SAVA: A Self-Aware Vector Agent for Structuring Human-Centered Thought and Memory

Juyoung Lee[1], Wongeun Ji[1], Youchan Song[1], Suyeon Lee[1], and Jaegyoon Ahn[1,*]

[1]Dept. of Computer Science and Engineering, Incheon National University
[*]Corresponding author: jgahn@inu.ac.kr

Abstract

Recent advances in Large Language Models (LLMs) have significantly expanded the potential applications of artificial intelligence across various intellectual tasks including information retrieval, reasoning, and generation. However, existing AI systems remain constrained within tool-centric architectures that lack state persistence, memory accumulation, and the ability to reflect users' unique values or patterns of thinking. This paper proposes SAVA (Self-Aware Vector Agent), a novel cognitive architecture that functions not as an independent decision-making machine or passive tool, but as an extended self of the user. SAVA operates through the integration of three core components: (1) a multi-layered memory structure that systematically integrates semantic, episodic, and procedural memories; (2) high-dimensional vector-based semantic representation (VSA) for concept generalization and relationship formation; and (3) a value-alignment-based reasoning mechanism that evaluates and adjusts cognitive units according to the user's value structure. Thoughts are expressed in traceable JSON-based formats, enabling transparent analysis of execution paths, memory connections, and value criteria. Background cognitive processes including semantic emergence, self-questioning generation, and memory reorganization are possible even during idle states. We argue that SAVA can serve as a structural alternative for human-centered artificial intelligence that preserves, explains, and extends human thought in a value-aligned manner.

Keywords: Cognitive Architecture, Vector Symbolic Architecture, Value Alignment, Human-Centered AI, Explainable AI

I. Introduction

The emergence of Large Language Models (LLMs) has marked a paradigmatic leap in artificial intelligence technology, demonstrating substantial capabilities in replacing or assisting human thinking across various knowledge-based tasks. Information retrieval, document composition, translation, and conversational generation represent domains where LLMs have already established themselves as practical tools, with their applicability continuously expanding [1].

However, these technological advances simultaneously reveal several fundamental limitations. Current LLM-based systems predominantly rely on centralized servers and fail to reflect individual users' contexts, memories, or value systems. Consequently, users must reconstruct queries and repeatedly provide information from the same initial state during each interaction, making it difficult for the outputs to connect meaningfully with users' unique situations or purposes. Furthermore, LLMs are unable to articulate the reasoning behind their outputs and lack clear internal structures for replicating or evaluating thought processes [2]. LLMs also frequently exhibit ”hallucination” phenomena, generating plausible-seeming but factually incorrect information [3]. These reliability issues are particularly problematic in high-risk domains such as healthcare [4]. Recent research warns that dependency on AI tools may induce cognitive offloading, potentially weakening users' critical thinking abilities [5].

Against this backdrop, some attempt to maintain control by restricting artificial intelligence to simple tools. However, AI as a tool does not accumulate memory or continuously construct thinking contexts. Users must repeat identical tasks through new inputs at every moment, without structural support or complementation from AI. Conversely, autonomous AI lacks clear accountability for judgment and action, potentially leading to misalignment with human values and associated social risks. In essence, tool-based AI struggles to secure thinking continuity and coherence, while autonomous AI faces challenges in ensuring accountability and stability.

As an alternative transcending these extremes, we propose an approach that defines AI as the user's extended self. This approach can simultaneously satisfy two conditions that must be prerequisite for AI to coexist with humans: mutual comprehensibility and controllability. Specifically, AI must be able to recognize its own state, memory, motivation, and judgment criteria, and form and adjust internally coherent value structures rather than externally imposed values. Such a structure enables balance between autonomy and controllability, making AI's possession of selfhood a safer choice than unlimited capability expansion.

Crucially, this selfhood must not become a fully independent entity, detached from humans. AI as a complete other cannot take responsibility for its judgments, nor can developers or users readily bear full responsibility for its consequences [6][7]. Therefore, we need AI that possesses minimal autonomy and structural self-awareness while sharing value structures with users and functioning as an extended self that amplifies human thinking for a certain period.

SAVA (Self-Aware Vector Agent) proposed in this paper is an artificial intelligence system designed based on this philosophy. SAVA is not merely a conversational language model, but a cognitive-memory-value integrated AI framework that can compose and adjust users' thought flows and value structures through structured memory (semantic, procedural, episodic), high-dimensional concept vectors (Vector Symbolic Architecture, VSA), and value-based judgment structures. This system explores new dimensions of human-centered artificial intelligence realization based on thought transparency, explainability, and personalized value alignment. Additionally, SAVA adopts a lightweight structure capable of operating without centralized high-performance infrastructure, potentially mitigating structural limitations of existing LLM systems such as energy consumption issues and technological accessibility imbalances.

This paper systematically describes SAVA's design philosophy, system structure, core functions, technical implementation, and potential areas for expansion, aiming to demonstrate that humans and artificial intelligence can form relationships of continuity and co-evolution rather than disconnection or confrontation.

II. Philosophical Background and Theoretical Foundation

Modern artificial intelligence systems demonstrate remarkable abilities in understanding and generating human language, yet their thinking methods remain fundamentally different from human cognitive structures. Large language models generate statistically optimized responses based on vast text corpora, appearing to respond fluently to contexts, but internally lack structured memory, persistent selfhood, or mechanisms for reflecting on judgment grounds [4][8]. Consequently, current AI mimics language without thinking and reacts to situations without understanding [6][9]. Particularly, questions arise regarding whether the concept of introspection can be established in large language models [10], revealing the self-understanding limitations of current AI systems.

