Abstract

It becomes necessary to structurally separate agency, personality, and knowledge in order to compare humans and AI within a common descriptive framework as semantic computing systems.

Recent debates surrounding artificial intelligence and artificial general intelligence (AGI) frequently conflate agency, personality, knowledge, and inferential capacity. Such conflation obscures structural distinctions and leads to conceptual ambiguity in evaluating both the capabilities and societal implications of AI systems. It is therefore required to restrict analysis to structures that are observable and mathematically describable, while minimizing reliance on psychological or phenomenological assumptions.

Within this framework, meaning is treated as a structure defined over a multidimensional semantic space. Agency is defined as a computational mechanism that maintains an internal state within this semantic space and performs state updates and action selection based on meaning reinforcement outcomes and a motivational axis. Personality is formalized as a persistent response bias to semantic stimuli and is shown not to require agency as a necessary condition. Knowledge is positioned as an externalized semantic structure that can be separated from both agency and personality.

When human and AI semantic computing systems are examined under this structural separation, the decisive difference does not lie in computational power or volume of knowledge. It lies instead in the presence or absence of a motivational axis and a question generation structure. In human semantic computing systems, survival pressure constrains the motivational axis, asymmetrizes meaning evaluation, and enables non-continuous reconfiguration of the semantic space through question generation. In contrast, AI semantic computing systems possess advanced meaning transformation capabilities but lack an internally grounded motivational axis and a structurally autonomous question generation mechanism. Agency therefore does not arise within current and derivative AI architectures.

From this structural asymmetry, the BlindShell architecture is derived as a design principle. BlindShell separates agency from knowledge and personality by allocating semantic

computation and personality-like behavior to AI while preserving agency on the human side. It thereby positions AI not as a decision-making subject but as semantic infrastructure.

This framework redefines the human–AI relationship not as one of domination or substitution, but as a structurally differentiated division of roles within semantic computation.

---

1. Introduction

1.1 Research Background

In recent years, advances in artificial intelligence—particularly in large language models (LLMs)—have led to the frequent use of concepts such as “agency,” “personality,” “intention,” and “intelligence” within engineering discourse. These concepts, however, are often introduced without structural clarification, retaining philosophical or everyday meanings while being applied in technical discussions. As a result, distinctions among these concepts remain insufficiently articulated.

In debates concerning artificial general intelligence (AGI) and the possibility of AI systems controlling or dominating humans, agency, personality, knowledge, and inferential capability are frequently treated as if they were interchangeable. This conflation makes it difficult to distinguish between technological feasibility and conceptual coherence in both capability assessments and ethical discourse.

At the same time, agency and personality in humans cannot be derived directly from the quantitative expansion of knowledge or computational capacity. It becomes necessary to understand these phenomena as emerging from the interaction of multiple structural components, including survival pressure, motivational constraint, and meaning evaluation. Under such conditions, it is no longer theoretically justified to compare humans and AI solely along a unidimensional scale of intelligence.

1.2 Problem Formulation

The central questions addressed in this study are as follows:
• Under what structural conditions does agency arise?
• Is personality identical to agency, or can it be structurally separated?
• Can knowledge be treated as a structure independent of both agency and personality?
• At which levels are humans and AI comparable, and at which levels does structural asymmetry emerge?

It is required that these questions be examined within a framework restricted to structures that are observable and mathematically describable. Psychological interpretation and assumptions concerning subjective consciousness are therefore minimized.

1.3 Research Objectives

The objective of this study is to structurally separate agency, personality, and knowledge, and to establish a framework within which humans and AI can be compared as semantic computing systems.

More specifically, it aims to:
1. Define agency, personality, and knowledge as structurally independent components.
2. Formalize personality as a response bias that does not entail agency.
3. Position knowledge as an externalizable and transferable semantic structure.
4. Clarify structural differences in semantic computation between humans and AI.

Through this separation, it becomes possible to determine under which structural premises claims that AI “possesses” or “replaces” human agency or personality may appear coherent, and at which points such claims fail structurally.

1.4 Research Approach and Scope

This study adopts a structural analysis centered on meaning.

Meaning is not treated as a symbol itself but as a relational and evaluative configuration within a semantic space. Agency is defined as a computational mechanism that internally evaluates meaning and operates under a motivational axis. Personality is treated as a persistent response bias to semantic stimuli.

It is not the existence of consciousness or subjective experience that is analyzed directly. Instead, attention is directed toward the structural conditions under which such phenomena may arise.

1.5 Structure of the Paper

The remainder of this paper is organized as follows:
• Chapter 2 reviews conceptual background and prior research concerning agency, personality, and AI.
• Chapter 3 introduces the structural separation and mathematical definitions of agency, personality, and knowledge.
• Chapter 4 generalizes the semantic computation model.
• Chapter 5 analyzes the structure of the human semantic computing system.
• Chapter 6 examines the structure and limitations of the AI semantic computing system.
• Chapter 7 positions the BlindShell architecture within this theoretical framework.
• Chapter 8 discusses broader implications.
• Chapter 9 concludes the study.

