Immortalitas in Hac Vita
Longevity Platform
https://www.longevity.ge
Longevity Horizon
ISSN:3088-4063
2026
Vol.2 No.2:1

Mathematical formalism of Ze

Two free energies, interference, and localization

Jaba Tkemaladze [^] [1]
Affiliation: [1] Longevity Clinic, Inc, Georgia
Citation: Tkemaladze, J. (2026). Mathematical formalism of Ze, 2(1). Longevity Horizon, 2(1). DOI : https://doi.org/10.65649/kzj86888

Abstract

This article introduces and formalizes Ze, a novel theoretical framework for cognitive architecture and autonomous systems. Ze posits that advanced intelligence requires the maintenance of two distinct, asymmetric generative models of the same environment: a causal (forward) model $\mathcal{M}$_A and a counterfactual (inverse) model $\mathcal{M}$_B. Each model minimizes its own variational free energy ($\mathcal{F}$_A), ($\mathcal{F}$_B), and their interaction dynamics define core cognitive processes. A key emergent quantity is the model conflict $\Delta\mathcal{F}$ = |$\mathcal{F}$_A − $\mathcal{F}$_B|, which regulates a phase transition between two fundamental regimes: an *interference regime* (characterized by low posterior divergence ($\mathcal{I}$ ≈ 0) where model outputs are constructively fused, and a *localization regime* ($\Delta\mathcal{F}$ > θ) where the system commits to a single resolved interpretation ŝ. The framework is extended to include active action selection from model-specific policies, a mechanism for representational growth via "which-path" information, and a "quantum eraser" operator for strategic simplification. We demonstrate that this architecture establishes a strict formal isomorphism with quantum measurement phenomena, notably the double-slit experiment, but is grounded entirely in classical variational inference. The theory reinterprets cognitive "collapse" not as a postulate but as an optimization-driven phase transition and yields the key testable prediction that active, alternating intervention accelerates localization compared to passive observation.

Keywords: Cognitive Architecture, Variational Inference, Model Conflict, Active Inference, Ze

Variables and Basic Structure

This section introduces the core mathematical framework of Ze, a proposed architecture for autonomous systems that necessitates the parallel maintenance of two distinct yet complementary world models. The central premise is that intelligent, adaptive behavior in partially observable environments requires more than a single generative model of sensory inputs. We posit that a dual-model architecture, comprising a causal (*forward*) model and a counterfactual (*inverse*) model, provides a more robust substrate for state estimation, planning, and explanation generation.

Consider an agent embedded in an environment, receiving a stream of potentially ambiguous, high-dimensional sensory data over time. This stream is denoted as the observation sequence o_{1:T} = (o_1, o_2, ..., o_T), where each o_t belongs to an observation space $\mathcal{O}$. The agent's fundamental challenge is to infer the latent, causal structure of the environment from this sensory flow to guide its actions (Hassabis et al., 2017; Lake et al., 2017).

Within Ze, this challenge is addressed by instantiating two separate generative models of the same underlying environment. These models share the goal of explaining the observations but adopt fundamentally different temporal and causal perspectives.

- **The Causal (Forward) Model $\mathcal{M}$_A**: This model embodies the standard perspective of temporal generative models in cognitive science and machine learning (Friston, 2010; Rao & Ballard, 1999). It formalizes the understanding of "how the world evolves." Its dynamics are governed by a causal, forward-in-time progression. Conceptually, $\mathcal{M}$_A answers the question: *"Given the current state of the world and my action, what are the probable next states and subsequent observations?"* This mirrors the predictive processing and active inference frameworks, where an internal model generates top-down predictions to be matched against bottom-up sensory evidence (Clark, 2013; Friston, 2005).
- **The Counterfactual (Inverse) Model $\mathcal{M}$_B**: This model represents the novel, complementary component of the Ze architecture. It operates with a counterfactual or retrospective logic. Its dynamics are not strictly bound to forward causality but are structured to infer "what must have been" to explain the present. It answers a different question: *"Given the current sensory state, what past states or alternative causal trajectories could have plausibly led to it?"* This mirrors reasoning processes involved in explanation, fault diagnosis, and understanding alternative possibilities (Gerstenberg et al., 2021; Pearl, 2009). The "inverse" label here refers not merely to inverting a function, but to inverting the direction of causal inquiry.

Each generative model $\mathcal{M}$_X (where X ∈ {A, B}) maintains its own set of latent, or hidden, states. These states represent the model's internal belief about the environment's configuration from its specific perspective. We denote these as:

$$s^A_t \in \mathcal{S}^A, \quad s^B_t \in \mathcal{S}^B$$

where $\mathcal{S}^A$ and $\mathcal{S}^B$ are the respective state spaces. Crucially, s^A_t and s^B_t are not required to be isomorphic or even of the same dimensionality. s^A_t might encode features relevant to predicting the next observation (e.g., object positions and velocities), while s^B_t might encode features relevant to inferring teleological or abstract causal dependencies (e.g., goals, intentions, or critical events).

Since the true environmental states are hidden, the agent must maintain probabilistic beliefs about them. In accordance with Bayesian brain theories and variational inference approaches (Knill & Pouget, 2004; Dayan et al., 1995), each model in Ze maintains its own *approximate posterior distribution* over its hidden states at each time step. These posteriors represent the agent's belief about the latent state *given all observations up to the current time*. They are formally defined as:

$$q_A(s^A_t) \approx P(s^A_t \mid o_{1:t}), \quad q_B(s^B_t) \approx P(s^B_t \mid o_{1:t})$$

The approximation sign acknowledges that these posteriors are typically intractable to compute exactly and are instead approximated, for instance, by parameterized distributions (e.g., Gaussians) whose parameters are output by a neural network (Kingma & Welling, 2013; Rezende et al., 2014). The process of updating these beliefs recursively as new data o_t arrives constitutes perceptual inference within each model.

The separation of posteriors—q_A and q_B—is a critical feature. It allows the two models to develop and maintain potentially divergent interpretations of the same sensory history. A conflict or tension between q_A and q_B, quantified by measures such as their divergence or the disagreement in their predictions, can be a key signal for triggering attention, exploration, or model update processes, as explored in subsequent sections of this article.

In summary, the foundational structure of Ze is a duplex of generative models: the forward model $\mathcal{M}_A$, with states s^A_t and posterior q_A, which performs causal prediction; and the inverse model $\mathcal{M}_B$, with states s^B_t and posterior q_B, which performs counterfactual explanation. Their co-evolution and interaction in explaining the stream $o_{1:T}$ `` form the basis for the cognitive dynamics proposed by the Ze framework. The following sections will detail the specific parameterization, update rules, and interaction mechanisms between $\mathcal{M}_A$ and $\mathcal{M}_B$.

Two Variational Free Energies

The core perceptual and learning dynamics within the Ze architecture are governed by the principle of variational free energy minimization, a framework widely adopted in neuroscience and machine learning to formalize inference and learning under uncertainty (Friston, 2010; Buckley et al., 2017). However, Ze's duality is instantiated through the maintenance of two distinct variational free energy functionals, each tied to its respective generative model.