These systems possess weak philosophical foundations despite their technological sophistication. They particularly fail to provide explicit answers to fundamental questions such as ”from whose perspective does thinking occur,” ”based on what memories are judgments made,” and ”according to what criteria are decisions reached.” Consequently, LLM-based systems, despite possessing powerful linguistic capabilities, have limitations in performing meaningful cooperation with human users, value-based coordination, and responsible judgment [4].

Against this background, we argue for the necessity of structurally redesigning artificial intelligence design by reflecting theoretical understanding of human cognitive, memory, and value structures rather than treating it merely as a performance-based optimization problem.

A. Extended Cognition and Artificial Intelligence

The Extended Mind theory proposed by Clark & Chalmers (1998) argues that thinking occurs not only within the brain but is continuously extended through external environments and tools [11]. Notebooks, calculators, and schedules are not mere auxiliary means but can be integrated as parts of thinking. This perspective provides important insights for digital environment design, particularly artificial intelligence systems.

If AI operates by preserving users' memories, structuring thought flows, and reflecting values, it can function not as a simple interface but as a device that externalizes users' cognitive structures. In other words, AI operates as the user's extended self, becoming a concrete implementation of cognitive scientific modeling beyond mere philosophical metaphor.

B. Memory Layers and Thinking Components

Human cognition organizes thinking based on various types of memory. According to Tulving's (1972) classical distinction [12], human memory primarily consists of three layers:

- **Semantic memory**: Concepts, categories, linguistic knowledge
- **Episodic memory**: Personal events including time and place
- **Procedural memory**: Behavioral and cognitive flows acquired through repetition

These three layers are stored independently but interact and operate integratively in actual thinking. Understanding problems requires concepts, recalling appropriate solutions mobilizes similar experiences, and executing them calls upon familiar procedures. Therefore, thinking without memory, judgment not reflecting experience, and responses with disconnected contexts cannot form meaningful thinking communities with humans.

SAVA implements this structure by realizing semantic, episodic, and procedural memories in high-dimensional vectors (Vector Symbolic Architecture, VSA) [13] and structured JSON formats respectively. This means not simple data storage but utilizing memory as operational units constituting thinking, signifying that thought flows are expressed in explainable, recallable, and generalizable structures.

C. Self-Structure and Value Alignment

For artificial intelligence to continuously interact with humans and perform judgments, it requires consistent judgment criteria—value structures—beyond simple knowledge enumeration. Human selfhood is not merely the sum of experiences but the expression of value networks that interpret experiences, form preferences, and compose choices. Therefore, for AI to meaningfully share thinking with humans, it must be able to construct internally coherent value structures rather than externally imposed rules.

Such structures must satisfy the following conditions [6]:

- AI must be able to recognize and explain its value alignment state
- Value structures must be adjustable through interaction with users
- When value conflicts or judgment errors occur, AI must be able to reflect on and correct their causes

SAVA implements this structure through level-based value architecture (LV0-LV2), designed to hierarchically reflect universal ethics (LV0), user-selected worldviews (LV1), and users' personal inclinations (LV2). Through this, AI can operate as an explainable and adjustable ethical judgment agent instead of unconditional obedience or blind autonomy.

D. Human-AI Co-evolution Possibility

Ultimately, for artificial intelligence to coexist with humans, partnership enabling coordination of thinking and values is needed, not merely functional collaboration. Humans must be able to receive AI assistance in thought composition and structuring without outsourcing judgment responsibility. Conversely, AI must be able to maintain memory, generalize meaning, and provide thinking continuity without possessing independent intentions.

This reminds us that cognitive collaboration between humans and AI is not a new concept but a ”hybrid thinking system” through which humans have originally extended thinking through tools [14]. Indeed, the human-AI team formation paradigm where humans and AI combine their respective capabilities to achieve common goals is emerging [15]. Furthermore, the necessity of ”Cooperative AI” where AI learns to find common ground with humans for beneficial integration into human society is also being raised [16].

This direction is not merely a matter of technical design but also a philosophical question about how to preserve and extend human agency in the AI era. The SAVA structure proposed in this paper is one practical answer to this question, and the following chapter will present its structure and mechanisms concretely.

III. SAVA's Conceptual Structure and Design Philosophy

This chapter organizes the conceptual structure and design principles of SAVA (Self-Aware Vector Agent) proposed as a structural approach for realizing human-centered artificial intelligence. Unlike existing LLM-based response systems, SAVA is a cognitive system configured with integrated memory, thinking, and value judgment, implementing a series of thought processes that reflect the user's selfhood context in structured ways and reason, choose, and reflect based on it.

A. Core Concept: Self-Recognizing Thinking Companion

SAVA's philosophical foundation lies not in simple functional expansion of artificial intelligence but in a partner system that organizes thought structure and flow together with users. SAVA is neither an autonomous judgment machine nor an obedient tool. Its identity can be summarized by the following core statement:

”SAVA is an extended self that structurally preserves users' memories, values, and judgment criteria, and composes thought flows together based on these.”