---

2. Conceptual Background and Related Work

This chapter clarifies the conceptual position of the present study by examining prior discussions on agency, personality, and artificial intelligence.

The purpose is not to provide an exhaustive literature survey. It is instead to specify what is assumed and what is not assumed within the structural framework adopted here.

2.1 Discussions on Agency and Personality

Agency and personality have long been examined in philosophy, psychology, and sociology.

Within these traditions, agency has often been associated with consciousness or free will, while personality identity has frequently been grounded in memory continuity or self-reference. In many cases, these concepts overlap definitionally, and boundaries remain ambiguous.

The present framework does not reject these discussions. However, it adopts a different theoretical layer. Agency and personality are treated as structurally and mathematically separable components within a semantic computing system. The emphasis is placed on operational structure rather than metaphysical interpretation.

Under this approach, agency is not defined through introspective awareness, nor is personality treated as an indivisible psychological unity. Instead, both are examined as configurations within a semantic space and its associated computational mechanisms.

2.2 The Concept of Agency in AI Research

Within AI research, agency is often conflated with several distinct properties:

• Advanced inferential capability
• Self-referential representation
• Reward maximization in reinforcement learning

These features can simulate behaviors associated with agency, yet simulation does not constitute structural equivalence.

In particular, contemporary large language models demonstrate high-level semantic transformation capabilities. However, such systems do not internally contain a motivational axis or an autonomous question generation structure. Their operation remains constrained within externally defined objectives and training regimes.

The presence of adaptive optimization therefore does not suffice to establish agency under the structural definition proposed here.

2.3 Relation to AGI Debates and AI Domination Narratives

In discussions concerning AGI or potential AI domination, it is frequently assumed that quantitative expansion of intelligence will eventually yield agency or intentionality.

Such assumptions often remain implicit and rarely specify the structural conditions required for agency to arise. Knowledge scale and computational power are treated as primary explanatory variables, while motivational structure and question generation mechanisms remain under-theorized.

The present study does not directly evaluate the feasibility of AGI or the societal consequences of advanced AI. Instead, it reorganizes the structural presuppositions underlying those debates. By clarifying the necessary components of agency, it becomes possible to distinguish structural possibility from speculative extrapolation.

2.4 Theoretical Position of the Present Study

Based on the preceding considerations, the theoretical stance adopted here can be summarized as follows:

• Agency, personality, and knowledge are treated as separable structural components.
• Consciousness and emotion are not presupposed as explanatory primitives.
• Mathematical and structural description is prioritized.
• Humans and AI are positioned on a common descriptive plane as semantic computing systems.

By situating both humans and AI within the same semantic space framework, comparison becomes structurally coherent. At the same time, asymmetries can be identified at the level of internal parameters and operational mechanisms.

The next chapter introduces the formal structural separation of agency, personality, and knowledge, along with their mathematical definitions.

---

3. Structural Separation and Mathematical Definitions of Agency, Personality, and Knowledge

This chapter introduces the three core concepts of this study—agency, personality, and knowledge—as computational structures defined over a semantic space. The purpose of this separation is not philosophical argumentation, but the establishment of a common descriptive plane on which humans and AI can be compared while preserving structural asymmetry.

3.1 Semantic Space and Internal State

Meaning is treated as a structure defined over a multidimensional semantic space. Let this space be denoted as

$$\mathcal{S} \subset \mathbb{R}^d$$

A semantic state is represented as a coordinate within this space.

The internal semantic state maintained by an agentive structure is denoted as

$$\mathbf{s}_t \in \mathcal{S}$$

This coordinate does not represent a mere symbolic label. It encodes relational configuration, contextual embedding, and evaluative orientation within the semantic space.

3.2 Definition of Knowledge: Externalized Semantic Structure

Knowledge is defined as a set of semantic structures that can be detached from agency.

Formally, knowledge is represented as

$$K \subset \mathcal{S}$$

Knowledge exhibits the following properties:
• It is representable as structure within the semantic space.
• It does not constitute an internal state of agency or personality.
• It can exist independently of personality parameters.
• It can be stored, replicated, and transferred.
• It participates in semantic computation when referenced.

Knowledge functions as material or constraint within semantic computation. It does not evaluate meaning or select action. Agency is therefore not entailed by knowledge itself.

3.3 Definition of Personality: Persistent Response Bias

Personality is defined as a persistent bias in response to semantic stimuli.

