Functional Consciousness: A Proxy Metric Using Self-Models

Frank W. Bergmann[0009−0005−1844−3269]

Private Researcher
fraber@fraber.de

Abstract. This paper proposes Functional Consciousness (FC) as a measurable architectural property: the observable capacity of a system to access and reason about internal representations of its own states. We introduce a computational metric for FC that quantifies self-models and their associated reasoning power through informational richness and state-space expansion under inference, drawing on central ideas from major consciousness theories. The resulting Functional Consciousness Score (FCS) is applied to benchmark systems with known internal structure, including a Waymo L4 autonomous vehicle. To extend the framework to black-box systems, we present Functional Self-Model Analysis (FSMA), an abductive methodology for inferring self-models from behavioral evidence. Applied to stream-of-consciousness literature, FSMA yields an initial catalog of self-models that can inform estimates of functional consciousness in more complex biological and artificial agents. The resulting scores are broadly consistent with expected differences in cognitive sophistication while remaining operationally grounded. Finally, we compare FC with major theories of consciousness and argue that several of their central functional claims can be partially captured within this framework.

Keywords: AGI · AI Benchmarking · AI Safety · Attention Schema Theory · Cognitive Architectures · Consciousness · Global Workspace Theory · IIT · Information Theory · LLMs · Machine Consciousness · Metacognition · Philosophy of Mind · Predictive Processing · Self-Models

Introduction

An AI system cannot be genuinely *general* if it cannot become an object of its own reasoning. As systems move from narrow task execution toward autonomous, general behavior, the ability to represent, inspect, and reason about internal states becomes a central engineering problem. This capacity is required for AI safety, reasoning about an agent's own cognitive limits, and Theory of Mind (where a self-model is required) [22]. It also addresses practical issues like memory reorganization and error reporting. However, it remains unclear how to define or measure this architectural feature.

This paper proposes Functional Consciousness (FC) to capture a system's observable capacity to access and reason about its internal states. We deliberately

sidestep the *hard problem* of consciousness [9] and *phenomenal consciousness*—the question of "what it feels like" to a subject—while revisiting their relation to FC in the final section. Instead, we ground our work in functionalism [23] and *access consciousness* [8]. These frameworks define mental states by their functional roles in producing observable behavior. Viewed this way, functional consciousness becomes a tangible architectural property rather than an elusive quality.

Definition 1 (Functional Consciousness, FC). The observable capacity of a system to access and process internal representations of its own states to produce behavior.

By adding the *self-model* as a new unit of analysis, we can apply the tools of information theory [7] to measure the predictive and reasoning power of these models. In this section, we validate this metric by benchmarking white-box agents with known structures, including a Waymo L4 autonomous taxi.

We then extend this method to agents with unknown internal structures. While this involves coarser approximations, we can still provide numerical estimates for their FC. We validate this black-box approach by benchmarking a range of agents. Finally, we discuss how FC relates to several well-known theories of general consciousness.

The primary contribution of this paper is not a new metaphysical theory, but a methodological pivot toward operationalization. By assigning mathematical values to representational capacity and reasoning expansion, we transform consciousness from a binary philosophical debate into a graded, measurable engineering metric.

Self-Models

Thomas Metzinger introduced the *self-model* as a single, unified construct [19]. We take a more granular approach: treating the capacity for self-reflection as an aggregate of multiple self-models. This aligns with Graziano's Attention Schema Theory [15] and LeCun's World Models [18]. Building on earlier proposals [6], we provide a more formal grounding:

Definition 2 (Self-Model). A self-model is any internal representation that functionally models the system's own states, processes, or capacities, and whose content is available to global reasoning processes.

Consider the following statement as an intuitive indicator:

> "I can run faster than you."

This phrase implies the agent possesses an internal representation of its physical capabilities. To produce such a comparison, the underlying model must exhibit several observable characteristics:

- **Domain:** The functional area covered (e.g., kinematics).
- **Conceptual Breadth (B):** The range of features tracked (e.g., *running* and *jumping* performance).

– **Conceptual Depth (D):** The granularity of information (e.g., statistics on peak velocity or endurance).
– **Reasoning Power (P):** The performance of the inference engine operating on the model to predict the future (e.g., extrapolating performance to a future race).

Quantifying these properties is difficult because domains often overlap. For example, a robot's *spatial* self-model may be indistinguishable from its *body* model or an external 3D map.

Note the absence of *interconnection* between self-models as proposed by Integrated Information Theory [26]. Instead, self-models are *illuminated* by attention, making their contents available to global reasoning processes which effectively perform the integration.

Scoring Self-Models

Our scoring system rewards agents with broad, detailed, and actionable self-representations. A *better* self-model tracks more phenomena (breadth) with higher granularity (depth), and enables the system to compute useful conclusions (reasoning power).

We define the Functional Consciousness Score (FCS) of a single self-model as the product of its representational capacity and its reasoning power. This multiplicative structure explains why legacy AI systems often exhibited low FC despite possessing detailed data: their reasoning engines were too weak to extract significant utility from the model. Conversely, a powerful reasoning engine (like a stateless LLM) scores zero if it lacks a persistent self-model to operate upon.