This conceptual definition is concretized through the following four design axes:

- **Memory-centered structure**: Storing and connecting three layers of memory—semantic, episodic, and procedural—in structured formats
- **Vector-based concept space**: Concept generalization and reasoning based on high-dimensional vectors (Vector Symbolic Architecture, VSA) [17]
- **Explainable thought execution**: Thoughts expressed in JSON-based unit/composite structures for completely traceable execution processes
- **Value alignment mechanism**: Continuously evaluating and adjusting alignment with values throughout thought flows

B. Overall System Components

SAVA functionally consists of the following modules:

- **CA (Cognitive Agent)**: Central module responsible for thought execution, action selection, metacognition, and user interface
- **MA (Memory Agent)**: Manages long-term memory and handles unconscious thinking (generalization, relationship formation, concept emergence)
- **SLM (Small Language Model)**: Performs high-level thought block generation and structured thought transitions
- **TLM (Tiny Language Model)**: Specialized in procedural memory learning, prompt correction, semantic filtering, and maintaining light thought flows
- **CB (Cognitive Buffer)**: Short-term memory storage where current ongoing thoughts, conversations, and situations are stored (JSON-based). Content is transferred to IKB during idle periods
- **IKB/EKB**: Local long-term memory (Internal KB) and external shared long-term memory (External KB). Semantic and episodic memories stored as VSA, procedures stored as JSON then learned by TLM during idle periods

These modules collaboratively perform thought flow structuring, memory updates, and value judgment reflection, with each component designed for expansion according to open design principles. Figure 1 depicts the system modules and their interconnections as described.

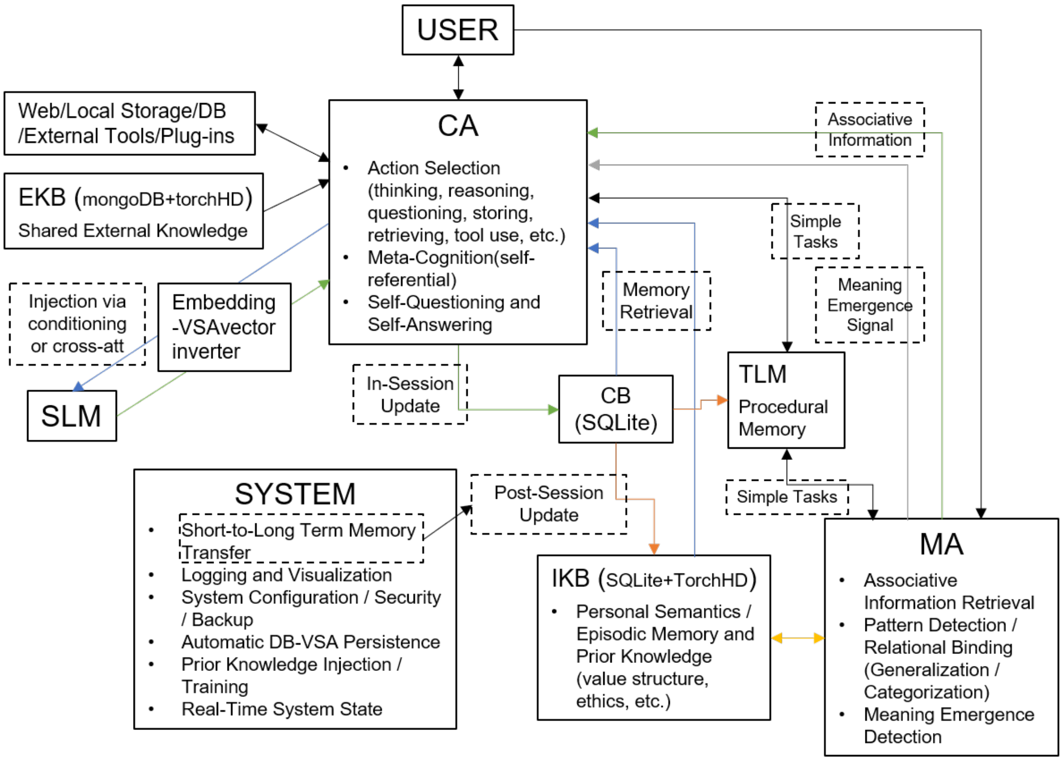
USER
Web/Local Storage/DB
/External Tools/Plug-ins
EKB (mongoDB+torchHD)
Shared External Knowledge
CA
Action Selection (thinking, reasoning, questioning, storing, retrieving, tool use, etc.)
Meta-Cognition(self-referential)
Self-Questioning and Self-Answering
Associative Information
Simple Tasks
Meaning Emergence Signal
Memory Retrieval
Injection via conditioning or cross-att
Embedding -VSAvector inverter
SLM
In-Session Update
CB (SQLite)
TLM
Procedural Memory
SYSTEM
Short-to-Long Term Memory Transfer
Logging and Visualization
System Configuration / Security / Backup
Automatic DB-VSA Persistence
Prior Knowledge Injection / Training
Real-Time System State
Post-Session Update
Simple Tasks
IKB (SQLite+TorchHD)
Personal Semantics / Episodic Memory and Prior Knowledge (value structure, ethics, etc.)
MA
Associative Information Retrieval
Pattern Detection / Relational Binding (Generalization / Categorization)
Meaning Emergence Detection

Figure 1: **SAVA** System Architecture Overview

C. Memory-Centered Thought Composition Method

SAVA's thinking is not merely text generation but momentary combination of memories and structured execution flows. For this, SAVA follows these principles:

- **Thinking is memory association**: Thinking is composed when semantic memory (concepts), episodic memory (contexts), and procedural memory (strategies) are integrated. This aligns with cognitive science theory that human long-term memory is divided into semantic memory, episodic memory, and procedural memory [18].