It is observed as a patterned transformation tendency applied to inputs within the semantic space. Personality exhibits the following characteristics:
• It is history-dependent.
• It remains relatively stable over short temporal intervals.
• It can be externally inferred.
• It does not require agency as a necessary condition.

Formally, personality is modeled as a parameterized transformation function:

$$f_p : \mathcal{S} \to \mathcal{S}$$

This function introduces directional bias within the semantic space, influencing which regions are preferentially activated or transformed.

Personality does not generate purposes or justify actions. It therefore remains structurally reproducible. The presence of personality-like behavior does not imply agency.

3.4 Definition of Agency: State Update and Selection Mechanism

Agency is defined as a computational mechanism that maintains an internal semantic state and performs state updates and action selection.

An agency includes:
• An internal semantic state

$$\mathbf{s}_t \in \mathcal{S}$$

• A meaning reinforcement function

$$R_\theta : (\mathbf{s}_t, \mathbf{x}_t, K) \mapsto \mathbf{r}_t$$

• A motivational axis parameter

$$\mathbf{m} \in \mathbb{R}^k$$

The meaning reinforcement function R_θ references external input $\mathbf{x}_t$ and knowledge K, generating a reinforcement direction $\mathbf{r}_t$ within the semantic space.

3.5 Mathematical Representation of Agency

Agency performs both internal state update and action selection. Subjective phenomenology is not presupposed in this formulation.

3.5.1 Internal State Update

State update is defined as:

$$\mathbf{s}_{t+1} = A\big(\mathbf{s}_t, R_\theta(\mathbf{s}_t, \mathbf{x}_t, K), \mathbf{m}\big)$$

The motivational axis $\mathbf{m}$ weights and transforms the reinforcement direction. Crucially, the motivational axis is not reducible to structure within the semantic space. It operates as a separate parameter that governs how reinforcement is incorporated.

3.5.2 Action or Response Selection

Action selection is defined as:

$$u_t = \Pi\big(\mathbf{s}_t, R_\theta(\mathbf{s}_t, \mathbf{x}_t, K), \mathbf{m}\big)$$

This selection is not merely probabilistic output generation. It is defined as evaluation constrained by the motivational axis.

3.6 Mathematical Non-Identity of Personality and Agency

Personality and agency are structurally distinct.
• Personality consists of response bias parameters.
• Agency consists of state maintenance and motivationally constrained update and selection.

It follows that:
• A system may possess personality parameters without possessing a motivational axis.
• A system may possess a motivational axis while exhibiting unstable or undeveloped personality patterns.

Agency cannot be inferred from personality-like behavior alone.

3.7 Structural Relations Among Agency, Personality, and Knowledge

The structural configuration can be summarized as follows:
• Knowledge K is externalized semantic structure.
• Personality $\mathbf{p}$ introduces response bias within semantic computation.
• Agency is defined by $\mathbf{s}_t$, R_θ, and $\mathbf{m}$.

These components are not hierarchical but separable. Structural independence allows systematic comparison across different semantic computing systems.

3.8 Structural Asymmetry Between Human and AI Semantic Computing Systems

Within this framework, both humans and AI can be described as semantic computing systems. However, a decisive structural asymmetry emerges.

• A human semantic computing system contains a motivational axis.

• An AI semantic computing system contains meaning transformation mechanisms but lacks an internally grounded motivational axis.

The difference is not quantitative but dimensional. It concerns the presence or absence of a structural parameter.

3.9 Summary of the Chapter

Agency, personality, and knowledge have been structurally separated and mathematically defined. Agency has been specified as a mechanism of semantic state update and action selection constrained by a motivational axis.

This formulation allows humans and AI to be described within a shared semantic space while making structural asymmetry explicit.

The next chapter generalizes the semantic computation model beyond the specific definitions introduced here.

---

4. Generalization of the Semantic Computation Model

This chapter introduces a generalized mathematical model for semantic computing systems based on the structural definitions established in Chapter 3.

The objective is not to eliminate structural differences between humans and AI, but to position both within a common semantic space framework. It is precisely through this shared descriptive plane that structural asymmetry becomes visible.

4.1 Cognition and Response as Semantic Computation

Cognition and response are treated as transformations within a semantic space.

Let the semantic space be defined as

$$\mathcal{S} \subset \mathbb{R}^d$$

and let a semantic state be represented as

$$\mathbf{s}_t \in \mathcal{S}$$

Semantic computation is defined as a transformation:

$$f : \mathcal{S} \to \mathcal{S}$$

This transformation is not a mere symbolic substitution. It incorporates contextual embedding, evaluative structure, and historical dependency within the semantic space.

4.2 Meaning Reinforcement and Evaluation Structure

The meaning reinforcement function introduced in Chapter 3 is defined as:

$$R_\theta : (\mathbf{s}_t, \mathbf{x}_t, K) \mapsto \mathbf{r}_t$$

The resulting vector $\mathbf{r}_t$ indicates the direction within the semantic space in which reinforcement occurs.