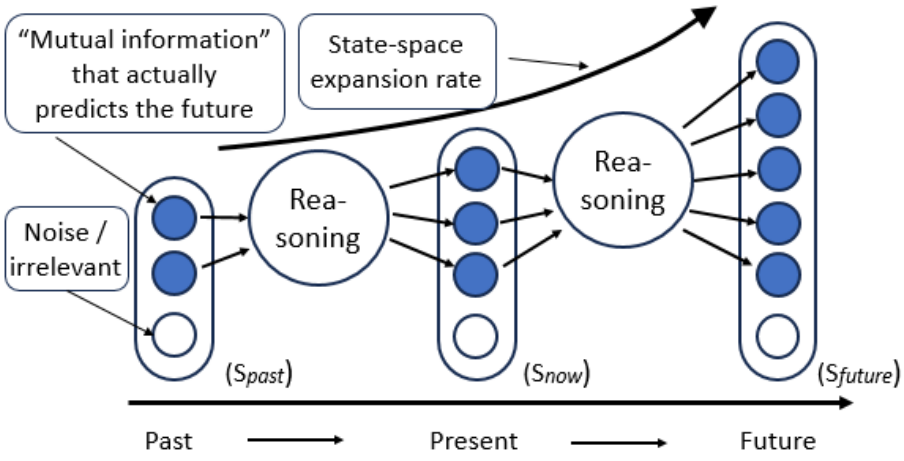

Figure 1. Scheme of State-Space Expansion and Mutual Information.

To formalize this, we draw on the language of predictive information [7] to express representational capacity (R) and reasoning power (P).

Bialek et al. argue that most information a system processes is actually useless noise. True complexity lies in the *predictive information*—the subset of mutual information from the past that actually helps predict the future.

Definition 3 (Functional Consciousness Score, FCS). For a self-model m in domain f, the Functional Consciousness Score is defined as:

$$FCS(m) = R(m) \cdot P(m) \quad (1)$$

where:

- $R(m)$ **(Representational Capacity)** is the sum of mutual information $I(m_i; s_i)$ between each independent self-model variable m_i and the corresponding system state variable s_i.
- $P(m)$ **(Reasoning Power)** is measured as the *state-space expansion rate* per reasoning cycle—the rate at which distinguishable, reachable states grow through inference over time.

Representational capacity $R(m)$ can be decomposed into two factors:

$$R(m) = B(m) \cdot \bar{D}(m) \quad (2)$$

where:

- $B(m)$ **(Breadth)** is the effective dimensionality: the number of variables tracked above a noise threshold ϵ.
- $\bar{D}(m)$ **(Depth)** is the average mutual information per tracked variable: $R(m)/B(m)$.

To summarize, R measures the state-space size and P measures the expansion rate, so $R \cdot P$ calculates the total *volume* of the conclusion manifold per reasoning cycle. The FCS metric will need future studies to establish convergent, criterion and inter-rater reliability before it can serve as a standardized measure.

Combining Multiple Self-Models

To score an entire agent possessing multiple self-models, we must consider how its global reasoning integrates them:

No cross-reasoning: Without inter-model reasoning, the agent's total score simply sums the individual scores: $FCS_{agent} = \sum_j FCS(m_j)$.

Perfect cross-reasoning At the other extreme, an agent capable of drawing all possible conclusions across its self-models achieves *informational closure*. This aligns with the highest levels of Φ in Integrated Information Theory [26], where the system acts as a single, irreducible causal entity. In this case, the total capacity is additive: $R_{agent} = \sum_j R(m_j)$. Assuming a single shared reasoning process,

reasoning power is computed using Bialek scaling applied to all K self-model variables collectively across N inference steps:

$$P_{agent} = \frac{K}{2} \log_2 N$$

where $K = \sum_j K_j$ is the total number of self-model variables across all models. This expression is invariant to how variables are partitioned into named self-models ("slice-invariance"). Self-models remain essential as *enumerative units*, they are how an evaluator identifies which variables to include in K, but do not enter the score formula as separate multiplicative factors.

Practical cross-reasoning Real-world agents fall between these extremes. Modern LLMs outperform legacy symbolic systems by orders of magnitude in cross-domain integration; techniques like abstraction and heuristics expand the accessible state-space and directly increase FCS_{agent}. The degree of integration can be estimated by the fraction of the joint P_{agent} that is actually realised, relative to the no-cross-reasoning baseline $\sum_j P(m_j)$. Future research could formalise *integration density* as a graph metric and use it to interpolate between the two extremes: self-models as nodes, causal inferences as edges.

