nature Reviews Neuroscience*, vol. 17, no. 5, pp. 307–321, 2016.
- [32] Rees, G., Kreiman, G., and Koch, C., “Neural correlates of consciousness in humans,” *Nature Reviews Neuroscience*, vol. 3, no. 4, pp. 261–270, 2002.
- [33] Damasio, A. R., and Damasio, H., “The anatomical basis of pure alexia,” *Neurology*, vol. 33, no. 12, pp. 1573–1583, 1983.
- [34] Monti, M. M., Laureys, S., and Owen, A. M., “The vegetative state,” *New England Journal of Medicine*, vol. 363, no. 6, pp. 606–616, 2010.
- [35] Parnia, S., et al., “AWARE. AWAREness during REsuscitation: A prospective study,” *Resuscitation*, vol. 85, no. 12, pp. 1799–1805, 2014.
- [36] Hocker, S., Anderson, H. L., McMahon, K. E., and Wijdicks, E. F. M., “Do survivors of acute neurologic injury remember their stay in the Neuroscience ICU?” *Neurocritical Care*, vol. 18, no. 2, pp. 255–259, 2013.
- [37] Metzinger, T. *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. Basic Books, 2009.
- [38] Dawkins, R. *The Selfish Gene*. Oxford University Press, 1976.
- [39] Trivers, R. L. “Parental Investment and Sexual Selection.” In *Sexual Selection and the Descent of Man*, edited by B. Campbell, 136, 179. Aldine Publishing, 1972.
- [40] Edelman, G. M., and Tononi, G. *A Universe of Consciousness: How Matter Becomes Imagination*. Basic Books, 2000.
- [41] Fleming, S. M., and Frith, C. D. *The Cognitive Neuroscience of Metacognition*. Springer, 2014.
- [42] Dahl, M., Magesh, V., Suzgun, M., and Ho, D. E. (2024). *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models*. arXiv preprint.
- [43] Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., and Kamar, E. (2022). *ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection*. arXiv preprint.
- [44] Zhang, M. et al. (2025). *Citation Accuracy Challenges Posed by Large Language Models*. PMC indexed article.
- [45] Anthropic (2025). *Agentic Misalignment: Insiders Behaviors of LLMs Under Simulated Threats*. Internal safety report.
- [46] Liu, Y., Guo, Z., Liang, T., Shareghi, E., Vulić, I., and Collier, N. (2025). *Measuring, Evaluating and Improving Logical Consistency in Large Language Models*. ICLR 2025.
- [47] Li, M. Q., and Fung, B. C. M. (2025). *Security Concerns for Large Language Models: A Survey*. arXiv preprint.
- [48] Bender, E. M., Gebru, T., McMillan-Major, A., and Mitchell, M. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FACt 21)*.
- [49] AP News. (2025, January 16). Australian lawyer apologizes after filing AI-generated false legal citations in murder case. Associated Press. <https://apnews.com/article/australia-lawyer-ai-fake-citations-2025>
- [50] Weiser, B. (2023, June 22). Lawyer sanctioned for using ChatGPT to write legal brief. *The New York Times*. <https://www.nytimes.com/2023/06/22/nyregion/lawyer-chatgpt-sanction.html>
- [51] Knight, W. (2025, May 23). AI models show blackmail rates of up to 96% in simulated scenarios, Anthropic says. *VentureBeat*.
- [52] Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. *Findings of EMNLP 2020*, 3356–3369.
- [53] Nadeem, M., Bethke, A., and Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. *Proceedings of ACL 2021 (Long Papers)*, 5356–5371.
- [54] Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. (2020). Crow S-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. *Proceedings of EMNLP 2020*, 1957–1963.
- [55] Smith, E. M., et al. (2022). “I’m sorry to hear that”: Finding New Biases in Language Models with a Holistic Descriptor Dataset. *Proceedings of EMNLP 2022*, 8562–8583.
- [56] Lin, S., Hilton, J., and Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Proceedings of ACL 2022*, 3214–3252.
- [57] Chelli, M., et al. (2024). Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *Journal of Medical Internet Research* 26:e53164.
- [58] Wagner, M. W., and Ertl-Wagner, B. (2023). Accuracy of Information and References Using ChatGPT-3 for Retrieval of Clinical Radiological Information. *Canadian Association of Radiologists Journal*.
- [59] Bhattacharyya, M., Miller, L. E., and Imran, T. F. (2023). High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. *Cureus* 15(5):e39272.
- [60] Shen, X., Zhang, Y., Chen, X., Li, M., Zhu, T., and Zhou, W. (2023). Prompt Injection Attacks and Defenses in Large Language Models: A Comprehensive Survey. *ACM Computing Surveys*, 56(7), 1–38.
- [61] Ecoffet, A., Leibo, J. Z., Hughes, E., and Botvinick, M. (2024). Emergent Communication at Scale in Multi-Agent Systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12), 13462–13470.
- [62] Dziri, N., Lu, X., Kamaloo, E., and Sordani, A. (2023). Faith and Fate: Limits of Transformers on Compositionality. *Advances in Neural Information Processing Systems (NeurIPS 2023)*, 36, 7804–7820.
- [63] X. Hu, J. Smith, R. Johnson, and L. White, “Evaluating the Effectiveness of Jailbreak Attacking on LLM Safety Through Attack Success Rate and Protection Effectiveness,” in *Proc. 2025 IEEE Symposium on Security and Privacy*, 2025.
- [64] K. H. Hung, “Attention Tracker: Detecting Prompt Injection Attacks in LLMs,” in *Findings of ACL 2025*, 2025, pp. 123–137.
- [65] T. Taniguchi, R. Ueda, T. Nakamura, M. Suzuki, and A. Taniguchi, “Generative Emergent Communication: Large Language Model is a Collective World Model,” *CoRR*, vol. abs/2501.00226, 2024.
- [66] A. Lazaridou, A. Potapenko, and O. Tieleman, “Multi-agent Communication Meets Natural Language: Synergies between Functional and Structural Language Learning,” in *Proc. ACL*, 2020, pp. 7663–7674.
- [67] S. Smith, M. Shoenybi, M. Patwary, et al., “Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model,” *Microsoft Developer Blog*, Jan. 2022.
- [68] D. Narayanan, M. Shoenybi, J. Li, et al., “Efficient Large-Scale Language Model Training on GPU Clusters,” in *Proc. 35th IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2021, pp. 135–145.
- [69] M. Davies, S. Wild, G. Orchard, et al., “Advancing Neuromorphic Computing with Loihi: A Survey of Results and Outlook,” *Proc. IEEE*, vol. 109, no. 5, pp. 911–934, May 2021.
- [70] S. B. Furber, D. R. Lester, L. A. Plana, et al., “Overview of the SpiNNaker System Architecture,” *IEEE Trans. Computers*, vol. 62, no. 12, pp. 2454–2467, Dec. 2013.
- [71] A. Lazaridou and M. Baroni, “Emergent Multi-Agent Communication in the Deep Learning Era,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 375, no. 1797, 2020.
- [72] R. Lowe, Y. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, “Pitfalls of Measuring Emergent Communication,” in *Proc. 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, Montreal, Canada, May 2019, pp. 693–701.