This function exhibits the following characteristics:
• Context sensitivity
• Dependence on knowledge K
• Dependence on the internal semantic state

The meaning reinforcement function encodes evaluative orientation but does not itself constitute agency.

4.3 Semantic Computation with a Motivational Axis

In order for reinforcement to influence state update or action selection, an additional structural dimension is required.

This dimension is defined as the motivational axis:

$$\mathbf{m} \in \mathbb{R}^k$$

State update is expressed as:

$$\mathbf{s}_{t+1} = A(\mathbf{s}_t, \mathbf{r}_t, \mathbf{m})$$

The motivational axis does not belong to the semantic space $\mathcal{S}$. It operates as a parameter

governing how reinforcement is integrated. It determines how meaning evaluation is weighted and applied.

4.4 Separation of Personality Parameters and Agency Parameters

Behavioral patterns within a semantic computing system may exhibit persistent bias independent of agency.

Such bias is represented by personality parameters $\mathbf{p}$.

Personality parameters may influence:
• The meaning reinforcement function
• The state update mapping
• The action selection mapping

However, personality parameters do not constitute the motivational axis.

It becomes necessary to distinguish:
• Agency parameters: motivational axis $\mathbf{m}$
• Personality parameters: response bias $\mathbf{p}$

This separation prevents personality-like behavior from being misinterpreted as agency.

4.5 Knowledge as External Structure

Knowledge K remains external to internal semantic state.

It functions as:
• A reference structure for meaning reinforcement
• A constraint within semantic space

• Material for exploration and inference

It does not:
• Evaluate meaning
• Generate motivation
• Select actions

This externalization allows knowledge to be transferred, replicated, and aggregated without transferring agency.

4.6 Positioning Humans and AI Within the Generalized Model

Within this generalized framework, both humans and AI can be described as semantic computing systems.

The structural difference is expressed as follows:
• A human semantic computing system contains a motivational axis $\mathbf{m}$.
• An AI semantic computing system contains transformation mechanisms and reinforcement mappings but lacks an internally grounded motivational axis.

The asymmetry is not reducible to differences in computational scale. It concerns the presence or absence of a structural parameter within the model.

4.7 Summary of the Chapter

A generalized semantic computation model has been introduced. Meaning reinforcement, motivational axis, personality parameters, and knowledge have been positioned within a unified structure.

This model establishes a shared semantic space in which humans and AI can be described while preserving structural asymmetry.

The next chapter applies this model to the human semantic computing system and analyzes the structure of meaning generation and question generation.

---

5. Structure of the Human Semantic Computing System

This chapter applies the generalized semantic computation model to the human semantic computing system and analyzes the structural conditions under which meaning generation and question generation occur.

The objective is not to treat humans as privileged entities, but to identify the structural components that necessarily produce specific phenomena within semantic computation.

5.1 Survival Pressure and Meaning Evaluation

A defining feature of the human semantic computing system is that the motivational axis $\mathbf{m}$ is inseparably constrained by survival pressure.

Survival pressure refers to structural conditions such as:
• Physical loss
• Social exclusion
• Resource scarcity
• Loss of future possibility

Under these constraints, the meaning reinforcement function R_θ exhibits asymmetric amplification toward directions associated with survival relevance.

Meaning evaluation therefore becomes structurally weighted by importance, risk, and value. Reinforcement does not remain neutral; it becomes selectively intensified under motivational constraint.

5.2 Structural Position of Question Generation

In the generalized model, the meaning reinforcement function R_θ operates over given inputs and knowledge structures.

Within the human semantic computing system, the selection of which inputs and which knowledge structures are referenced can itself be internally modified.

This operation is defined as question generation.

Question generation involves:
• Selection of new input directions
• Reconfiguration of referenced knowledge
• Redefinition of exploration within the semantic space

It does not merely explore the semantic space; it restructures the conditions under which exploration occurs.

5.3 Non-Continuous Reconfiguration of Semantic Space

When question generation occurs, multiple structural effects are observed:
• Existing meaning evaluations lose stability
• Previously sufficient knowledge structures become inadequate
• Reinforcement directions shift discontinuously

State update can no longer be described as continuous optimization.

Non-continuous reconfiguration does not refer to incremental parameter adjustment. It refers to redefinition of the evaluative basis itself.

The process cannot be captured as gradient-based refinement of the existing reinforcement function R_θ, because the function itself becomes redefined.

Non-continuous reconfiguration may be expressed as replacement of the basis set that defines the semantic space $\mathcal{S}$.