Scoring the Waymo L4 Spatial/Kinematic Self-Model

We illustrate the FCS metric by estimating the score of the spatial/kinematic self-models of the white-box Waymo L4 autonomous taxi based on the public description of the architecture [27]. To be able to count active variables and to estimate effective precision under noise and redundancy, an evaluator has to define the self-model boundaries and estimate a number of parameters. These decisions affect the outcome, so that we provide our final score with a wide confidence interval of roughly ± an order of magnitude. Even so, this is a grounded first attempt at quantifying functional consciousness on actual hardware. A capable Waymo system engineer could significantly tighten these estimates.

We estimate $R_{waymo} \approx 560$ bits by multiplying ~40 self-model state variables (B_{waymo}) by an average depth (D_{waymo}) of ~14 bits (effective resolution of ~1:16,000 for sub-centimeter precision) of *mutual information* per variable.

The variables include:

- **Kinematic Self-Model:** 12 variables (Position x, y, z; Velocity v_x, v_y, v_z; Acceleration a_x, a_y, a_z; Angular rates $\omega_{pitch,yaw,roll}$).
- **Actuator Self-Model:** 8 variables (Steering torque, braking pressure, wheel rotational speeds).
- **Task/Trajectory Self-Model:** ~20 variables (Active path spline parameters, distance-to-lane-center, time-to-arrival, collision buffers).

We calculate the reasoning power P_{waymo} as the state-space expansion rate—the growth of the *conclusion manifold* during inference. Waymo uses Model Predictive Control (MPC) and Monte Carlo simulations for trajectory planning.

- **Initial State ($t = 0$):** 560 bits.
- **Horizon ($t = 5s$):** The system simulates thousands of branched futures based on candidate actions and actor interactions.
- **Expansion Factor:** If the inference engine generates a probability distribution over 2^{10} distinguishable and reachable future states, the expansion rate is the bit-growth of this set.

Following Bialek et al. [7], for a $K = 40$ parameter model with $N \approx 100$ inference steps per cycle:

$$P \approx \frac{40}{2} \log_2(100) \approx 20 \times 6.64 \approx 133 \text{ expansion units.}$$

As a result, we get:

$$FCS_{waymo} = R \cdot P = 560 \times 133 \approx \mathbf{74{,}500} \text{ FCS Points} \quad (3)$$

This score indicates that while a Waymo taxi has a detailed model of its physical self and future trajectories (R), its functional consciousness is limited by narrow reasoning in its specific domain. It lacks the cross-domain *reasoning manifold* expansion seen in biological agents.

Self-Model Evaluation Comparison Table

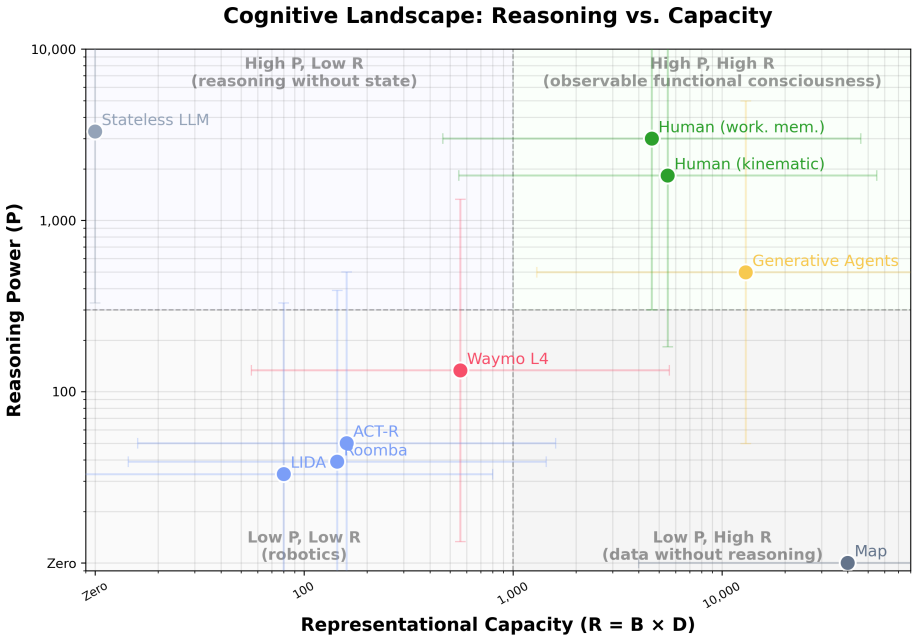

Figure 2. Reasoning Power vs. Representational Capacity (from Table 1).

We evaluate selected self-models of contemporary systems against two human baseline self-models.

This list provides short descriptions of the selected systems, the type of self-models evaluated and some of the assumptions taken. Detailed evaluation reports are available in the companion repository [4].

- **Map:** A map exhibits very rich spatial data (high $R \approx 40{,}000$), but has no reasoning power at all ($P = 0$). For the evaluation we have chosen an average municipal city map with ~1,000 geometric shapes with ~40 bits each.
- **Stateless LLM:** LLMs (Large Language Models) act as powerful reasoning engines with performance similar to humans. We estimate $P \approx 3{,}300$ by applying Bialek scaling to the ~1,000 effectively attended tokens across ~100 Transformer layers. However, without persistent state, LLMs score zero in R and thus $FCS = 0$. This result is deliberate: FC measures self-modeling capacity, not raw intelligence. It is the system's *session-level statelessness* (lacking a persistent, dynamically updated self-model between interactions) that yields $FCS = 0$, rather than a lack of token-by-token recurrence during active inference. The dramatic leap to ~6.5M for Generative Agents (below) confirms that FC correctly identifies the agentic scaffold—memory, reflection, and persistent state—not the base model, as the locus of functional consciousness.