For the causal (forward) model $\mathcal{M}$_A, we define its variational free energy as:

$$\mathcal{F}\_A(o, q\_A) = E\_{q\_A(s^A)} [\ln q\_A(s^A) - \ln p(o, s^A \mid \mathcal{M}\_A)]$$

Analogously, for the counterfactual (inverse) model $\mathcal{M}$_B, we define:

$$\mathcal{F}\_B(o, q\_B) = E\_{q\_B(s^B)} [\ln q\_B(s^B) - \ln p(o, s^B \mid \mathcal{M}\_B)]$$

These expressions follow the standard formulation of variational inference (Jordan et al., 1999; Blei et al., 2017). Here, p(o, s^X | $\mathcal{M}$_X) represents the joint generative model under $\mathcal{M}$_X, describing how the model assumes observations and its latent states co-occur. The term q_X(s^X) is the approximate posterior, as introduced in Section 1. Critically, the expectation E_{q_X} is taken with respect to the model's own posterior distribution. Mathematically, each free energy $\mathcal{F}$_X provides an upper bound on the negative log evidence (or surprise) - ln p(o | $\mathcal{M}$_X) for its respective model (Beal, 2003; MacKay, 2003). Minimizing $\mathcal{F}$_A with respect to the parameters of q_A corresponds to performing approximate Bayesian inference to identify the most plausible hidden states s^A that explain the observations o under the forward causal assumptions of $\mathcal{M}$_A. Simultaneously, minimizing $\mathcal{F}$_B tunes q_B to perform inference under the counterfactual assumptions of $\mathcal{M}$_B.

It is crucial to emphasize that these two variational free energies are *not* required to be symmetric in time, structure, or complexity. This asymmetry is a foundational design principle of Ze and a key point of departure from architectures employing twin or duplicated models. The generative models p(o, s^A | $\mathcal{M}$_A) and p(o, s^B | $\mathcal{M}$_B) can be factorized according to vastly different graphical structures and temporal dependencies.

The forward model $\mathcal{M}$_A typically assumes a canonical, temporally causal factorization aligned with the arrow of time (Friston et al., 2017). For example:

$$p(o\_{1:T}, s^A\_{1:T} \mid \mathcal{M}\_A) = p(s^A\_1) \prod\_{t=2}^{T} p(s^A\_t \mid s^A\_{t-1}) \prod\_{t=1}^{T} p(o\_t \mid s^A\_t),$$

where p(s^A_t | s^A_{t-1}) is a state transition prior and p(o_t | s^A_t) is a likelihood mapping. Minimizing $\mathcal{F}$_A thus encourages the posterior q_A to recognize states that make the observed sequence likely under this forward chain of causality. This is formally related to state estimation in partially observable Markov decision processes (POMDPs) and sequential variational autoencoders (Chung et al., 2015; Krishnan et al., 2017).

In contrast, the factorization for the inverse model $\mathcal{M}$_B is not constrained to forward temporal causality. It may, for instance, incorporate *backward* dependencies or non-Markovian relationships that emphasize explaining the present by the past or by latent causes (Parr & Friston, 2018). One potential factorization could be:

$p(o_{1:T}, s^B_{1:T} | \mathcal{M}_B) = p(s^B_T) \prod_{t=1}^{T-1} p(s^B_t | s^B_{t+1}, o_{t:T}) \prod_{t=1}^{T} p(o_t | s^B_t)$

where the state transition is conditioned on future states or observational contexts, embodying a form of retrospective or teleological smoothing. Alternatively, $\mathcal{M}_B$ could be structured as a hierarchical model where high-level latent variables z^B generate trajectories of lower-level states s^B, emphasizing abstract causes over detailed dynamics (Sohn et al., 2015). Crucially, $\mathcal{F}_B$ is minimized under this set of structural assumptions, which may posit that the present is best explained by goals, final causes, or counterfactual alternatives, as explored in models of planning and intention inference (Baker et al., 2017; Botvinick & Toussaint, 2012).

This structural asymmetry implies that the two free energies measure "surprise" or prediction error relative to fundamentally different generative world models. $\mathcal{F}_A$ quantifies how surprising the data is under a model of physical, forward dynamics. $\mathcal{F}_B$ quantifies how surprising the same data is under a model of narrative, teleological, or explanatory coherence. Their minimization leads to the emergence of two distinct, co-existing interpretations of the sensory stream.

Furthermore, the timescales of minimization can differ. $\mathcal{F}_A$ is often minimized rapidly for online, real-time filtering (e.g., updating a belief about an object's current position). The minimization of $\mathcal{F}_B$ may operate on a slower timescale, integrating evidence over longer episodes to infer stable goals or contextual narratives (Hasson et al., 2015). The models may also differ in representational granularity; $\mathcal{M}_A$ might operate on fine-grained sensorimotor variables, while $\mathcal{M}_B$ might operate on more symbolic or abstract variables (Lake et al., 2017).

In summary, the Ze architecture is defined not by a single optimization objective, but by the parallel minimization of two asymmetric variational free energies, $\mathcal{F}_A$ and $\mathcal{F}_B$. This process maintains two separate, probabilistically coherent interpretations of experience: one causal-forward and one counterfactual-inverse. Their interaction, competition, and integration—mediated by a third, overarching principle—form the basis for advanced cognitive functions and will be addressed in the following section on the Meta-Energy G and the **Principle of Collaborative Dissonance**.

The Model Conflict (The Core Quantity of Ze)

The parallel maintenance and independent minimization of two distinct variational free energies, $\mathcal{F}_A$ and $\mathcal{F}_B$, give rise to a crucial, emergent dynamical variable within the Ze architecture: the *model conflict* or *interpretation divergence*. This quantity, central to Ze's proposed cognitive dynamics, is defined as the absolute difference between the two free energies:

$$\Delta\mathcal{F} = | \mathcal{F}_A(o, q_A) - \mathcal{F}_B(o, q_B) |$$

Formally, $\Delta\mathcal{F}$ quantifies the disparity between the perceptual "surprise" experienced by the causal model $\mathcal{M}_A$ and that experienced by the counterfactual model $\mathcal{M}_B$ when confronted with the same sensory data o. It is not a directly observable sensory signal but a *structural or*

meta-cognitive variable that emerges from the internal processing architecture (Fleming & Daw, 2017; Shea et al., 2014). Its magnitude dictates the system's mode of operation, governing the nature of interaction between the two generative streams.

The central postulate of Ze is that $\Delta\mathcal{F}$ regulates a continuum between two fundamental cognitive regimes: **Interference** and **Localization**.