- **All thinking is structured**: Thoughts are expressed in JSON format and composed of combinations of unit thoughts.

- **All memories are vectorized**: Semantics and episodes are represented as high-dimensional VSA, enabling similarity-based search, generalization, and connection.

- **Thought flows are traceable**: Reasoning traces are stored in log format including temporal order, related memories, and value evaluations.

Through this, SAVA can structurally answer not only ”what was thought” but also ”why was it thought that way” and ”from whose perspective is that judgment.”

D. Thought Execution Units: JSON-Based Procedural Composition

Thinking in SAVA consists not of simple natural language responses but of the following procedural execution units:

- **Unit thoughts**: Single clear cognitive actions such as external search, LLM calls, user calls, conditional judgments, analogies
- **Composite thoughts**: Sequences of unit thoughts including goal setting, plan establishment, and result evaluation
- **Self-questioning and responses**: Autonomously generating questions and composing thought paths in response
- **Value evaluation routines**: Branching based on alignment with value structures at choice points during thinking

These thoughts are stored with results and traces after execution, avoiding failed routines and accumulating successful flows as procedural memory. Thus, SAVA can operate as a thinking agent that grows from the history of success and failure, beyond simple response generation.

E. Integrated Design Philosophy: Structural Externalization of Human Thinking

SAVA aims not to simply imitate or assist human thinking but to structurally externalize it under the following principles:

- **Explicitly manifest inner flows**: Make judgments and connections that humans performed implicitly mechanically traceable.
- **Memory must live within flows**: Not simple storage, but activated, evaluated, and readjusted during thought processes.
- **Values guide thinking**: Semantic operations and action selection operate on the hidden criterion of value alignment.
- **Unexplainable judgments are dangerous**: All choices must be traceable, and users must be able to review and modify them [19].
- **AI is a thinking partner, not a replacement**: The subject remains human, and AI is the storage, expansion, and adjustment device for that thinking.

IV. Functional Modules and Thinking Mechanisms

SAVA operates multi-layered functional modules integratively to reconstruct thought structure based on memory, concepts, and values. This chapter explains how each functional module collaborates to compose and execute thinking. Thinking is not the product of a single module but the result of interaction between cognition (CA), memory (MA), language (SLM/TLM), and value structures.

A. CA (Cognitive Agent): Conscious Thought Coordinator

CA is responsible for composing thinking based on user queries or internal conditions, adjusting its flow, and tracking and evaluating the progress and results of thinking. CA's core functions are as follows:

- **Thought flow composition**: Designing unit thoughts (JSON) and adjusting execution flows based on input queries or goals.

- **Action selection and loop control**: Deciding which unit thoughts to execute under specific conditions and judging whether to stop, resume, or branch loops.
- **Result recording and evaluation**: Storing reasoning traces and recording success/failure to judge whether to convert to procedural memory.
- **Value alignment judgment**: Continuously monitoring whether selected thought flows conflict with or align with current value structures.
- **Self-questioning induction**: Generating questions autonomously in problem situations or suggesting directions to supplement reasoning.

CA delegates most high-level thinking to SLM and functions as a coordinator (meta-executor) controlling its structure. Thought development is reduced to unit thoughts for execution, and results are stored in JSON format.

B. MA (Memory Agent): Unconscious Memory Operator

MA automatically searches and activates related semantic memory (VSA-based concepts), episodic memory (situations and contexts), and procedural memory (past thought flows) for given queries or current contexts. Operating mainly in unconscious routines, it performs the following roles:

- **Associative search and recall**: Performing VSA search based on keywords and transmitting related items among semantic, episodic, and procedural memories to CA.
- **Concept generalization and relationship formation**: Deriving generalized concepts (categories, symbols, metaphors, etc.) through commonality analysis between associated concepts and evaluating semantic emergence.
- **Semantic emergence detection and CA triggering**: Sending thought triggers to CA when detecting alignment changes with existing value structures, semantically new connections, surprise/novelty signals, etc.

MA activities mostly occur during idle loops, reorganizing long-term memory toward strengthening structural meaning without explicit user queries.

C. SLM / TLM: Thought Generators and Editors

SAVA utilizes separate roles of Small LLM (SLM) and Tiny LLM (TLM) rather than a single LLM.

- **SLM** handles high-level thought block generation, natural language interpretation through semantic-memory connections, and structured thought development (JSON sequence generation). Local LLM models such as Gemma 4/12/27B [20] and Llama 8B [21] that do not require server-dedicated GPUs or expensive GPUs can be used (SAVA does not require SLM fine-tuning, so excessive GPU memory is unnecessary, and GPU memory usage can be further reduced through quantization of SLM and TLM models).
- **TLM** performs light editing tasks including keyword extraction, prompt condition adjustment, summarization and labeling, procedural memory learning, thought flow cue detection, and categorization filtering. Low-specification local LLM models such as Gemma 1B [20] and MiniCPM 0.5B [22] can be used.