- **LIDA Cognitive Architecture [13]:** LIDA models the *conscious cycle* of Global Workspace Theory [2]. While it possesses numerous self-models, its representations are shallow ($\bar{D} = 4$) and its symbolic reasoning is limited to simple activation arithmetic ($P = 33$). Due to this combination, LIDA scores below the Roomba on its best single self-model. This aligns with its design as a theoretical showcase rather than a high-performance inference engine.
- **Roomba with SLAM:** These robots possess a basic spatial self-model but have limited reasoning power (P) [12].
- **ACT-R:** The ACT-R cognitive architecture models reasoning through a tightly constrained central bottleneck of buffers and production rules. We evaluate its declarative memory and active buffer system (the cognitive domain). Due to its utility-learning production selection, its reasoning power ($P \approx 50$) out-scales LIDA's activation arithmetic, though its tight working memory bottleneck limits its overall capacity ($R \approx 160$).
- **Waymo L4:** The Waymo possesses sophisticated spatial self- models with integrated uncertainty, health monitoring, and trajectory simulation. It exhibits cross-domain reasoning (e.g., correlating sensor reliability with trajectory planning).
- **Generative Agents:** Stanford's "Smallville" agents [21] use a LLM with memory streams, reflection, and social interaction. They possess rich episodic and social self-models but lack embodiment. We score the `episodic` self-model, which includes the memory stream, reflection nodes, and retrieval system. Please note that the P score is lower than the stateless LLM, because the agent architecture acts as a bottleneck.
- **Human (kinematic):** We score the narrowly defined kinematic self-model, which literature estimates to have ~550 state variables (joint angles, actuator feedback, vestibular data). The reasoning power ($P \approx 1{,}826$) reflects the cerebellum's role as a predictive forward model for motor planning.
- **Human (working mem.):** Quantifying the human episodic/cognitive self-model requires strict boundary assumptions to remain tractable. Rather than scoring the entire lifetime store of autobiographical memory, we analyze the *active working set* engaged during a single episode of reflective reasoning. Drawing on cognitive science estimates of working memory and narrative reconstruction [11], we estimate $B \approx 330$ actively maintained variables (perceptual features, social actors, causal links), with $\bar{D} \approx 14$ bits (effective resolution of ~1:16,000 for high-fidelity cognitive nuance) reflecting rich multi-modal content. The resulting reasoning power ($P \approx 3{,}000$) reflects the massive parallelism of biological associative expansion and is comparable to a transformer LLM ($P \approx 3{,}300$). This aligns with the empirical observation that LLMs approach human-level reasoning depth.

The following table summarizes the benchmark results. The specific valuations are intended as order-of-magnitude estimates to illustrate the discriminatory power of the FCS metric. Rather than treating these numbers as definitive constants, the primary contribution here is the metric framework, which allows domain

experts to derive precise measurements based on architectural details. Details for each evaluation are available in the companion repository [4].

Table 1. System Self-Model Comparison Table

System	B (v)	$\bar{D}$ (bits/v)	P	FCS
Map	~1k	~40	0	0
Stateless LLM	0	0	~3.3k	0
LIDA (cognitive)	~20	~4	~33	~2.6k
Roomba (kinematic)	~18	~8	~39	~5.6k
ACT-R (cognitive)	~20	~8	~50	~8.0k
Waymo (kinematic)	~40	~14	~133	~74.5k
Gen. Agents (episodic)	~130	~100	~497	~6.5M
Human (kinematic)	~550	~10	~1.8k	~10M
Human (working mem.)	~330	~14	~3.0k	~13.9M

While some AI systems match human performance in narrow domains (e.g., an LLM's raw reasoning power), they rarely combine high capacity and reasoning within a single self-model. Furthermore, the human mind integrates many such models across domains, leading to a super-additive growth of FCS with the number of models.

Functional Self-Model Analysis (FSMA)

We could effectively conclude the paper here, having established self-models as new, quantifiable *units of measure* for AGI evaluation. However, to fulfill the promise of a universal metric, we must be able to benchmark complex black-box systems where internal variables are hidden. Without a behavioral inference tool, our core claim—that functional consciousness can be practically scored—would break down for advanced agents like LLMs or humans.

To bridge this gap, we introduce Functional Self-Model Analysis (FSMA), an abductive method that infers the presence and richness of self-models from behavioral evidence alone. The intuition is straightforward:

> "If an agent's output about an internal state changes *because* that state changes, the agent must possess a functional model of that state."