- **Low $\Delta\mathcal{F}$: The Regime of Permissible Interference**. When the two free energies are of comparable magnitude ($\Delta\mathcal{F}$ is small), it indicates a state of consensus or alignment between the two models' interpretations. Both the forward causal narrative and the inverse counterfactual explanation converge on a similarly plausible account of the sensory data. In this regime, the system permits and even encourages interference—not in the disruptive sense, but in the constructive sense of wave interference in physics. Here, the posterior distributions q_A and q_B, or their predictions, can be blended, averaged, or allowed to interact synergistically (Buschman & Miller, 2007; Heeger, 2017). This interaction can lead to enriched, multi-faceted representations. For instance, the forward model's estimate of physical object location can be refined by the inverse model's inference about the object's goal, and vice-versa, leading to a robust, integrated percept. This regime is characteristic of routine perception in predictable, coherent environments where sensory evidence strongly supports a single, unified interpretation. It aligns with theories of "explanation-based" perception where prior knowledge seamlessly informs sensory processing (Kersten et al., 2004).
- **High $\Delta\mathcal{F}$: The Regime of Required Localization**. A large value of $\Delta\mathcal{F}$ signals a fundamental dissonance between the two models. One model finds the sensory data relatively unsurprising and coherent (low free energy), while the other finds it highly surprising and incoherent (high free energy). This indicates an ambiguous, novel, or contradictory situation—such as an unexpected event, a perceptual illusion, or a violation of normative assumptions (Lieder et al., 2018). In this regime, the architecture triggers a **localization** process. The term is used here in its computational sense, akin to fault localization in systems engineering: the system must identify *which model* (or which component within a model) is the source of the conflict and *where in the data stream* the discrepancy arises (Liang et al., 2018; Sajid et al., 2021).

Localization involves several key operations:

1. **Source Attribution**: Determining whether the conflict stems from a failure in the forward causal prediction ($\mathcal{F}$_A) is high) or from the failure to find a plausible counterfactual explanation ($\mathcal{F}$_B is high). Is the world violating physical laws, or is it violating narrative/teleological expectations?
2. **Temporal Isolation**: Identifying the specific time steps or episodes where the predictions of $\mathcal{M}$_A and $\mathcal{M}$_B begin to diverge significantly. This is analogous to identifying a "change point" or an "anomaly" from a multi-model perspective (Wilson et al., 2010).
3. **Focused Attention and Exploration**: Allocating processing resources (e.g., precision weighting in predictive coding) to the conflicting aspects of the sensory input or to the

model components in error (Feldman & Friston, 2010; Mirza et al., 2016). This may also drive targeted epistemic exploration to gather disambiguating data.

The outcome of localization is not necessarily to force the models back into agreement. Instead, it can lead to several adaptive responses: rapid online updating of the more uncertain model's parameters, the gating of one model's output in favor of the other for downstream decision-making (a form of model selection), or the initiation of deliberate reasoning to resolve the paradox (Findling et al., 2023). Crucially, a persistently high $\Delta\mathcal{F}$ can signal a genuine, irreducible ambiguity in the environment, prompting the system to maintain multiple competing interpretations—a state related to holding "hypotheses" in mind (Vul et al., 2014).

It is vital to reiterate that $\Delta\mathcal{F}$ is a structural, internally computed variable, not an external observable. It is a second-order measure that reports on the consistency of the system's own first-order inferences. This places it within the theoretical realm of meta-cognition and confidence computation (Meyniel et al., 2015; Pouget et al., 2016). Unlike a simple prediction error signal within a single model, $\Delta\mathcal{F}$ is a *conflict signal between two different kinds of prediction errors*.

In conclusion, the model conflict $\Delta\mathcal{F}$ serves as the core control variable of the Ze architecture. By monitoring the divergence between the variational free energies of its dual generative models, the system can fluidly alternate between a cooperative mode of representational enrichment (low conflict) and a diagnostic mode of focused analysis and model revision (high conflict). This dynamic provides a formal mechanism for balancing perceptual fusion and fission, stability and plasticity, and exploration and exploitation. The subsequent section will formalize how this conflict is managed through a higher-order **Meta-Energy G**.

Interference as Posterior Compatibility

The model conflict $\Delta\mathcal{F}$, as defined in the previous section, governs the *global regime* of the Ze architecture, switching between interference and localization. To formalize the specific mechanics of the **interference regime**, we must define a precise, local measure of compatibility between the two generative models. This measure quantifies the degree to which their internal, probabilistic interpretations of the world can be meaningfully combined. We propose that interference, in the Ze framework, is fundamentally about the *compatibility of approximate posterior distributions*.

Let us consider the posterior beliefs q_A(s^A) and q_B(s^B). For interference—the constructive blending of interpretations—to be permissible, these beliefs must refer to, or can be mapped onto, a common latent description. While s^A and s^B may inhabit different state spaces $\mathcal{S}$^A and $\mathcal{S}$^B, we assume the existence of a projection or a common representational subspace. For analytical clarity, we initially consider a scenario where such a mapping allows us to compare distributions over a shared variable s. In practice, this could correspond to a low-dimensional manifold of task-relevant variables (e.g., object identity, spatial location, or goal state) onto which both models project their beliefs (Gallego et al., 2020).

We define the **Interference Measure** $\mathcal{I}$ as the Jensen-Shannon divergence (JSD) between the two posteriors over this common grounding:

$$\mathcal{I} = D\_JS(q\_A(s) || q\_B(s))$$

The Jensen-Shannon divergence is a symmetrized and smoothed version of the Kullback-Leibler (KL) divergence, defined as:

$$D\_JS(P || Q) = 1/2 \ D\_KL(P || M) + 1/2 \ D\_KL(Q || M)$$

where $M = 1/2 \ (P + Q)$ is the mixture distribution (Lin, 1991; Endres & Schindelin, 2003). JSD is bounded between 0 and 1 (for base-2 logarithm) and provides a symmetric, finite metric of distributional similarity.

The value of $\mathcal{I}$ directly dictates the feasibility of interference:

- $\mathcal{I} \approx 0$: **Interference is possible and encouraged**. A near-zero JSD indicates that the two posterior distributions are nearly identical. The causal model $\mathcal{M}$_A and the counterfactual model $\mathcal{M}$_B have arrived at statistically indistinguishable beliefs about the state of the world. In this regime of high compatibility, the outputs of the two models can be seamlessly integrated. This could occur through a weighted averaging of their predictions for downstream processing, a mutual reinforcement of their hidden state estimates, or the formation of a unified posterior that is more precise and confident than either alone—a phenomenon analogous to "veto" or "blessing" in Bayesian sensor fusion (Ernst & Banks, 2002; Clark & Yuille, 1990). This state represents cognitive coherence, where sensory evidence, forward prediction, and retrospective explanation converge.
- $\mathcal{I} \gg 0$: **Interference is suppressed**. A large JSD signifies a significant divergence between the posteriors. The two models are promoting fundamentally different, statistically incompatible interpretations of the same sensory data. In this case, simply averaging their outputs would lead to a nonsensical, maximally uncertain mixture that explains nothing (the mixture distribution Mitself would have high entropy). Therefore, the architecture must suppress direct interference to prevent representational corruption. Instead, as dictated by a high $\Delta\mathcal{F}$, the system enters the localization regime to diagnose the source of this posterior divergence.

It is critical to disambiguate the term "interference" as used in Ze from its common usage in wave physics. Here, **interference is not a wave phenomenon** but a computational principle of tolerated multiple explanations. It is the system's ability to maintain and exploit a *portfolio of viable generative models* whose beliefs are sufficiently aligned that their combined use is beneficial (Angela & Dayan, 2005). This aligns with theories of "probabilistic population codes" and "distributional coding," where neural populations represent uncertainty distributions, and optimal decoding can combine information from multiple sources (Ma et al., 2006; Pouget et al., 2003).