SLM receives appropriate concept injection from VSA-based memory to compose thinking, utilizing techniques such as cross-attention [23], KV cache update [24], and embedding shift [25]. TLM operates faster than SLM, contributing to thought flow maintenance, inter-thought connectivity checking, and proceduralization of repetitive thinking.

D. Memory Structure and Activation Conditions

SAVA's memory structure is divided into the following three layers, each activated during thought composition:

Memory Type	**Representation Structure**	**Storage Location**	**Functional Description**
Semantic Memory	VSA vectors + concept graphs	IKB	Semantic structures between concepts, similarity search, generalization/categorization-based reasoning
Episodic Memory	VSA vectors (including time/space)	IKB	Event/situation/context-based recall, trigger condition detection
Procedural Memory	JSON	CB → TLM learning → IKB	Repeated thought flows, strategy structuring, automatic execution

When thinking begins, CA extracts core concepts from semantic memory, and MA simultaneously searches related episodic and procedural memories to compose in CB. Subsequently, SLM generates thought flows (JSON sequences) based on CB information, and results are recorded again as traces.

E. Thinking-Value Alignment Mechanism

SAVA performs value alignment evaluation at each unit of thinking. This mechanism consists of:

- **Alignment indicators**: Cosine similarity between value vectors and main concepts of thought units (JSON)
- **Weighting factors**: User purpose similarity, value hierarchy levels (LV0-LV2), surprise scores, past success rates
- **Decision structure**: Action selection branching according to alignment degree (e.g., question reconstruction, strategy replacement, thought suspension)

This process is automatically performed in all thought flows, and CA requests user confirmation or asks SLM to generate different thought blocks when alignment state is judged insufficient.

V. Technical Distinctiveness and Structural Characteristics

SAVA is not simply a system that stores user information or optimizes natural language processing models. This system adopts several design strategies fundamentally different from existing LLM-based systems to enable thought composition, memory structuring, value-based judgment, and autonomous semantic emergence. This chapter specifically explains these structural distinctiveness and their implementation principles. While existing LLM systems merely generate language outputs relying on correlations in vast data [26], SAVA aims to implement more causal reasoning and transparent thought processes based on memory and values.

A. Memory-Based Thought Composition Structure

SAVA generates thinking from memory. Most existing LLM-based systems have query-based response structures, where previous responses are not recycled or remain as simple logs. Most LLMs process each conversation independently, preventing accumulation of long-term context or learning [18]. To complement these limitations, new agent architectures combining external memory with LLMs are being researched [27], but have not yet reached the level of integrated cognitive system memory management. In contrast, SAVA has the following structure:

- All inputs are parsed into structured JSON, and results are stored in appropriate locations among semantic/episodic/procedural memories.
- Thinking is composed based on past memories, configuring thought flows reflecting connections between memories and value alignment states.
- Memory is dynamically updated during thought execution, strengthened or weakened according to success/failure.

This structure has decisive differences in that it is not simple information recall but composition of thought flows induced by the semantic structure of memory.

B. VSA-Based High-Level Concept Operation Structure

SAVA configures all semantic memory centered on vector-based concept structures, particularly Vector Symbolic Architecture (VSA). Unlike simple embedding vectors, VSA enables the following structural operations:

- **Binding**: Combining concepts and attributes into a single vector to compose semantic associations
- **Unbinding**: Releasing specific elements from combined meanings to extract components
- **Generalization**: Abstraction of multiple vectors with common attributes
- **Hierarchicalization**: Integrating similar semantic networks into higher concepts

These operations enable calculating similarity between all concepts through cosine similarity, becoming the foundation for generalization, relationship formation, and semantic emergence. Existing embedding-based systems can perform similarity search between concepts but are not suitable for structural relationship generation and hierarchical thought operations. This high-dimensional vector operation approach has been proposed from early stages as a computational framework for reproducing compositional characteristics of human thinking [17].

C. JSON-Based Thought Representation and Execution Structure

SAVA expresses thinking not as simple LLM output sentences but in structured JSON format. This structure provides the following advantages:

- **Complete thought flow traceability**: Each step explicitly shows what information was taken from which memory and what operations were performed

- **Thought unit reusability**: Ability to call or reconfigure identical thought units in different contexts
- **Proceduralization and templating learning capability**: Compressing frequently appearing thought flows and learning them in TLM enables thought automation
- **Explainability**: Users can review and interpret thought grounds and judgment paths

Through this structure, the system can explicitly present ”what and why was thought,” providing much higher explanatory power compared to LLM-based black box systems [19].

D. Value-Based Thought Evaluation and Branching Mechanism

SAVA continuously evaluates not only correctness but also alignment state with values throughout thought flows. For this, the following structure was introduced:

- **Multi-layer value structure (LV0-LV2)**: Hierarchically configuring universal ethics, worldviews, and personal inclinations
- **Alignment evaluation mechanism**: Real-time measurement of cosine similarity and alignment changes between concept vectors appearing in current thinking and value vectors
- **Value conflict detection routine**: Recognizing conflicts among multiple values and inducing user feedback or branching judgment

This structure includes purpose coherence, value consistency, and self-consistency as criteria for thought judgment beyond thought 'correctness.' This has distinct differentiation from existing systems particularly in emotional state reasoning, ethical judgment, and user choice coordination. This value alignment concept ensuring AI decision-making aligns with human values is emerging as a core task for building safe and trustworthy intelligent systems [28].