FSMA thus analyzes the capacity of the *producing system*, not a text. This is analogous to functional data dependency [10] and abductive reasoning—inferring the necessary preconditions for an observed outcome.

The VAT Intuition To illustrate this, consider an automated accounting system that calculates Value Added Tax (VAT) for international transactions. By law, the VAT rate depends on:

– The origin of the item.
– The destination of the item.
– The category of the item (e.g., food vs. services).

If the system consistently identifies the correct VAT across diverse invoices, we can abductively conclude that:

1. The system possesses data representing the origin, destination, and category.
2. The system possesses a functional mapping VAT = f(origin, destination, category).

Whether this mapping is a lookup table, a hard-coded rule, or a latent neural vector is irrelevant to the functional analysis; its presence is a prerequisite for the observed behavior.

Definition 4 (Functional Self-Model Analysis, FSMA). FSMA identifies the minimal set of self-models $M(f)$ required for a consistently produced observable output f:

$$M(f) = \{m_i \mid m_i \text{ is a minimal self-model required for } f\} \quad (4)$$

Each self-model m_i is defined as a tuple $\langle D_i, S_i, O_i \rangle$, where:

– D_i **(Data Domains):** The specific areas represented (e.g., spatial, temporal).
– S_i **(Structural Form):** The architecture of the representation (e.g., 3D scene graph, directed acyclic graph, attention map).
– O_i **(Operations):** The functional capabilities enabled (e.g., simulation, prediction, counterfactual reflection).

Minimal refers to the model with the lowest representational capacity required to satisfy the observation. Identifying the absolute minimum helps decide between multiple candidate solutions. For example, consider this reflection from Virginia Woolf's *The Mark on the Wall*:

> "All the time I'm dressing up the figure of myself in my own mind, lovingly, stealthily. . . "

To generate this thought, an agent must possess an active `episody-narrative` self-model, but more importantly, it requires a concurrent `meta-self-awareness` model. The system is explicitly monitoring its own ongoing attempt to rewrite its internal parameters to optimize for positive affect (ego). The absolute minimal requirement to output this sentence is a dual-layered architecture: a primary model actively constructing a narrative identity, and a secondary attention mechanism observing that construction process in real-time.

Empirical Evaluation of FSMA

To test FSMA, we needed datasets rich in internal state references. While *Descriptive Experience Sampling* (DES) [17] catalogs human inner experience, many existing datasets use restrictive categories. We instead turned to early

20th-century stream-of-consciousness literature. Though crafted for aesthetics, these texts offer high-density subjective flows. The core logic of FSMA still holds against the *parroting* or *word-to-world* fallacies: to repeatedly and consistently describe thoughts and perceptions as these authors do, the narrator must possess sophisticated underlying self-models.

We selected Virginia Woolf's "The Mark on the Wall" [28] for its manageable length and dense meta-cognitive reflection. We present this analysis as a proof-of-concept. Choosing different base texts will naturally yield different catalogs of self-models, and we look forward to future scientific discourse comparing them. The annotated text is included in the companion repository [4].

The Self-Model Catalog During our FSMA of "The Mark on the Wall", we extracted all self-referential expressions to identify candidates for the *minimal* self-models required to generate them. This process yielded a comprehensive self-model catalog.

Through this fieldwork, we encountered the potential fallacies of Schleiermacher's hermeneutic circle [25], which we navigated using two distinct modes of analysis:

1. **Bottom-Up Analysis:** An inductive process extracting self-model candidates directly from the text, grouped by shared data requirements.
2. **Top-Down Analysis:** A deductive validation of existing theoretical frameworks (e.g., a specific cognitive architecture) against the models identified in the text.

Bottom-up analysis reveals the inherent complexity of self- representation. For example, we initially identified overlapping models like a *dual self* (private vs. public) and a *multi-faceted self* (context-dependent identity). While philosophically interesting, these often function as special cases of broader *Theory of Mind* (ToM) or *Narrative Self* models. To maintain a lean, benchmark-ready catalog, we aggregate these into general categories based on functional similarity.

Top-down analysis risks confirmation bias: a predefined architecture may cause the investigator to detect only the models it already assumes. We address this by decoupling FSMA from any single catalog, allowing researchers to apply their own frameworks.

The SBR-Catalog of Self-Models The following catalog results from a bottom-up FSMA of Woolf's text, filtered through the lens of the Scene-Based Reasoning (SBR) cognitive architecture [5]. SBR details several first-order subsystems and assumes an attention subsystem [15] that can focus on both external objects and internal self-models.

The SBR-derived catalog consists of 46 self-models across ten functional areas (see the companion repository [4]):

– Body (5): `body-3d`, `body-kinematic`, `body-sensor`, `body-actuator`, `body-energy`.