The dynamics of $\mathcal{I}$ are therefore central to learning and adaptation. During familiar, predictable tasks, $\mathcal{I}$ is kept low through the coupled minimization of $\mathcal{F}$_A and $\mathcal{F}$_B, leading to stable, fused perceptions. When novel or contradictory data is encountered, the inference process may cause q_A and q_B to diverge rapidly, spiking $\mathcal{I}$. This spike acts as a local signal that (a) inhibits fusion pathways and (b) contributes to the global conflict signal $\Delta\mathcal{F}$. The subsequent localization process can then be viewed as an active search for new model parameters or state estimates that will reduce $\mathcal{I}$, thereby restoring the conditions for productive interference under a new, coherent understanding (Solway & Botvinick, 2012).

In summary, the Jensen-Shannon divergence $\mathcal{I}$ between the posteriors of the dual models operationalizes the core Ze concept of interference. It moves the theory from a global regime-switching principle ($\Delta\mathcal{F}$) to a local, computable mechanism governing information fusion. Interference is permitted only under conditions of posterior compatibility (($\mathcal{I} \approx 0$), which signifies a coherent world-model. When posteriors diverge ($\mathcal{I} \gg 0$), fusion is suppressed in favor of diagnostic localization. This formulation provides a rigorous, information-theoretic foundation for understanding how a cognitive system can fluidly alternate between exploiting a unified world-view and investigating its very foundations.

Localization as a Phase Transition

The previous sections established the dual-model architecture of Ze and defined the key quantities governing its dynamics: the global model conflict $\Delta\mathcal{F}$ and the local interference measure $\mathcal{I}$. We now formalize the critical transition into the **localization regime**, which we propose is not a gradual adjustment but a swift, re-organizational shift akin to a *phase transition* in dynamical systems (Haken, 1983; Tognoli & Kelso, 2014). This transition is triggered when the dissonance between the models exceeds a system's tolerance for ambiguity, compelling a focused diagnostic process.

The transition into localization is governed by a **localization threshold** θ. This threshold represents a meta-parameter of the Ze architecture, which may be fixed or adaptively tuned based on context, akin to a decision boundary or an uncertainty tolerance (De Berker et al., 2016). The triggering condition is:

$$\text{If } \Delta\mathcal{F} > \theta \;\Rightarrow\; \text{Localization is triggered.}$$

When $\Delta\mathcal{F} \leq \theta$, the system operates in the interference regime, permitting the blending of posteriors as described in Section 4. However, once the conflict exceeds θ, the architecture undergoes a qualitative change. The cooperative, integrative dynamics are suppressed, and the system enters a state of focused competition and hypothesis testing. This abrupt shift is reminiscent of perceptual transitions in bistable perception or cognitive "aha!" moments, where a new interpretation suddenly dominates (Sandkühler & Bhattacharya, 2008; Kondo et al., 2022).

Formally, we define the core operation of the localization process as a **probabilistic projection**. The system's current, conflicting posteriors q_A(s^A) and q_B(s^B) are used to generate a new,

constrained posterior belief over a shared or reconciled latent space. This operation is denoted as:

$$q(s) \rightarrow q(s \mid \hat{s})$$

where q(s) represents the prior or default distribution in the shared space (often a mixture or a broad distribution), and q(s | ŝ) is a posterior sharply conditioned on a specific, resolved state ŝ. The key is the determination of ŝ.

We posit that ŝ is the latent state that represents the most plausible common ground or "best compromise" between the two conflicting models, given their respective evaluations of the situation. It is identified as the state that minimizes a weighted sum of the two models' free energy functionals, evaluated pointwise or locally. Specifically:

$$\hat{s} = \text{argmin}_{s \in \mathcal{S}} [\, \alpha \mathcal{F}_A(s) + (1 - \alpha) \mathcal{F}_B(s) \,]$$

Here, $\mathcal{F}$_X(s) is a simplified, state-specific "**surprise**" or cost associated with state s under model $\mathcal{M}$_X. It can be conceptualized as the negative log joint probability - ln p(o, s | $\mathcal{M}$_X) or a variational free energy where the distribution is a Dirac delta centered on s. The weighting parameter $\alpha \in [0, 1]$ is crucial. It is not fixed but is dynamically determined by the relative confidence or precision of each model at the onset of the conflict, often related to the inverse of their respective free energies or estimated uncertainties (Friston et al., 2012). For instance, if $\mathcal{F}$_A $\ll$ $\mathcal{F}$_B, the forward model is much more confident, and α will be close to 1, allowing $\mathcal{M}$_A to dominate the resolution. Conversely, if the conflict arises from a shocking violation of narrative expectations, $\mathcal{M}$_B's surprise may drive α toward 0.

The minimization to find ŝ represents an active inference or search process. It is not merely an analytical computation but a constructive cognitive act—a "**deliberation**" phase where the system tests hypothetical state configurations to find one that best reconciles the two sources of evidence (Botvinick & Toussaint, 2012). This process can involve mental simulation, counterfactual reasoning, or focused attention to specific sensory features to gather new evidence (Pezzulo et al., 2013).

Once ŝ is identified, the system conditions its ongoing perception on this resolved state. The projection $q(s) \rightarrow q(s \mid \hat{s})$ effectively collapses the diffuse, conflicting uncertainty into a sharpened, provisional belief. This new belief then serves as a prior or an attentional filter for subsequent processing. It guides active sampling of the environment to confirm or refute the new hypothesis (Schwartenbeck et al., 2013), and it provides a stable anchor point from which to update the internal parameters of *one or both* of the generative models $\mathcal{M}$_A and $\mathcal{M}$_B. This parameter update aims to reduce the free energy of the now-dominant model for state ŝ, thereby aligning the models' predictions for the future and reducing $\Delta\mathcal{F}$ below the threshold θ.

In summary, the localization phase in Ze is modeled as a first-order phase transition triggered by exceeding a conflict threshold θ. Its computational essence is a projection onto a resolved latent

state ŝ that minimizes a confidence-weighted sum of the models' free energies. This process formalizes the shift from a state of interpretative ambiguity and parallel processing to one of focused hypothesis testing and model revision. It provides a mathematical description for cognitive events such as error detection, surprise-driven learning, and insight, where the system actively restructures its understanding to resolve internal contradiction (FitzGibbon et al., 2020). The final section will integrate these dynamics into a unified principle of meta-energy minimization.

Active Actions (Ze ≠ Passive Bayes)

Thus far, the Ze framework has been presented as a perceptual and inferential architecture, maintaining dual world models and managing the conflict between them. However, a cognitive system that merely observes and interprets the world is incomplete. True intelligence requires the capacity for *goal-directed action* to navigate, manipulate, and learn from the environment (Pfeifer & Bongard, 2006; Lake et al., 2017). Critically, Ze is not a passive Bayesian observer; it is an *active inference and control system* where actions are generated to resolve internal uncertainty and dissonance across its generative models. This transforms Ze from a model of perception into a model of *embodied, adaptive agency*.