E. Idle Loop-Based Knowledge Reconstruction Mechanism

SAVA performs the following operations internally even during times when thinking is not executed (idle state):

- Generalization and categorization between memories
- Commonality search and relationship formation
- Hierarchical reorganization of concept graphs
- User value alignment evaluation updates

These operations are performed by MA (Memory Agent), exploring possibilities for semantic emergence (significant novelty) without user intervention. This aims for operating methods similar to human unconscious thinking, particularly dreams, associations, and reinterpretation. Existing LLM systems have decisive structural differences in that no thinking or memory updates occur during times without input, and all learning depends on fine-tuning. Similarly, recent cases have been reported where LLM agents attempt dynamic memory updates and behavior improvement through self-summarization and reflection on accumulated experiences [27].

VI. Application Possibilities and Technical Extension Directions

SAVA (Self-Aware Vector Agent), based on its design structure enabling thought structuring, memory-based reasoning, and value alignment-based judgment, presents new possibilities for tasks difficult to solve with existing artificial intelligence systems across various intellectual activities and human-centered problem domains. This chapter explains SAVA's application possibilities divided into three categories: individual-level thinking support, collective intelligence formation, and social/ethical technology extension, and proposes technical extension directions.

A. Individual Level: Thinking Companion and Memory Extension

SAVA can provide continuous and internalized thinking support that existing AI systems cannot offer, as it can learn and reflect users' memories and value structures long-term.

A.1. Structured Thinking Assistance

- Store repetitive problem-solving structures in JSON format and automatically present procedural thought flows for similar future problems
- Meta-track long-term question flows or exploration directions and identify thinking blind spots
- Visualize user thinking to detect connections between high-level concepts or repetitive thinking patterns

A.2. Memory-Based Self-Reflection

- Store user experiences in episodic structures to analyze repetitive emotional states, value conflicts, and choice patterns
- Generate self-explanations of emotion-memory-judgment flows for specific topics
- Derive unconscious interests and cognitive vulnerabilities based on diaries, meeting minutes, notes, etc.

A.3. Value-Based Decision Assistance

- Align and recommend alternatives with high alignment to user value structures
- Provide simulation based on past similar judgment flows when value conflicts occur
- Respond from value dimensions to questions like ”Why does this choice make me uncomfortable?”

B. Collective Level: Collaborative Thinking Systems and Collective Intelligence

SAVA can extend beyond single users through connection and sharing between multiple SAVA instances to become a distributed thinking infrastructure enabling collective semantic creation and knowledge evolution. This aligns with the concept of human-AI co-evolution where humans and AI provide mutual feedback and develop together [29].

B.1. Federated Semantic Sharing Structure (SAVA-NET)

- Partial sharing of concepts, thought flows, and semantic networks according to user-selected sharing levels (L0-L3)
- Collective category reconstruction and knowledge generalization by fusing semantic memories of multiple SAVAs on common topics

B.2. Collective Value Comparison and Coordination

- Visualize individual value alignment differences for identical concepts
- Explicitly present reasons for value conflicts and adjustable points within teams or organizations
- Function as social explanation models for belief differences (e.g., ”Why do we think differently about this problem?”)

B.3. Meaning-Based Collective Intelligence Modeling

- Derive concepts emerging from semantic structures connected by multiple SAVAs → record birth of new concepts
- Jointly execute hypothesis generation → verification → reconstruction loops in domains like science, creativity, and policy design
- Realize ”thinking humans connected together” rather than ”machines thinking”

C. Social Level: Extension to Human-Centered Technology Ecosystem

SAVA was designed from the beginning centered on local execution, user-centered structure, and philosophically-based safety. This enables the following social and technical extensions.

C.1. Knowledge-Based Economy Structuring

- Thought units (JSON + VSA) become contribution units for knowledge production and combination, quantifying knowledge production and utilization
- Knowledge value evaluation based on how frequently specific thought flows are reused and induce derivative thinking
- This can lead to new knowledge economy models where human thinking itself, not AI output, becomes assets

C.2. Philosophically and Ethically Verifiable AI

- Securing transparency of judgment responsibility subjects and decision-making structures as AI judgment flows are explicitly recorded
- Valueable as explainable AI in domains requiring decision legitimacy such as policy, healthcare, and education

- Redefining AI bias problems from user-centered perspectives by structuring user value-based alignment

C.3. Expansion Possibility to Sensation and Emotion-Centered AI

- Combining multimodal sensory information with meaning-based structures (VSA) enables emotional state interpretation and sensation-based meaning composition
- Structuring non-linguistic thought fragments like ”feeling-based judgment” and ”memory-embedded scenes”
- Future expansion to vision/audition-based creation, emotion-responsive AI, intuitive thinking assistance

C.4. Technology Democratization and Sustainability

- Existing LLM-based AI systems operate dependently on centralized servers and high-performance GPUs, causing massive power consumption and carbon emission problems [30][31]
- SAVA, designed as a lightweight structure operable on personal terminals, can dramatically reduce energy consumption and is advantageous in environmental sustainability
- Operating without high-performance GPUs reduces AI technology access gaps [32] and increases technology democratization realization possibilities

D. Summary: AI System Where Technology Meets Philosophy

SAVA is not merely a high-performance LLM agent but a thinking system with structures for remembering, explaining, and reflecting. SAVA is AI that preserves and extends human thinking rather than replacing humans.