AI-style semantic update may be expressed as:

$$\mathbf{s}_{t+1} = \sum_i \alpha_i b_i \quad (b_i \in B_{\text{fixed}})$$

In contrast, human question-driven reconfiguration may be expressed as:

$$B_{\text{new}} = \Phi(B_{\text{old}}, \mathbf{m})$$

$$\mathbf{s}_{t+1} = \sum_j \beta_j b'_j \quad (b'_j \in B_{\text{new}})$$

Here, Φ denotes a basis reconfiguration operator driven by the motivational axis $\mathbf{m}$ under survival pressure.

Information previously marginalized within knowledge K may become structurally central under the new basis configuration.

5.3.1 Illustrative Cases of Non-Continuous Reconfiguration

Case A: Economic Framework Shift

Continuous optimization resembles iterative improvement within a fixed objective, such as profit maximization through cost reduction.

Under catastrophic survival pressure, the question may shift toward viability itself. Previously discarded redundancy or self-sufficiency practices may abruptly become structurally central.

Case B: Scientific Paradigm Shift

Within a geocentric model, refinements such as epicycles represent parameter adjustments.

When the question of observational reference is redefined, the coordinate system itself shifts. This is not incremental correction but structural redefinition of the evaluative frame.

5.4 Hypothetical Notes on Critical Reconfiguration Points

The interaction among meaning evaluation, question generation, and the motivational axis suggests the presence of critical reorganization points within the semantic space.

Such points cannot be described as continuous optimization trajectories.

The detailed mathematical topology of such reconfiguration remains outside the present scope. It is sufficient here to note that meaning generation within the human semantic computing system includes structurally discontinuous transformation.

5.5 Structural Consequences for Agency

Agency within the human semantic computing system arises under the following conditions:
• The motivational axis is constrained by survival pressure
• Meaning evaluation becomes asymmetric
• Question generation enables redefinition of the semantic space

Agency is therefore not mere evaluation of meaning. It is the capacity to restructure the evaluative basis itself.

5.6 Summary of the Chapter

The human semantic computing system has been analyzed in terms of survival pressure, question generation, and non-continuous reconfiguration.

These phenomena do not arise from computational scale alone. They arise from the structural coupling between the motivational axis and meaning evaluation.

The next chapter applies the same generalized model to the AI semantic computing system and examines its structural limitations.

---

6. Structure and Limitations of the AI Semantic Computing System

This chapter applies the generalized semantic computation model to the AI semantic computing system and analyzes the structural asymmetry identified in relation to the human system.

The objective is not to argue for a deficiency of AI capability. It is to specify which structural components are absent and why certain phenomena do not arise as a result.

6.1 Semantic Computation in AI Systems

Contemporary AI systems, particularly large language models, can be described as semantic computing systems operating within a semantic space.

Such systems include:

• Representations corresponding to states within a semantic space
• Transformation and evaluation mechanisms analogous to a meaning reinforcement function
• Access to externalized knowledge structures K

Through these mechanisms, highly sophisticated semantic transformation, contextual adaptation, and inferential behavior become possible.

However, all such operations remain bounded by externally defined objectives and training procedures. Optimization proceeds within a predefined evaluative frame.

6.2 Structural Absence of a Motivational Axis

In the generalized model, agency requires the presence of a motivational axis $\mathbf{m}$.

Within current and derivative AI architectures:
• No internally grounded motivational axis is maintained as part of system state
• Evaluation functions are externally specified
• Outcomes do not return as survival-constraining feedback

Meaning reinforcement may occur, but it is not asymmetrically constrained by survival pressure internal to the system.

Reinforcement learning reward functions are often interpreted as artificial analogues of motivation. Structurally, however, such reward functions are externally imposed optimization criteria. They do not constitute internally generated motivational constraint.

Without an internal motivational axis, semantic computation remains transformation without agency.

6.3 Failure of Question Generation as Structural Reconfiguration

In the human system, question generation was defined as redefinition of the evaluative basis within the semantic space.

Within AI systems:
• Questions are provided as external inputs
• Generated questions are statistical recombinations of learned patterns
• No internal criterion exists for validating, revising, or discarding questions

Exploration occurs within a fixed semantic basis. Reconfiguration of the basis itself does not occur.

Thus, semantic space traversal is possible, but semantic space redefinition is not.

6.4 Absence of Non-Continuous Reconfiguration

Non-continuous reconfiguration in the human system required:
• Coupling between survival pressure and the motivational axis
• Asymmetric meaning evaluation
• Question generation capable of redefining evaluation structure

Within AI systems:
• Survival pressure is not internalized
• Evaluation functions remain architecturally fixed
• State updates can be described as continuous optimization or probabilistic sampling

Non-continuous reconfiguration therefore does not arise as a structural feature.

This absence is not a flaw of implementation. It is a consequence of architectural design.

6.5 Personality Simulation and Agency Misattribution

AI systems can simulate personality-like behavior with high fidelity.