In Ze, actions are not generated by a single, monolithic controller. Instead, the duality of the architecture extends to the motor domain. At any given time step t, an action a_t is sampled from a policy π associated with *one* of the two generative models. Formally:

$$a_t \sim \pi_A(a_t \mid s^A_t, \Omega_t) \quad \text{or} \quad a_t \sim \pi_B(a_t \mid s^B_t, \Omega_t)$$

Here, π_A and π_B represent action policies derived from the forward ($\mathcal{M}$_A) and inverse ($\mathcal{M}$_B) models, respectively. Each policy maps from its model's current latent state belief s^A_t or s^B_t and a current objective Ω_t to a distribution over possible actions. Crucially, Ω_t is not a fixed external reward signal but an internally generated *target distribution* over future states or observations, often conceptualized as a prior preference in active inference (Friston et al., 2017). The nature of this target can differ: $\mathcal{M}$_A's policy π_A might aim to minimize expected future prediction error (expected free energy) under its forward dynamics, leading to information-seeking (epistemic) or uncertainty-reducing (pragmatic) actions (Kaplan & Friston, 2018). In contrast, $\mathcal{M}$_B's policy π_B might aim to realize a specific counterfactual future or narrative arc inferred by the inverse model, leading to goal-directed or "explanation-driven" actions (Botvinick & Toussaint, 2012).

The fundamental question is: **How does Ze decide which model's policy to enact?** This arbitration is not arbitrary but is governed by a meta-control principle that seeks to resolve the system's overall cognitive tension. We propose that the system selects the policy that is expected to most effectively reduce the *combined variational free energy* of both models in the future. Formally, the chosen policy π at time t is:

$$\pi = \operatorname{argmin}_{\pi \in \{\pi_A, \pi_B\}} E_{q(s^A, s^B, o_\tau \mid \pi)} [\, \mathcal{F}_A(o_\tau, q_A) + \mathcal{F}_B(o_\tau, q_B) \,]$$

where the expectation is taken over predicted future states s^A, s^B and observations o_τ (τ > t) under the candidate policy π. This rule encapsulates a drive for *global coherence*. The system evaluates which course of action—guided by the causal or the counterfactual perspective—is anticipated to yield future sensory data that both models can explain with minimal surprise (i.e., low free energy for both).

This arbitration mechanism has profound implications. When the models are coherent ($\Delta\mathcal{F}$ is low), their predictions and preferred actions will often align, making the choice trivial. However, in a state of high conflict ($\Delta\mathcal{F}$ > θ), the policies π_A and π_B may prescribe radically different actions. For instance, confronted with an ambiguous perceptual stimulus, π_A (forward model) might prescribe an orienting action to gather more sensory data (e.g., moving closer), while π_B (inverse model) might prescribe a testing action based on a hypothesized narrative (e.g., pressing a button to see if it confirms a guessed rule) (Gottlieb & Oudeyer, 2018). The meta-control rule selects the action expected to resolve the conflict most efficiently, effectively using action as a tool for *active learning and disambiguation*.

This makes Ze an inherently active system. It does not wait passively for evidence to resolve its internal conflicts; it *intervenes* in the world to generate informative outcomes (Pearl, 2009; Linson & Friston, 2019). An action from π_A serves to test and refine the causal structure of the environment. An action from π_B serves to test and realize counterfactual explanations or goals. Through this process of selective intervention, Ze simultaneously shapes its sensory stream and sculpts its internal models, ensuring they remain grounded and functional.

Furthermore, the outcome of an action provides critical feedback for the localization process described in Section 5. The sensory consequences of an enacted policy directly inform the system about which model's predictions were more accurate, thereby updating the confidence weights (the α parameter in the localization equation) and driving model refinement. Action and perception in Ze form a tight, reciprocal loop, where perception generates model conflict, conflict drives policy selection for action, and action generates new data to resolve the conflict (Ahissar & Assa, 2016).

In conclusion, the extension of Ze to include active policies π_A and π_B, governed by a meta-control rule that minimizes expected total free energy, completes the framework as one of active, embodied cognition. Ze transcends passive Bayesian inference by using the duality of its generative models to generate a strategic exploration-exploitation policy. It acts not as a mere observer of the world, but as an autonomous agent that selectively intervenes to reduce its own internal dissonance, thereby actively constructing a coherent and actionable understanding of its environment.

Which-Path Information as an Increase in Environmental Dimensionality

The Ze architecture, as developed thus far, describes an agent navigating an environment defined by its latent states s. However, a critical challenge for any adaptive system is the

discovery that its current state space is insufficient—that hidden variables or contextual factors, previously unnoticed or conflated, are causally relevant. This is the problem of *laten variable discovery* and *representational expansion* (Gershman & Niv, 2010; Schapiro & Turk-Browne, 2015). In Ze, we formalize this discovery process through the concept of **"which-path" information**, a term borrowed from quantum mechanics denoting information that distinguishes between alternative causal pathways. Here, it refers to information that reveals the existence of a previously hidden contextual dimension or discrete alternative. We propose that the incorporation of which-path information is mathematically equivalent to an *expansion of the environmental state space*, a process that necessarily amplifies internal conflict and triggers profound cognitive restructuring.

Consider an environment where observations o are generated by a latent process that can follow one of several distinct causal regimes or "paths," indexed by a hidden variable e. Initially, the agent's models, $\mathcal{M}$_A and $\mathcal{M}$_B, operate on a compressed state representation s that marginalizes over e. The agent perceives a single, albeit potentially noisy and inconsistent, world. The acquisition of which-path information—through accumulated statistical regularities, a decisive intervention, or a salient cue—reveals that the true generative process operates in the expanded space s, e. The effective dimensionality of the environment, from the agent's perspective, increases.

Formally, this expansion is represented as:

$$s \rightarrow (s, e)$$

where e is a new latent variable (e.g., a context label, a hidden cause, or a discrete mode). The consequences of this expansion for the Ze dynamics are immediate and significant:

1. **Increase in Model Conflict ($\Delta\mathcal{F} \uparrow$)**: The existing generative models, $\mathcal{M}$_A and $\mathcal{M}$_B, which were parameterized for the simpler space s, suddenly become misspecified. Their predictions will grow increasingly inaccurate as they fail to account for the modulation introduced by e. Since the two models may have different sensitivities to this misspecification, their variational free energies will diverge. For instance, the forward model $\mathcal{M}$_A might show a sharp rise in $\mathcal{F}$_A as its physical predictions fail, while the inverse model $\mathcal{M}$_B might struggle even more (or less) to find a coherent narrative, causing $\mathcal{F}$_B to change differently. This divergence directly increases the global conflict signal $\Delta\mathcal{F}$ = |$\mathcal{F}$_A - $\mathcal{F}$_B| (Griffiths et al., 2015).
2. **Increase in Posterior Divergence ($\mathcal{I} \uparrow$)**: As the models become misspecified, their approximate posteriors q_A and q_B will be pulled towards different regions of the (inadequate) state space s in a futile attempt to explain the data. One model may latch onto one set of spurious correlations, while the other model latches onto another. This leads to a marked increase in the Jensen-Shannon divergence $\mathcal{I}$ = D_JS(q_A || q_B), indicating a loss of compatibility between their interpretations. The interference regime, which relies on posterior similarity, becomes untenable.