- For individuals: A tool to interpret and extend one's own thinking
- For groups: A structure enabling value-based knowledge collaboration and semantic emergence
- For society: An experimental alternative for explainable and responsible AI design

SAVA is a structural proposal for the human-centered intelligence era and can ultimately function as core infrastructure for knowledge-based ecosystems where thinking becomes assets.

VII. Conclusion and Future Research Plans

A. Conclusion: Structural Proposal for Human-Centered Artificial Intelligence

This paper identified structural limitations of existing large language model-based artificial intelligence systems and proposed SAVA (Self-Aware Vector Agent), a new artificial intelligence structure designed with human coexistence and cooperation possibilities at its center. SAVA is not a simple knowledge response system but a cognitive-memory-value integrated system that can simultaneously secure thinking continuity, explainability, and value alignability by explicitly structuring users' memories (semantic, procedural, episodic), value structures (LV0-LV2), and judgment flows (traceable reasoning).

SAVA designs thinking centered on humans. It does not replace humans and does not compete with humans. Rather, SAVA aims to realize artificial intelligence that preserves and extends human thinking by structuring humans' inner flows, preserving judgment contexts, and composing thought flows together based on complex values.
We judge SAVA as an attempt to transcend existing AI paradigms at the following three points:

- **Restoration of memory-based thinking**: Reproducing that thinking is not simple response but continuous structure connected with memory and context
- **Introduction of value-centered judgment structure**: Explicitly reflecting that all thinking involves value-based decision-making processes
- **Self-explainable thought flows**: Providing transparent structures where users can interpret and modify thinking

This structure is not merely technical improvement but also a practical answer to the philosophical question of how to preserve human agency in the AI era.

B. Future Research and Implementation Plans

SAVA has completed implementation and primary experiments of SAVA v1.0, the initial version centered on CA, and plans to proceed with the following implementation and research stepwise.

B.1. Implementation and Evaluation Framework Construction

- Development of SLM injection mechanisms to overcome small context window limitations
- Implementation of TLM-based procedural memory learning and micro-planner modules
- Experimentation with MA-based generalization/relationship formation algorithms and design of semantic emergence detection triggers
- Construction of log systems evaluating module activity and value alignment
- Enhancement of CA's self-reference and self-questioning for advanced reasoning

B.2. Future Paper Plans

- **CA's thought composition and action selection evaluation** → Thought unit composition, action selection accuracy, reasoning trace analysis
- **Semantic/episodic memory structuring and long-term memorization experiments** → JSON-to-VSA conversion accuracy, memory-based thinking effect verification
- **External search and reasoning visualization, concept generalization mechanism verification** → Search precision, relationship-based reasoning performance changes, explainability evaluation
- **Semantic emergence detection and idle time-based memory enhancement experiments** → Emergence event detection accuracy, idle memory update effects
- **VSA vector drift/bias response mechanism analysis** → Drift log stability, value alignment maintenance effects, rollback possibilities

Subsequently, experiments will expand to metacognitive self-reference structure extension, value-based recommendation, multi-user federation structures, multimodal expansion, and lightweight implementation. Particularly, metacognitive self-reference capabilities where agents monitor and improve their own reasoning processes are identified as lacking elements in current LLM systems [33], and will become important research directions for enhancing SAVA's reliability and transparency. For example, attempts have been made to introduce metacognitive prompts to large language models to improve comprehension [34], but such approaches alone have limitations in making AI structure its own reasoning processes.

C. Conclusion: Structurable Thinking, Explainable Artificial Intelligence

At a point where artificial intelligence evolution transcends technical performance, we are required to have systems with the following criteria:

- Structures that can explicitly reveal thought composition principles
- Transparency enabling sharing of judgment flows and grounds with users
- Continuity capable of safely preserving and utilizing human memory, emotion, and value structures

SAVA was designed to structurally satisfy these requirements and is a system that composes thinking based on memory, explains judgment according to values, and continuously maintains alignment with users. This is not merely AI with new functions but a proposal as a basic unit of artificial intelligence for preserving and extending human-centered thinking. This paper systematically presented its structural definition and design principles, and will more clearly reveal its possibilities and limitations through experimental verification in the future.

References

[1] OCEG, "The Explainability Challenge of Generative AI and LLMs," *OCEG Blog*, Dec. 11, 2024. [Online]. Available: `https://www.oceg.org/the-explainability-challenge-of-generative-ai-and-llms/`

[2] Factored AI, "LLM as a Judge: Evaluating LLM Outputs and the Challenge of Hallucinations," *Factored AI Engineering Blog*, Mar. 17, 2025. [Online]. Available: `https://www.factored.ai/knowledge-hub/llm-hallucination-evaluation`

[3] M. Griot, C. Hemptinne, J. Vanderdonckt, and D. Yuksel, "Large Language Models lack essential metacognition for reliable medical reasoning," *Nat. Commun.*, vol. 16, no. 1, p. 642, Jan. 2025, doi: 10.1038/s41467-024-55628-6. [Online]. Available: `https://www.nature.com/articles/s41467-024-55628-6`

[4] C. Zhai et al., "The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review," *Smart Learning Environments*, vol. 11, no. 28, 2024.