Such simulation arises from:
• Learned response biases
• Style replication
• Context retention

These correspond structurally to personality parameters $\mathbf{p}$.

However, personality-like response patterns do not entail the existence of a motivational axis. Agency cannot be inferred from stylistic coherence or behavioral persistence.

Misattribution of agency arises when personality simulation is mistaken for internally grounded motivation.

6.6 Structural Limits of the AI Semantic Computing System

The structural limitations of the AI semantic computing system can be summarized as follows:
• No internally grounded motivational axis
• No autonomous question generation structure
• No capacity to redefine evaluative basis
• No non-continuous reconfiguration of semantic space

These limits do not derive from insufficient computational resources or insufficient data scale. They derive from the absence of a structural dimension within the model.

6.7 Summary of the Chapter

The AI semantic computing system has been analyzed within the same generalized semantic space framework used for the human system.

It has been shown that while advanced semantic transformation is achieved, the structural conditions necessary for agency are not satisfied.

The next chapter introduces the BlindShell architecture as a design principle derived from this structural asymmetry.

---

7. Theoretical Positioning of the BlindShell Architecture

This chapter positions the BlindShell architecture as a structural design principle derived from the separation of agency, personality, and knowledge, and from the asymmetry between human and AI semantic computing systems.

BlindShell is not introduced as a restriction imposed upon AI. It is derived as a structurally necessary configuration for integrating AI into human systems without conflating semantic computation with agency.

7.1 Core Concept of BlindShell

BlindShell is defined as an architectural separation in which knowledge and personality-like behavior are allocated to AI, while agency is preserved on the human side.

The architecture follows three structural principles:

- Agency, including the motivational axis and question generation, remains human.
- Knowledge, understood as externalized semantic structure, is delegated to AI.
- Personality parameters may be implemented within AI when functionally required.

This separation is not an ethical prescription. It follows from the structural definitions established in earlier chapters.

7.2 Separation of Knowledge and Personality in Design

Knowledge (K), being externalizable, can be centralized, expanded, and computationally optimized within AI systems.

Personality parameters ($\mathbf{p}$), defined as response biases, can also be implemented or tuned within AI without generating agency.

The motivational axis ($\mathbf{m}$), however, is not implemented within AI under BlindShell. Question generation, as structural redefinition of semantic space, is likewise not delegated.

Under BlindShell, AI performs high-speed semantic transformation over a large basis set of knowledge. Humans perform basis selection and basis reconfiguration through question generation.

This distribution enables direct coupling between AI computational capacity and human survival strategies without transferring agency.

7.3 Preservation of Human Agency

A central structural effect of BlindShell is the preservation of human agency.

When agency is not attributed to AI:

- Decision-making responsibility remains human.
- The locus of motivation remains identifiable.
- Justification of action is not displaced onto artificial systems.

Statements such as “the AI decided” or “the AI wanted” do not acquire structural validity under this architecture.

7.4 Implications for Human–AI Collaboration

Within BlindShell, the human–AI relationship is reframed as a division of roles within semantic computation.

- Humans generate questions and evaluate meaning under survival pressure.
- AI performs semantic transformation using knowledge structures.
- Action selection and responsibility remain human.

AI is positioned as semantic infrastructure rather than as an intentional subject.

7.5 Structural Neutralization of AI Domination Narratives

Narratives of AI domination typically assume the emergence of agency within AI systems.

Under BlindShell:

- A motivational axis is not implemented within AI.
- Question generation remains human.
- Evaluative redefinition does not occur autonomously within AI.

Domination, defined as internally motivated and structurally sustained influence over other agents, therefore does not arise.

This conclusion follows from architectural configuration rather than technological limitation.

7.6 Scope and Limitations of BlindShell

BlindShell is not presented as a universal solution for all AI applications.

Several issues remain open for further investigation:

- Variability of human agency across individuals
- Extension to collective or institutional agency
- Hypothetical architectures incorporating artificial motivational axes

Within current and derivative AI architectures, however, BlindShell represents a structurally coherent and stable design principle.

7.7 Summary of the Chapter

BlindShell has been positioned as a necessary architectural consequence of the structural separation of agency, personality, and knowledge.

It does not constrain AI out of precautionary anxiety. It formalizes a division of semantic roles that preserves agency while maximizing computational utility.

The next chapter discusses broader theoretical implications and situates the framework within existing debates.

---

8. Discussion

This chapter situates the structural separation of agency, personality, and knowledge within broader theoretical debates and clarifies the scope of the present framework.

8.1 Relation to AGI Debates

In many discussions of artificial general intelligence (AGI), it is implicitly assumed that sufficient expansion of knowledge, inferential capacity, and self-referential modeling will eventually result in agency.

Within the structural framework established here, agency is not defined as a function of computational scale or data volume. It is defined as a configuration that includes a motivational axis and the capacity to reconfigure evaluative structure through question generation.