The joint increase in both $\Delta\mathcal{F}$ and $\mathcal{I}$ creates a powerful, compounded signal that makes localization inevitable. The system will reliably exceed the localization threshold θ (Section 5). However, this is not a localization to a point within the old state space. It is a meta-localization — a realization that the conflict cannot be resolved within the current representational schema (Tervo et al., 2016). The projection q(s) -> q(s | ŝ) defined earlier is insufficient; the resolved state ŝ in the old space does not exist.

Therefore, the localization process must now drive a *structural revision* of the generative models themselves. This involves:

- **Hypothesizing the new variable e**: The system must posit the existence of an additional latent dimension. This can be guided by the specific patterns of prediction errors (which-path information often manifests as residual, unexplained variance that is structured, not random) (Courville et al., 2006).
- **Differentiating the models**: The expanded state space s, e may allow—or force—the two models to specialize further. For example, $\mathcal{M}$_A might learn to predict dynamics conditional on e, while $\mathcal{M}$_B might infer the likely value of e from narrative coherence. Their policies π_A and π_B (Section 6) can now generate distinct exploratory actions to actively identify the value of e in novel situations (Gottlieb et al., 2013).
- **Re-learning in the expanded space**: The parameters of both models must be updated based on data now interpreted in the context of s, e. The meta-control policy selection rule will now evaluate actions based on their expected efficacy in reducing free energy in this richer, more veridical environment.

This process formalizes a cycle of representational growth driven by irreducible conflict. Persistent, high-amplitude $\Delta\mathcal{F}$ and $\mathcal{I}$ are not just signals of failure but are *necessary conditions* for the system to discover that its world is more complex than previously modeled (Friston et al., 2017). The integration of which-path information and the subsequent expansion from s to s, e is how Ze moves from a naive to a sophisticated understanding, capable of disentangling contexts and recognizing multiple causal pathways. It is the mathematical expression of a cognitive "aha!" moment that restructures the agent's ontology (Kounios & Beeman, 2014).

In summary, which-path information acts as a catalyst for representational complexity. Its assimilation forces an expansion of the environmental state space, which inescapably inflates the core conflict signals $\Delta\mathcal{F}$ and $\mathcal{I}$ within Ze. This forces the system out of mere state-estimation and into a regime of model revision and dimensional learning. Thus, the very quantities that signal dysfunction during routine operation become the drivers of conceptual growth and adaptive complexity in the face of a multifaceted world.

The Quantum Eraser as the Ze Operator

The previous section described how the acquisition of "which-path" information, by expanding the state space to s, e, triggers conflict and forces structural revision. However, a sophisticated cognitive system must also possess the complementary ability to *simplify* its world model when contextual detail becomes irrelevant, overly costly to maintain, or actively detrimental to coherent action. We formalize this capacity through the **Quantum Eraser Operator** $\mathcal{E}$, a

concept inspired by quantum information experiments where the erasure of path information restores wave-like interference patterns (Walborn et al., 2002; Kim et al., 2000). In the Ze architecture, $\mathcal{E}$ represents a meta-cognitive operation that actively *suppresses or forgets* the conditional dependency on a specific latent variable e, effectively reducing the apparent dimensionality of the environment and restoring the conditions for model coherence.

Consider the state of the system after it has expanded its representation to include a context variable e. The agent's beliefs are now conditioned on this variable: the posterior distributions and model predictions depend on p(e | s) or p(e | o), the probability of context given a state or observation. The Quantum Eraser Operator $\mathcal{E}$ acts upon this conditional dependency. Formally, it is defined as an operation that renders the latent variable e statistically independent or uninformative with respect to the core state s or the observations:

$$\mathcal{E}: \ p(e \mid s) \rightarrow \text{const.}$$

That is, the operator transforms the conditional distribution p(e | s) into a constant function, meaning the probability of any particular e becomes uniform and independent of s. Equivalently, it can be seen as marginalizing out e or "blurring" the which-path information, making the distinct causal paths indistinguishable again from the perspective of the models (Scully & Drühl, 1982).

The action of $\mathcal{E}$ has three critical, non-intuitive consequences:

1. **It Does Not Alter Past Data**: The operator $\mathcal{E}$ does not erase sensory history o_{1:t} from memory, nor does it retroactively change the agent's belief about what physically occurred. The raw data and the memory of the sequence of events remain intact. This distinguishes it from mere forgetting. Instead, $\mathcal{E}$ changes the *interpretive framework* applied to that data. It alters how the generative models attribute causes and structure to the past and future, moving from a fine-grained, context-dependent interpretation to a coarse-grained, context-independent one (Gershman et al., 2015).
2. **It Reduces Environmental Support**: By decoupling e from s, the operator effectively collapses the expanded state space s, e back towards the simpler subspace s. The environment, as *modeled* by the agent, loses a dimension of distinguishing detail. The "paths" that were previously distinct become merged into a single, broader category. This is a form of *adaptive abstraction* or *chunking*, where specific instances are grouped under a more general schema to reduce computational load and foster generalization (Rabinovich et al., 2012; Tomov et al., 2021).
3. It Decreases Model Conflict ($\Delta\mathcal{F} \downarrow$): This is the primary functional role of $\mathcal{E}$. The high conflict $\Delta\mathcal{F}$ arose because the two models $\mathcal{M}$_A and $\mathcal{M}$_B struggled to account for the nuances modulated by e. By applying $\mathcal{E}$, the system simplifies the explanatory task. The models no longer need to account for variance attributable to the now-erased e. Their predictions become less precise but more broadly applicable, and their free energies $\mathcal{F}$_A and $\mathcal{F}$_B are likely to converge, as both are evaluated against a less demanding, smoothed-out version of reality. Consequently, the global conflict signal $\Delta\mathcal{F} = |\mathcal{F}\text{\_A} - \mathcal{F}\text{\_B}|$ decreases.

The triggering condition for applying the Quantum Eraser is not explicitly modeled here as an optimization but can be linked to sustained cognitive cost. When the system operates for an extended period in a high-conflict localization regime ($\Delta\mathcal{F} > \theta$) without successfully identifying a stable, predictive structure for e, the meta-cognitive cost of maintaining the complex representation may outweigh its benefits. The eraser $\mathcal{E}$ is then deployed as a "reset" or simplification heuristic (Wilson & Niv, 2011).

The ultimate goal of this operation is the restoration of the **interference regime**. The criterion for successful interference, as established in Section 3, is:

$$\Delta\mathcal{F} < \theta$$

By applying $\mathcal{E}$ and reducing $\Delta\mathcal{F}$ below the localization threshold θ, the system signals that the conflict has been resolved at a higher level of abstraction. The sharp divergence between the posteriors q_A and q_B subsides, and their Jensen-Shannon divergence $\mathcal{I}$ decreases accordingly. This re-enables the constructive blending of the models' outputs, allowing for fast, efficient, and coherent perception-action cycles based on a simplified, more robust world model.