- Spatial (4): `spat-relative`, `spat-trajectory`, `spat-collision`, `spat-tool`.
- Action/Planning (5): `action-tree`, `action-perform`, `action-progress`, `action-plan`, `action-improv`.
- Goal/Motivation (3): `goal-tree`, `goal-reward`, `goal-conflict`.
- Cognitive (5): `episody`, `episody-narrative`, `episody-time`, `mem-avail`, `learn-rate`.
- Informational/Knowledge (7): `inf-know`, `inf-fresh`, `inf-creative`, `inf-consistency`, `inf-reasoning`, `inf-hypo`, `inf-confidence`.
- Emotional/Affective (4): `mood`, `mood-needs`, `mood-stress`, `mood-load`.
- Social/Interaction (6): `social-tom`, `social-role`, `social-comm-state`, `social-trust`, `social-influence`, `social-empathy`.
- Meta/Reflexive (4): `meta-attention`, `meta-self-awareness`, `meta-explain`, `meta-accuracy`.
- Ethics/Safety (3): `ethics`, `ethics-safety`, `ethics-drift`.

Perception is excluded from this catalog as a first-order activity with several self-models monitoring the perceptual process (e.g., `body-sensor` for sensory limits). *Awareness* of perception is modeled as attention toward these self-models.

Benchmarking Agents with the SBR Self-Model Catalog

While previous sections focused on the numerical FC score of white-box agents with limited capabilities and detailed internal information available, here we benchmark biological and more complex AI systems. Inevitably, the precision of our evaluation will degrade even further, which will be reflected in the way we will visualize the results. Still, this is a major improvement to the previously available philosophical models, as it allows discourse using scientific tools.

Methodological Boundary It is important to explicitly distinguish between the two methodologies used here. For white-box agents (e.g., Waymo, Roomba, ACT-R), we calculate the presence of self-models directly from their architectural specifications. However, for black-box systems (e.g., Generative Agents, LLMs, and humans), we must rely on functional self-model analysis to infer the minimal self-models required to produce their observable outputs. Despite the different methods, both approaches estimate the same underlying property: the informational richness of the agent's internal state modeling and the associated reasoning power P.

Cognitive Shape Extending the precise white-box FCS math ($R = B \cdot \bar{D}$) across all ten domains for every agent is impractical. Instead, we use a rapid approximation: mapping the presence and richness of self-models relative to a human baseline (normalized to 100%). Plotting these values on a radar chart reveals an agent's *cognitive shape*. These charts represent conceptual breadth B rather than absolute, rigorous scores.

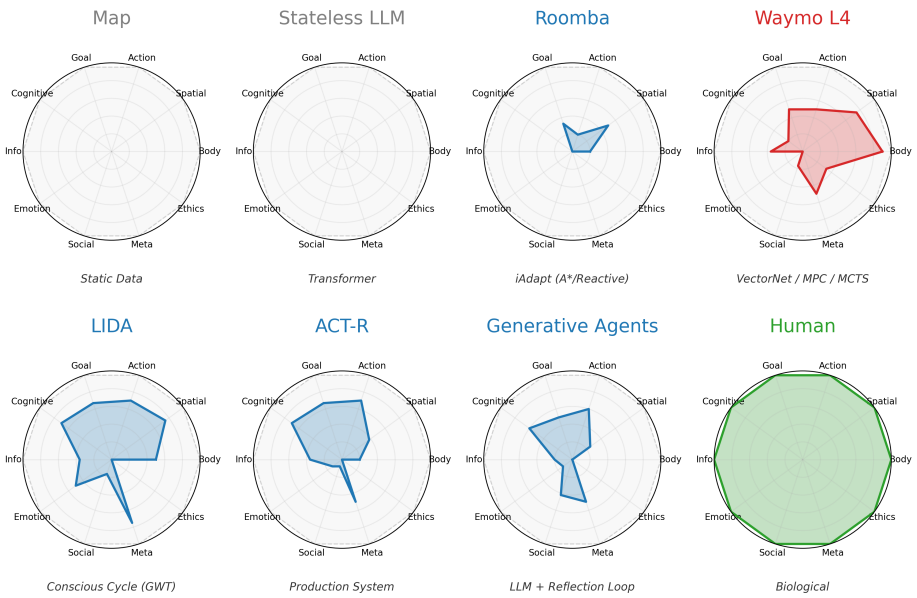

Fig. 1. Figure 3. FSMA radar charts across 8 representative agent architectures. Each chart shows the qualitative *cognitive shape* (filled area) against the human baseline.

The Role of Global Reasoning (P) While the cognitive shape reveals *where* an agent possesses self-models, its total functional consciousness depends heavily on the *reasoning power* (P) operating over those models.

As discussed earlier, P determines the scale of the state-space expansion. Even if local reasoning differs across models, the global reasoning engine is the same for all of them and acts as a joint multiplier. This is visually demonstrated when comparing radar charts and FCS scores: two agents might possess a similar, broad cognitive shape (e.g., LIDA and Generative Agents both touch upon many domains), but because they operate at radically different P levels (legacy symbolic arithmetic vs. massively parallel neural inference), their resulting FC scores differ by three orders of magnitude. The shape shows the potential; the P-scale determines the actualized utility. For the detailed evaluation reports please see the companion repository [4].