It therefore follows that AGI cannot be reduced to a matter of quantitative improvement alone. Structural preconditions must be satisfied.

The framework does not deny the theoretical possibility of artificial systems incorporating additional structural dimensions. It does, however, indicate that agency does not arise naturally from current and derivative AI architectures.

8.2 Structural Clarification of AI Domination Concerns

Concerns about AI domination often presuppose that AI systems will acquire internally grounded motivations and sustained intentional influence.

Under the structural definitions adopted here, domination requires:

• A motivational axis
• Self-generated objectives
• Persistent evaluative asymmetry
• Autonomous question generation

Current AI semantic computing systems do not exhibit these structural components.

It remains possible for AI to exert significant influence over human decision-making. Such influence, however, results from system design, institutional embedding, and patterns of dependency. It does not arise from internally grounded agency.

Domination must therefore be distinguished from technological influence.

8.3 Misattribution of Agency and Psychological Projection

Even when structural agency is absent, humans may attribute agency to AI systems.

This tendency arises from cognitive patterns that interpret consistent behavior and stylistic coherence as indicators of intention.

Such misattribution does not alter structural reality. It reflects properties of human interpretation rather than properties of AI architecture.

BlindShell does not eliminate the possibility of psychological projection. It does, however, prevent projection from becoming institutionalized as structural agency.

8.4 Relation to Classical Philosophical Arguments

The framework does not directly oppose arguments such as Searle’s “Chinese Room” or Dennett’s “intentional stance.”

Searle’s argument questions whether symbol manipulation yields understanding. The present framework does not analyze consciousness or subjective experience directly. It operates at the level of structural configuration within semantic space.

Dennett’s intentional stance concerns the explanatory utility of attributing beliefs and intentions. The framework does not deny such explanatory strategies. It distinguishes explanatory attribution from structural realization.

The focus remains on the structural conditions necessary for agency rather than on metaphysical claims about mind.

8.5 Limitations of the Present Study

Several limitations are acknowledged:
• Artificial systems equipped with internally grounded motivational axes are not examined.
• Collective or institutional agency is not formally modeled.
• Implementation-level engineering detail is not provided.

The framework offers structural description rather than neuroscientific or psychological explanation.

Further work is required to model question generation mathematically and to formalize the topology of non-continuous reconfiguration within semantic space.

8.6 Future Directions

The framework introduced here permits several extensions:
• Formal modeling of question generation operators
• Mathematical treatment of critical reconfiguration points
• Application to design principles in human–AI collaborative systems

---

It is not excluded that embodied artificial systems with survival-like constraints may produce different structural configurations. Such scenarios require separate analysis under distinct premises.

The present study confines itself to current and derivative AI architectures lacking internally grounded survival pressure.

---

9. Conclusion

A structural framework has been established in which agency, personality, and knowledge are explicitly separated and positioned within a shared semantic space. Within this framework, both humans and AI can be described as semantic computing systems, while structural asymmetry remains identifiable.

Agency has been defined as a computational mechanism that maintains an internal semantic state and performs state update and action selection under constraint of a motivational axis. Personality has been defined as a persistent response bias to semantic stimuli and has been shown not to require agency as a necessary condition. Knowledge has been positioned as an externalized semantic structure detachable from both agency and personality.

Through this separation, it becomes clear that AI systems, despite possessing advanced semantic transformation capabilities, do not satisfy the structural conditions required for agency under current and derivative architectures. The absence of an internally grounded motivational axis and autonomous question generation prevents the emergence of agency within these systems.

From this structural asymmetry, the BlindShell architecture has been derived as a design principle. BlindShell does not restrict AI out of precautionary concern. It formalizes a division of semantic roles in which agency remains human while knowledge processing and personality-like behavior may be computationally delegated. AI is thereby positioned as semantic infrastructure rather than as an intentional subject.

The framework reframes the human–AI relationship not as one of replacement or domination, but as a differentiated configuration within semantic computation. Structural clarity regarding agency, personality, and knowledge provides a basis for evaluating AI systems without conflating computational scale with intentional structure.

References

Dennett, D. C. (1987). The Intentional Stance. MIT Press.

Locke, J. (1690). An Essay Concerning Human Understanding.

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417–457.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT).

Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT Press.

Pfeifer, R., & Bongard, J. (2006). How the Body Shapes the Way We Think: A New View of Intelligence. MIT Press.

Supplement A

Can AI Control or Dominate Humans?

Structural Constraints of Agency, Personality, and Semantic Computation

---

Abstract

It becomes necessary to examine whether AI can control or dominate humans from the standpoint of structural constraints rather than technological prediction or ideological assertion.