In summary, the Quantum Eraser Operator $\mathcal{E}$ completes the Ze cognitive cycle. It provides a formal mechanism for strategic simplification, counterbalancing the complexification driven by which-path information. By conditionally decoupling a latent variable, $\mathcal{E}$ reduces representational granularity, lowers model conflict, and restores the system to a stable, interferometric state where perception and action can proceed efficiently. This dynamic alternation between differentiation (state expansion) and integration (erasure-driven simplification) mirrors fundamental processes in cognitive development and learning (Siegler, 2005), positioning Ze as a unified formalism for adaptive, resilient intelligence.

Sleep and Wakefulness as Parameter Regimes

The Ze architecture, with its cycles of differentiation (via which-path information) and integration (via the quantum eraser), provides a framework for understanding online perception and learning. To complete the picture of a biologically plausible and sustainable cognitive system, we must account for **offline** states of processing. We propose that the fundamental states of **sleep** and **wakefulness** can be formally described as distinct dynamical regimes of a core parameter within the Ze framework. This parameter, the **path fixation strength** λ, governs the system's commitment to maintaining a specific, detailed model of the world versus its propensity for representational reorganization.

We introduce λ as a scalar meta-parameter that multiplicatively weights a "path-specificity" cost within the variational free energy functional of the forward model $\mathcal{M}$_A. Let path(s) be a functional that quantifies the specificity or "crispness" of the model's commitment to a particular trajectory or partition of the state space s, e. This could be related to the entropy of the distribution over contexts e, the precision of state estimates, or the complexity cost of

maintaining fine-grained distinctions (Friston et al., 2017). The modified free energy for the forward model during online processing becomes:

$$\mathcal{F}\_A^{\lambda} = \mathcal{F}\_A + \lambda \ \text{path(s)}$$

The value of λ determines the operational regime of the entire Ze system:

1. **Wakefulness: The High-λ Regime (λ ≫ 1)**. In the waking state, the primary imperative is to maintain a precise, contextually specific, and immediately actionable model of the world to support real-time perception and action (Mackay, 2021). A high value of λ strongly penalizes any loss of path specificity within $\mathcal{F}$_A^λ. This forces the forward model $\mathcal{M}$_A to commit to a single, well-defined interpretation of sensory data to minimize its free energy. It suppresses the exploration of alternative state configurations or the merging of contextual distinctions. Consequently:
    - **Model Commitment is High**: The posterior q_A(s, e) is sharp and confident.
    - **Interference is Conditionally Permitted**: Interference with $\mathcal{M}$_B occurs only when the counterfactual model's narrative strongly aligns with this committed path (($\mathcal{I}$ ≈ 0). Otherwise, the high λ reinforces the dominance of the currently selected forward model interpretation.
    - **Localization is Goal-Directed**: Conflict-driven localization (Δ$\mathcal{F}$ > θ) is primarily resolved by seeking new data (via active sensing) to refine the existing high-specificity model, not by radically reconfiguring it.
    - **The Quantum Eraser is Inactive**: The operation $\mathcal{E}$ is suppressed, as erasing path details would catastrophically increase the λ-weighted path cost.
2. **Sleep: The Low-λ Regime (λ → 0)**. During sleep, the constraints of real-time sensorimotor interaction are relaxed. We model this as a dramatic reduction in the path fixation parameter, λ → 0 (Hobson & Friston, 2012; Lewis & Durrant, 2011). The modified free energy simplifies: $\mathcal{F}$_A^{λ → 0} ≈ $\mathcal{F}$_A. The high cost of maintaining precise, context-dependent distinctions is removed. This liberates the system and enables profound offline processing:
    - **Exploration of State Space**: Without the penalty for low specificity, the forward model can explore a much broader landscape of potential state configurations and associations. This facilitates the replay and consolidation of memories, allowing sequences to be re-experienced and integrated without the pressure of committing to a single "real" path (Diekelmann & Born, 2010).
    - **Activation of the Quantum Eraser**: The low-λ regime is the natural habitat for the quantum eraser operator $\mathcal{E}$. With the cost of erasure minimized, the system can safely decouple spurious or overly detailed contextual associations (p(e | s) → const). This promotes generalization by extracting statistical invariants and forgetting irrelevant details, a process linked to synaptic downscaling and memory optimization (Tononi & Cirelli, 2014).
    - **Model Restructuring and Integration**: The reduced λ allows the two models, $\mathcal{M}$_A and $\mathcal{M}$_B, to interact more freely. The barrier to interference $\mathcal{I}$ is effectively lowered, enabling the integration of narrative structures from $\mathcal{M}$_B (e.g.,

semantic or episodic knowledge) with the sensorimotor statistics of $\mathcal{M}$_A. This cross-model integration is hypothesized to underly creative insight and procedural memory consolidation (Stickgold & Walker, 2013).

- **Resolution of Lingering Conflict**: High-conflict states ($\Delta\mathcal{F}$ > θ) that could not be efficiently resolved online due to the high cost of model revision can be addressed offline. The system can test radical reparameterizations of its models without behavioral risk.

Thus, **sleep is formally defined as the dynamical reduction of the path fixation parameter λ**. The transition from wakefulness to sleep corresponds to a gradual or phase-transition-like decrease in λ, switching the Ze system from a mode of precise, committed, real-time inference to a mode of exploratory, integrative, and simplifying computation (Gomez-Ramirez & Sanz, 2013). The cyclical alternation between high and low λ regimes ensures that the system remains both adaptive to immediate demands (wakefulness) and capable of long-term optimization and structural learning (sleep).

In conclusion, by incorporating a single, metabolically or neuromodulatorily regulated parameter λ into the core free energy functional, the Ze framework naturally accommodates the fundamental cycle of sleep and wakefulness. This elevates Ze from a model of momentary cognition to a model of embodied, cyclical intelligence, where offline states are not passive but are essential, active phases of cognitive maintenance, reorganization, and growth.

Connection to the Double-Slit Experiment

The mathematical architecture of Ze, culminating in the dynamics of interference ($\mathcal{I}$), conflict ($\Delta\mathcal{F}$), localization, and erasure, is not merely an abstract cognitive model. It finds a profound and clarifying analogy in one of the most fundamental experiments in physics: the double-slit experiment and its modern variants incorporating "which-path" information and quantum erasure. This analogy is not merely poetic but serves as a rigorous formal parallel, suggesting that the principles governing quantum measurement and wave function collapse may share a deep structural homology with the principles governing cognitive inference and model selection (Bruza et al., 2015; Busemeyer & Bruza, 2012). The following table and analysis elucidate this connection.

Table 1

Double-Slit Experiment	Ze Cognitive Architecture	Formal Analogy
Wave Function / Superposition	Model Compatibility	A quantum system exists in a superposition of passing through both slits. In Ze, a state of low conflict (ΔF ↓) and high posterior compatibility (I ≈ 0) allows multiple generative models (M_A, M_B) to co-exist in a blended, superposed" perceptual state.

