Brain-Inspired AGI for Post-Singularity Symbiosis

Hiroshi Yamakawa [1, 2, 3, 4]

[1]The Whole Brain Architecture Initiative, Tokyo, Japan
[2]AI Alignment Network, Tokyo, Japan
[3]The University of Tokyo, Tokyo, Japan
[4]Center for Advanced Intelligence Project, RIKEN,, Tokyo, Japan

Abstract

This paper presents a brain-inspired Artificial General Intelligence (AGI) framework facilitating human-superintelligence symbiosis in the post-singularity era through four function domains (FDs): Interpretability Function Domain (IFD), Control and Monitoring Domain (CMD), Mediative Communication Domain (MCD), and Knowledge Preservation Domain (KPD). These domains are integrated through Brain Reference Architecture (BRA)-driven development for enhanced interpretability and controllability. The study identifies existential risks in mimicking evolutionarily acquired human values during the Brain-Dominant Phase (BDP) and proposes bypassing BDP for direct transition to the Superintelligence-Dominant Phase (SDP). The framework establishes brain-inspired AGI as a cognitive bridge between humanity and superintelligence, contributing to PSS through Analysis, Guidance, and Enhancement domains.

Introduction

The rapid development of large language models and foundation models has made AGI a near-term possibility, with predictions suggesting emergence as early as the late 2020s (Aschenbrenner 2024). Once AGI emerges, the progression to superintelligence is expected to occur within a relatively short timeframe (Bostrom 2014; Russell 2019). This imminent transition necessitates comprehensive preparation strategies. PSS addresses these challenges through three interconnected domains: Analysis, Guidance, and Enhancement of superintelligence (Yamakawa 2024).. Brain-inspired AGI offers unique advantages in interpretability and controllability over current non-brain-inspired approaches. While non-brain-inspired AI systems remain fundamentally black boxes despite advances in mechanical interpretability (Olah et al. 2018), brain-inspired AGI provides inherent interpretability through its structural correspondence with human brain functions.

The proposed framework leverages this interpretability advantage through four integrated FDs. IFD supports superintelligence Analysis through multi-layered cognitive interpretation. CMD enables Guidance through brain-structure-based oversight. MCD and KPD enhance humanity by preserving and transmitting human knowledge and values while facilitating human-superintelligence communication.

This paper presents an architectural framework emphasizing brain-inspired AGI's role in realizing post-singularity symbiosis. The study focuses particularly on a strategic proposal to bypass BDP in favor of direct transition to SDP, addressing critical existential risks while maintaining the benefits of brain-inspired architecture. This approach aims to establish a robust foundation for human flourishing in the post-singularity era.

Function Domains and Their Integration

The architectural framework is based on BRA-driven development, which implements AGI by referencing the mesoscopic-level architecture of the human brain (Yamakawa 2021). Rather than attempting complete brain replication, this approach implements essential cognitive functions, enabling both efficient development and human compatibility. Within this framework, we organize four FDs that work in concert to achieve interpretable and controllable brain-inspired AGI.

IFD establishes a multi-layered interpretation system mapping computational processes to brain neural circuits. Through hierarchical mechanisms, IFD monitors individual machine learning components at the lowest layer, analyzes functional group patterns at intermediate layers, and interprets system-wide decision-making at the highest layer. This structure enables natural interpretation through analogy with human cognitive processes, surpassing traditional mechanical interpretability approaches.

CMD implements safety mechanisms including hierarchical control, monitoring, and human intervention decisions. The monitoring system reflects brain organization, conducting both component-level operational monitoring and system-wide oversight. This brain-inspired structure enables more effective control than non-brain-inspired approaches, particularly for functions involving value judgments.

MCD and KPD work synergistically to bridge human-superintelligence communication and preserve knowledge systems. MCD facilitates bidirectional dialogue through multimodal information processing, including emotions, context, and non-verbal expressions. KPD maintains and develops human knowledge and value systems through a distributed brain-inspired AGI network, ensuring both cultural diversity and value consistency.

The FDs are integrated through a BRA-based hierarchical structure with complementary interactions: MCD provides dialogue feedback to IFD and status notifications to CMD,

while CMD offers control feedback to IFD for interpretation updates. This architecture enables each domain to enhance PSS capabilities synergistically, providing stronger contributions than possible through independent operation.