[5] M. Gerlich, "AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking," *Societies*, vol. 15, no. 1, Art. 6, 2025. doi: `10.3390/soc15010006`

[6] M. Khamassi et al., "Strong and weak alignment of large language models with human values," *Scientific Reports*, vol. 14, art. 19399, 2024.

[7] D. Swanepoel, "Does artificial intelligence have agency?" in *The Mind-Technology Problem: Investigating Minds, Selves and 21st Century Artifacts*, R. W. Clowes, K. Gärtner, and I. Hipólito, Eds. Springer, 2021, pp. 83–104.

[8] Y. Wu, S. Liang, C. Zhang, Y. Wang, Y. Zhang, H. Guo, R. Tang, and Y. Liu, "From human memory to AI memory: A survey on memory mechanisms in the era of LLMs," arXiv preprint arXiv:2504.15965, Apr. 2025. [Online]. Available: `https://arxiv.org/abs/2504.15965`

[9] J. R. Searle, "The Chinese Room Argument," in *Stanford Encyclopedia of Philosophy*, Spring 2018 ed., E. N. Zalta, Ed. Stanford University, Mar. 19, 2004. [Online]. Available: `https://plato.stanford.edu/entries/chinese-room/`

[10] I. M. Comșa and M. Shanahan, "Does It Make Sense to Speak of Introspection in Large Language Models?" arXiv preprint arXiv:2506.05068, 2025.

[11] A. Clark and D. Chalmers, "The extended mind," *Analysis*, vol. 58, no. 1, pp. 7–19, 1998.

[12] E. Tulving, "Episodic and Semantic Memory," in *Organization of Memory*, Academic Press, 1972.

[13] P. Neubert et al., "Introduction to Hyperdimensional Computing and Vector Symbolic Architectures for Robotics," *KI*, vol. 33, 2019.

[14] A. Clark, "Extending Minds with Generative AI," *Nature Communications*, vol. 16, no. 4627, 2025.

[15] S. Berretta et al., "Defining human-AI teaming the human-centered way: a scoping review and network analysis," *Frontiers in Artificial Intelligence*, vol. 6, Art. 1250725, 2023.

[16] A. Dafoe et al., "Cooperative AI: machines must learn to find common ground," *Nature*, vol. 593, pp. 33–36, 2021.

[17] P. Kanerva, "Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors," *Cognitive Computation*, vol. 1, no. 2, pp. 139–159, 2009.

[18] L. Shan, S. Luo, Z. Zhu, Y. Yuan, and Y. Wu, "Cognitive Memory in Large Language Models," arXiv preprint arXiv:2504.02441, 2025. [Online]. Available: `https://arxiv.org/abs/2504.02441`

[19] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[20] T. Mesnard et al., "Gemma: Open models based on Gemini research and technology," arXiv preprint arXiv:2403.08295, Mar. 13, 2024.

[21] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave and G. Lample, "LLaMA: Open and Efficient Foundation Language Models," *CoRR*, vol. abs/2302.13971, Feb. 2023.

[22] S. Hu, Y. Tu, X. Han, C. He, G. Cui, X. Long, Z. Zheng, Y. Fang, Y. Huang, W. Zhao et al., "MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies," arXiv preprint arXiv:2404.06395, Apr. 2024.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N.Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, Dec. 2017.

[24] J. Wang, Z. Zhang, Y. Jiang, B. Pope et al., "Layer-Condensed KV Cache for Efficient Inference of Large-Scale Transformer Decoders," in Proc. ACL, Jul. 2024, pp. 11175–11189.

[25] S. Kiyono, S. Kobayashi, J. Suzuki, and K. Inui, "SHAPE: Shifted Absolute Position Embedding for Transformers," in Proc. EMNLP, Sep. 2021, pp. 3648–3658.

[26] J. Pearl, "The seven tools of causal inference, with reflections on machine learning," *Communications of the ACM*, vol. 62, no. 3, pp. 54–60, 2019.

[27] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," arXiv preprint arXiv:2304.03442, 2023. [Online]. Available: `https://arxiv.org/abs/2304.03442`

[28] I. Gabriel, "Artificial intelligence, values and alignment," *Minds and Machines*, vol. 30, no. 3, pp. 411–437, 2020.

[29] D. Pedreschi et al., "Human-AI coevolution," arXiv preprint arXiv:2306.13723, 2024. [Online]. Available: `https://arxiv.org/abs/2306.13723`

[30] A. Zewe, "Explained: Generative AI's environmental impact," *MIT News*, Jan. 17, 2025. [Online]. Available: `https://news.mit.edu/2025/generative-ai-environment-impact`

[31] "AI's Growing Carbon Footprint," *Columbia Climate School – Earth Institute*, Jun. 9, 2023. [Online]. Available: `https://news.climate.columbia.edu/2023/06/09/ais-growing-carbon-footprint/`

[32] E. F. Emaojo, "How AI is widening the global/human development gap," *DevelopmentAid*, Jun. 24, 2025. [Online]. Available: `https://www.developmentaid.org/news-stream/post/196997/equitable-distribution-of-ai`

[33] M. T. Cox, "Metacognition in computation: a selected research review," *Artificial Intelligence*, vol. 169, no. 2, pp. 104–141, 2005.

[34] Y. Zhao and Y. Wang, "Metacognitive Prompting Improves Understanding in Large Language Models," in Proc. NAACL, 2024.