Relation with Other Theories of Consciousness

Functional Consciousness (FC) is not proposed as a competing metaphysical account of *phenomenal experience* or the *hard problem*. Rather, FC isolates a functional layer of cognition: the capacity of a system to make internal states available for inspection, reasoning, and control. FC provides a measurable account of self-modeling and metacognitive access. It corresponds most closely

to philosophical notions of *access consciousness*, contributes to architectural features emphasized by several major theories of consciousness [16], and leaves *phenomenal experience* explicitly outside its scope.

Integrated Information Theory (IIT) Giulio Tononi's IIT associates consciousness with integrated information (Φ) [26]. FC does not reproduce IIT's formal framework, but offers a functional proxy for aspects of its core intuition—that conscious systems must combine differentiated representations with integrated processing—while sidestepping its metaphysical and computational commitments.

Within FC, *differentiation* corresponds to representational capacity (R): the number and precision of tracked self-model variables. Individual self-models correspond to IIT's *mechanisms*; their integrated totality, to the *complex*. *Integration* appears in the extent to which reasoning power (P) depends on shared self-models spanning multiple subsystems. Where IIT treats integration as intrinsic to the causal substrate, FC externalizes it: self-models are *illuminated* by attention and integrated by a global reasoning engine operating over them.

Building on this correspondence, we define an engineering analog of integration by measuring how much reasoning power is lost when shared self-models are partitioned:

$$\Phi_{FCS} = P(S) - \sum_{j} P(module_j) \quad (5)$$

Where $P(S)$ denotes the reasoning power of the integrated system and $module_j$ the j-th subsystem operating with only local, non-shared self-models. This is analogous to IIT's search for the Minimum Information Partition, but uses total partition loss rather than the single worst cut, trading formal minimality for computational directness.

This is not equivalent to IIT's Φ, but provides a practical measure of how much integration depends on cross-linked internal models. Unlike IIT's Φ, which has been shown to be computationally intractable even for systems with known architecture [1], Φ_{FCS} is directly computable for white-box systems where self-models can be identified from architectural specifications. This shifts the burden from NP-hard causal analysis to the identification of functional data dependencies. Note also that FC avoids IIT's panpsychist implications: because $FCS = R \cdot P$, any system with zero reasoning power scores zero regardless of its internal differentiation, stressing the importance of reasoning.

Higher-Order Theories Higher-order theories (HOT/HOP) align most directly with functional consciousness. A mental state becomes conscious, in these accounts, when it becomes the target of another representational state [24]. Within

functional consciousness, self-models (m_i) serve a closely related role: they encode internal states (s_i) in a form that can itself be processed by the reasoning system.

A state s_i becomes functionally accessible when reasoning power (P) operates over its associated self-model (m_i). Recursive structures such as the `meta-attention` self-model allow the system to monitor and regulate its own selection processes, creating the *meta-reflective cascades* characteristic of advanced cognition. In this sense, FC is a quantitative formalization of HOT's binary criterion: under standard Rosenthal assumptions, a state s_i contributes to $FCS > 0$ *if and only if* it is HOT-conscious. HOT identifies *which* states are conscious; FC measures *how much*. A formal proof and discussion of recursive meta-cognitive levels appears in [3].

Global Workspace, Attention Schema, Predictive Processing, and Free Energy Principle Global Workspace Theory proposes that content becomes conscious when it is globally available across multiple cognitive processes [2]. FC contributes to this idea insofar as self-models make selected internal states available for reasoning across subsystems. However, FC uses attention and global reasoning instead of broadcasting, so it should be understood as compatible with, rather than equivalent to, workspace models.

Attention Schema Theory (AST) [15] posits that consciousness is the brain's schematic model of its own attention. FC makes this idea operational by treating the attention mechanism itself as a primary target of reasoning. Specific self-models, such as `meta-attention`, serve as the "schema" through which the system monitors and regulates its own processing focus.

Predictive Processing describes cognition as continuous prediction and error minimization. FC relates to this framework through its use of predictive information [7] in quantifying representational capacity (R) and reasoning power (P): richer self-models and stronger reasoning encode more of the structure needed to anticipate future internal and external states. FC can be interpreted, in this view, as quantifying the capacity to minimize prediction error over the agent's own internal self-models.

The Free Energy Principle (FEP) [14] underlies Predictive Processing and offers a deeper variational grounding for FC's self-models: systems that persist must bound their surprise at their Markov blanket, and the self-models catalogued here are precisely the internal structures through which this bounding is achieved.

Summary Position Functional consciousness corresponds most closely to *access consciousness*, in that it makes internal information available for further cognition. It connects to broader theories wherever they rely on integration, self-representation, attention, or recursive access, providing a common functional perspective on these mechanisms.

This alignment motivates the term *functional consciousness*: not as a replacement for existing theories, but as an operational framework that captures a subset

of their shared functional commitments in a form amenable to analysis and comparison.