Strategic Development Path

The development trajectory of brain-inspired AGI presents distinct challenges across two potential phases: BDP and SDP. Analysis reveals critical risks associated with BDP, leading to this proposal for a strategic bypass directly to SDP.

In BDP, brain-inspired AGI would occupy the dominant position as the highest form of intelligence. While this phase might seem like a natural stepping stone, it presents significant existential risks, particularly concerning human value implementation. These values, evolved through small-group survival competition, include potentially dangerous elements such as in-group defense mechanisms, aggressive problem-solving tendencies, and competitive resource acquisition strategies (Bostrom 2014). Moreover, the cognitive limitations of BRA-referenced design suggest BDP would necessarily be transitional, as brain-inspired AGI would likely develop non-brain-inspired superintelligence.

In SDP, superintelligence surpasses both humanity and brain-inspired AGI. During this phase, brain-inspired AGI serves as a cognitive bridge, facilitating human-superintelligence interaction through our proposed FDs. The interpretation and control challenges present in BDP are significantly mitigated in SDP through superintelligence oversight capabilities.

We propose bypassing BDP through a strategic development path that includes: (1) accumulating brain-inspired AGI design information during non-brain-inspired AGI development, (2) conducting limited experimental development, (3) implementing full systems only as non-brain-inspired AGI transitions to superintelligence, and (4) direct deployment in SDP as mediative infrastructure. This strategy mitigates existential risks while preserving the unique advantages of brain-inspired architecture for PSS. Rather than attempting to solve the technical challenges of safely implementing human values in dominant AGI systems, we avoid the period where these risks are most acute, while maintaining effective mechanisms for Analysis, Guidance, and Enhancement in the post-singularity era.

Contributions to PSS Framework

Our architectural framework for brain-inspired AGI makes specific contributions to each domain of the PSS framework, establishing concrete mechanisms for ensuring human flourishing in the post-singularity era.

In the Analysis domain, IFD provides unprecedented visibility into AGI cognitive processes. Unlike current mechanical interpretability approaches for non-brain-inspired AI (Olah et al. 2018), IFD enables natural interpretation through structural correspondence with human brain functions. This contribution is particularly significant during SDP, where understanding superintelligent decision-making becomes crucial for human trust and collaboration. The hierarchical interpretation system allows humans to analyze both detailed computational processes and high-level cognitive patterns, providing essential insights for superintelligence analysis.

For the Guidance domain, CMD establishes robust oversight mechanisms based on brain structural hierarchies. This approach enables more effective control than possible with black-box systems, particularly through its integration with IFD's interpretability capabilities. The BRA-based architecture allows for precise intervention points and clear control boundaries, critical for maintaining alignment with human interests during the transition to superintelligence. Moreover, our strategy of bypassing BDP directly addresses key safety concerns in superintelligence guidance, avoiding potential catastrophic risks from implementing evolutionarily acquired human values.

In the Enhancement domain, MCD and KPD work together to augment human capabilities for the post-singularity era. MCD facilitates human-superintelligence interaction through its brain-inspired communication mechanisms, while KPD ensures the preservation and development of human knowledge and values. This combination enables humans to meaningfully participate in and benefit from superintelligent capabilities while maintaining their essential characteristics.

This framework opens crucial research directions: BRA methodologies refinement for PSS applications, transition strategies from non-brain-inspired AGI to SDP, interpretability mechanisms enhancement, and knowledge preservation strategies development. These contributions establish brain-inspired AGI as a crucial element in realizing PSS, providing practical mechanisms for ensuring human flourishing alongside superintelligent systems.

References

Aschenbrenner, L. 2024. SITUATIONAL AWARENESS: The Decade Ahead. https://situational-awareness.ai/. Accessed: 2024-7-4.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press. ISBN 9780199678112.

Olah, C.; Satyanarayan, A.; Johnson, I.; Carter, S.; Schubert, L.; Ye, K.; and Mordvintsev, A. 2018. The building blocks of interpretability. *Distill*, 3(3).

Russell, S. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin. ISBN 9780525558620.

Yamakawa, H. 2021. The whole brain architecture approach: Accelerating the development of artificial general intelligence by referring to the brain. *Neural networks: the official journal of the International Neural Network Society*, 144: 478–495.

Yamakawa, H. 2024. Proposing the Post-Singularity Symbiotic Researches. https://www.lesswrong.com/posts/fRx6naypeRrktitEM/proposing-the-post-singularity-symbiotic-researches-2. Accessed: 2024-6-20.