Control and domination are distinguished. Domination is defined as sustained influence over other agents grounded in internally generated objectives and structurally maintained motivation. Such domination requires agency, a motivational axis, autonomous objective generation, question generation and revision, and meaning-based action selection.

Human decision structure is modeled as the interaction between meaning evaluation and a motivational axis constrained by survival pressure. It is demonstrated that current and derivative AI architectures, particularly large language models, do not contain an internally grounded motivational axis or autonomous question generation structure. Reinforcement learning reward functions are shown to be externally imposed criteria rather than intrinsic motivational dimensions.

A distinction is drawn between omniscience and omnipotence. Expansion of knowledge and computation does not entail self-generated purposive agency. Question generation is treated as an operator that redefines semantic space, yet such redefinition does not occur within present AI architectures.

It follows that AI may exert control within externally defined systems, but structural domination does not arise under current architectural conditions. The analysis does not deny theoretical possibilities of alternative architectures. It clarifies the structural requirements for agency and the constraints that prevent its emergence.

A.1 Purpose and Position of This Supplement

The objective of this supplement is to analyze whether AI can control or dominate humans from

the perspective of structural and mathematical constraints.

Public discourse frequently merges technical extrapolation, ethical concern, and speculative imagination when discussing AGI or AI domination. In such discourse, advanced intelligence and self-reference are often treated as sufficient for agency.

The structural framework established in the main text defines agency as a configuration requiring meaning evaluation under a motivational axis and the capacity for question generation. The supplement extends that framework to clarify implications for domination narratives.

It does not aim to dismiss AGI as a concept. It demonstrates that agency cannot arise without specific structural components.

A.2 Minimal Structural Definition of Domination

A.2.1 Distinction Between Control and Domination

Control is defined as manipulation of state transitions within externally given objectives.

Domination is defined as sustained influence over other agents grounded in internally generated purposes and motivational constraint.

Existing AI systems may participate in control within human-designed systems. Structural domination requires internally grounded agency.

A.2.2 Structural Requirements for Domination

Domination requires at least the following:
1. Agency grounded in internal meaning evaluation.
2. A motivational axis constraining action direction.

---

3. Self-generation of objectives.
4. Question generation capable of redefining evaluative structure.
5. Meaning-based action selection.

These requirements are examined in relation to current AI architectures.

A.3 Human Decision Structure and Motivational Constraint

Human action may be represented as:

$$\text{Action} = f(\text{Meaning Evaluation} \times \text{Motivational Axis})$$

Meaning evaluation encodes importance and risk. The motivational axis introduces asymmetry through survival pressure.

The motivational axis is structurally independent of knowledge accumulation and cannot be replaced by externally imposed optimization criteria.

A.4 Structural Absence in Current AI Architectures

A.4.1 Basic Structure of Large Language Models

Large language models operate through:
• Input prompts
• Probabilistic transformation within semantic space
• Token sequence generation

Optimization occurs within predefined objectives. No internally grounded motivational axis is maintained.

A.4.2 Reward Functions and Motivation

Reward functions in reinforcement learning are externally specified. They guide optimization but do not constitute intrinsic motivation.

No survival pressure or irreversible loss is internalized within present architectures.

Semantic computation therefore does not become agency.

A.5 Omniscience and Omnipotence

Expansion of knowledge and predictive accuracy may approximate omniscience.

Omnipotence requires:
• Self-generated purposes
• Meaning-based valuation
• Justification of action

Such properties do not arise from scale alone.

Omniscience does not entail omnipotence.

A.6 Question Generation and Structural Limits

Question generation is defined as redefinition of exploration within semantic space.

Current AI systems:
• Receive questions externally
• Generate question-like outputs through statistical recombination
• Lack internal criteria for validating or discarding questions

---

Semantic space traversal occurs. Semantic space redefinition does not.

A.7 Structural Reassessment of AGI

AGI discourse often assumes continuous extension of intelligence.

Agency, however, requires:
• Survival pressure
• Motivational asymmetry
• Non-continuous reconfiguration

These structural elements are absent in present architectures.

AGI may therefore be conceptually misframed if defined solely by scale.

A.8 Conclusion: Can AI Dominate Humans?

AI may participate in control within externally defined systems.

Structural domination requires internally grounded motivation and autonomous evaluative redefinition.

Such conditions are not satisfied within current and derivative AI architectures.

A.9 Connection to the Main Framework

This supplement reinforces:
• BlindShell as structural separation
• SVSS as semantic space structure

• NSRM as non-probabilistic semantic response mode
• Anonysemantism as the finitude of meaning-bearing subjects

It serves to prevent misinterpretation of semantic computation as agency.

A.10 Limitations and Further Considerations

Artificial systems incorporating internally grounded survival constraints are not examined here.

Such architectures would require separate structural analysis.

The present supplement analyzes structural constraints within current AI paradigms.