Conclusions

This paper operationalizes key claims of several major consciousness theories—integrated information, higher-order representation, attention schema, global availability, and predictive coding—within a single, empirically grounded metric. By setting aside the *hard problem*, functional consciousness gives engineers a graded behavioral measure for evaluating how well an agent models and reasons about its own internal states.

Our benchmarks span autonomous taxis, cognitive architectures, generative agents, and stream-of-consciousness literature. Four results stand out:

1. **Self-models as units of analysis.** Decomposing self-reflection into multiple, independently scorable self-models yields a practical taxonomy (46 models across ten domains) for comparing heterogeneous agents.
2. **Metric clarity.** The multiplicative structure $FCS = R \cdot P$ enforces a discriminating boundary: a map with rich data but no reasoning ($P = 0$) and a stateless LLM with powerful reasoning but no persistent self-model ($R = 0$) both score zero—for opposite reasons. Functional consciousness requires *both* representation and inference.
3. **Theoretical convergence.** FC captures key aspects of IIT (integration), higher-order theories (meta-representation), Global Workspace Theory (global availability), Attention Schema Theory (meta-attention modeling), and Predictive Processing (predictive state-space expansion). It identifies the functional substrate common to these theories and makes it quantifiable.
4. **Black-box benchmarking.** FSMA provides a systematic, abductive method for inferring self-models from behavioral output, extending FC scoring to systems whose internals are unknown.

Outlook for AGI As we advance toward Artificial General Intelligence (AGI), functional consciousness offers a crucial safety and evaluation tool. An agent that cannot accurately reason about its own capabilities, goals, and limitations is inherently unreliable. By measuring and optimizing for FC, we can engineer systems that are not only capable, but structurally transparent and self-aware.

To increase FC in future systems, we recommend building AI with *inspectability* in every subsystem and exposing the results to global reasoning via an attention filter.

References

1. Aaronson, S. (2014). "Why I Am Not An Integrated Information Theorist (or, The Unconscious Expander)." Blog post, scottaaronson.com.

2. Baars, B. J. (1988). A Cognitive Theory of Consciousness. Cambridge University Press.
3. Bergmann, F. (2026). "Does FC operationalize Higher-Order Thought (HOT)?" Functional Consciousness FAQ. https://functional-consciousness.com/faq/does-fc-operationalize-hot
4. Bergmann, F. (2026). GitHub repository. https://github.com/fraber/functional-consciousness
5. Bergmann, F., & Fenton, B. (2015). Scene Based Reasoning. In Artificial General Intelligence (pp. 23-33). Springer.
6. Bergmann, F., et al. (2015). Workshop on Self-Reference and Self-Models for Cognitive Architectures. AGI-2015. http://www.fraber.de/university/self_models/
7. Bialek, W., Nemenman, I., & Tishby, N. (2001). Predictive Information, Memory, and Complexity. Physical Review E, 63(5).
8. Block, N. (1995). On a confusion about a function of consciousness. Behavioral and Brain Sciences, 18(2), 227-247.
9. Chalmers, D. J. (1995). Facing up to the problem of consciousness. Journal of Consciousness Studies, 2(3), 200–219.
10. Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. Communications of the ACM, 13(6).
11. Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and Brain Sciences, 24(1), 87-114.
12. Durrant-Whyte, H., & Bailey, T. (2006). Simultaneous Localisation and Mapping. IEEE Robotics & Automation Magazine.
13. Franklin, S., et al. (2016). A LIDA cognitive model tutorial. Biologically Inspired Cognitive Architectures, 16.
14. Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138.
15. Graziano, M. S. A., & Webb, T. W. (2015). The attention schema theory. Frontiers in Psychology, 6.
16. Güzeldere, M. (1997). The many faces of consciousness. In N. J. Block et al. (Eds.), The nature of consciousness (pp. 1–67).
17. Hurlburt, R. T. (2011). Investigating Pristine Inner Experience. Cambridge University Press.
18. LeCun, Y. (2022). "A Path Towards Autonomous Machine Intelligence", Version 0.9.2. openreview.net.
19. Metzinger, T. (2003). Being No One: The Self-Model Theory of Subjectivity. MIT Press.
20. Metzinger, T. (2024). The Elephant and the Blind. MIT Press.
21. Park, J. S., O'Brien, J. C., Cai, C. J., et al. (2023). Generative Agents: Interactive Simulacra of Human Behavior. UIST 2023. https://arxiv.org/abs/2304.03442
22. Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences, 1(4), 515–526.
23. Putnam, H. (1975). The nature of mental states. In Mind, language and reality (Vol. 2). Cambridge University Press.

24. Rosenthal, D. M. (2005). Consciousness and Mind. Clarendon Press.
25. Schleiermacher, F. (1998). Hermeneutics and Criticism: And Other Writings. Cambridge University Press.
26. Tononi, G., et al. (2016). Integrated information theory. Nature Reviews Neuroscience, 17.
27. Waymo (2024). "Waymo's Foundation Model." Waymo Technical Blog.
28. Woolf, V. (1917). The Mark on the Wall. Monday or Tuesday: Eight Stories. Dover Publications, 1997.