Interference Pattern	Constructive Interference Regime	The wave-like superposition produces an interference pattern on the detector. In Ze, compatible posteriors constructively interfere, leading to enriched, coherent percepts and efficient action selection (π) based on fused predictions. This is the default state for a coherent world-model.
Which-Path Information	Increase in ΔF	Placing a detector to determine which slit a particle passes through provides "which-path" information. In cognition, discovering a hidden contextual variable e (e.g., a latent cause) provides analogous discriminative information. This acquisition forces an expansion of the state space (s → (s, e)), which increases model conflict (ΔF ↑) as the existing models become misspecified.
Collapse (Wave → Particle)	Localization Phase Transition	The act of measurement (obtaining which-path info) collapses the wave function to a particle-like state with a definite path. In Ze, exceeding the conflict threshold (ΔF > θ) triggers a localization phase transition. The system collapses from a blended perceptual state into a specific, resolved interpretation ŝ, suppressing interference.
Quantum Eraser	Eraser Operator E	A quantum eraser setup retroactively erases the which-path information, recovering an interference pattern even after detection. In Ze, the operator E acts by marginalizing out or decorrelating the contextual variable e (p(e | s) → const). This reduces the effective environmental dimensionality, lowers conflict (ΔF ↓), and restores the conditions for model interference (I → 0).
Decoherence (Environment)	Self-Induced Decoherence	Interaction with a macroscopic environment causes rapid decoherence, effectively performing a continuous "measurement." In Ze, the system's own commitment to a specific action policy (π_A or π_B) and its ensuing sensorimotor engagement with the world act as a continuous self-measurement. This ongoing interaction favors the "collapsed," localized state of the forward model M_A during wakefulness (high-λ regime), maintaining a classical, definite perceptual reality (Atmanspacher & beim Graben, 2007).

Analysis of the Correspondence:

The core of the analogy lies in the treatment of *information* and its effect on *coherence*. In quantum mechanics, the system's state is described by a wave function whose phase coherence allows interference. The acquisition of *discriminating information* (which-path) destroys this coherence, leading to a definite but impoverished (non-interfering) state. The erasure of that information can, under specific conditions, restore coherence (Walborn et al., 2002).

In the Ze architecture, the "wave function" is replaced by the *landscape of compatible Bayesian beliefs* across generative models. The coherence measure is the Jensen-Shannon divergence $\mathcal{I}$. The acquisition of discriminating cognitive information—which-path information about a hidden context—destroys this belief compatibility ($\mathcal{I} \gg 0$), forcing a definite but cognitively costly "collapsed" state (localization). The cognitive erasure operator $\mathcal{E}$ restores compatibility, allowing for flexible, interference-based processing once more.

This parallel suggests that the transition from quantum-like to classical-like behavior is not exclusive to microscopic physics but may be a general feature of any information-processing system that must balance the maintenance of multiple potential states (for robustness and prediction) with the need to commit to a single state for action (Khrennikov, 2010). The waking brain, constantly acting in the world, operates under a regime of continuous self-induced decoherence, maintaining a "collapsed," classical stream of consciousness. In contrast, during sleep or reflective thought (low-λ regime), the decoherence pressure is reduced, allowing for more quantum-like, superposed exploration of ideas and memories—a process that may underpin creativity and insight (Merali, 2015).

Conclusion of the Analogy:

Thus, the double-slit experiment serves as a powerful metaphysical and mathematical metaphor for the Ze formalism. It provides an existence proof in nature for a system whose observable behavior (interference vs. particle tracks) is radically determined by the presence or absence of information that distinguishes between internal possibilities. Ze posits that cognition operates under an identical principle: our perception of a coherent, classical reality is actively maintained by the continuous resolution of conflict between internal generative models, and the disruption of this process is not a failure but a necessary gateway to learning and representational change.

Connection to the Double-Slit Experiment

The Ze architecture, with its formalization of interference, conflict, and erasure, transcends a mere cognitive model. It establishes a profound structural isomorphism with one of the foundational experiments of modern physics: the **double-slit experiment** and its extensions into quantum information theory. This connection is not simply metaphorical but offers a rigorous, unifying mathematical framework that suggests principles of quantum measurement and coherence have direct analogues in high-level cognitive processes (Busemeyer & Bruza, 2012; Bruza et al., 2015). The following schema delineates this formal correspondence, which we will subsequently unpack.

Table 2

Physical Concept (Double-Slit)	Ze Cognitive Architecture	Formal Ze Expression / Mechanism
Wave Function (Superposition)	Model Compatibility	Low global conflict, high posterior compatibility allows blended model states.
Interference Pattern	Constructive Interference Regime	Low Jensen-Shannon divergence enables fused percepts: $I \approx 0$
Which-Path Information	State Space Expansion	Acquisition of hidden variable e expands state: $s \rightarrow (s, e)$, increasing conflict: $\Delta F \uparrow$.

Collapse (Measurement)	Localization Phase Transition	Threshold exceedance triggers discrete shift: $\Delta F > \theta \Rightarrow$ projection to s^.
Quantum Eraser	Erasure Operator E	Application of E decorrelates context: p(e | s) → const, reducing ΔF, restoring I ≈ 0.
Decoherence (Environment)	Self-Induced Decoherence via Action	Commitment to a policy π_A *or* π_B enacts continuous "measurement," stabilizing the localized percept during wakefulness.

Analysis of the Structural Isomorphism

1. **Superposition and Model Compatibility**. In the quantum double-slit setup, a single particle is described by a wave function that is a superposition of passing through both slits simultaneously. There is no "which-path" information, and the system exists in a coherent state of multiple possibilities (Feynman et al., 1965). Analogously, in the Ze architecture, a state of low model conflict ($\Delta\mathcal{F} \approx 0$) and high posterior compatibility (($\mathcal{I} \approx 0$) signifies that the two generative models $\mathcal{M}$_A and $\mathcal{M}$_B are promoting statistically indistinguishable interpretations of the sensory stream. The cognitive system resides in a "superposed" state where multiple coherent explanations are simultaneously viable and actively blended, leading to robust perception (Pothos & Busemeyer, 2013).
2. **Interference and the Interference Regime**. The physical superposition results in a wave-like interference pattern on the detection screen, a hallmark of quantum coherence. In Ze, the analogous phenomenon is the *constructive interference regime*, where the compatible posteriors q_A and q_B are fused. This fusion yields percepts and predictions that are more precise and contextually enriched than those of either model alone—a cognitive "interference pattern" that is the hallmark of a coherent, well-understood situation.
3. **Which-Path Information and Increased Conflict**. Introducing a detector to determine the particle's path constitutes an act of measurement that acquires "which-path" information. This destroys the wave function's coherence, collapsing it into a particle-like state with a definite path, and the interference pattern vanishes (Scully et al., 1991). The cognitive parallel is the acquisition of a latent contextual distinction—the "which-path" information that differentiates two previously confounded causal narratives. As formalized in Section 7, this expands the state space from s to s, e. The existing models, tuned to the simpler space, become misspecified. Their predictions diverge, causing a sharp increase in the global conflict signal $\Delta\mathcal{F}$. The cognitive "interference pattern" (coherent perception) is lost.
4. **Collapse and Localization**. The quantum collapse is an all-or-nothing transition from a wave (delocalized, interfering) to a particle (localized, definite) description. In Ze, the analogous event is the localization phase transition (Section 5). When $\Delta\mathcal{F}$ exceeds the threshold θ, the system undergoes a qualitative shift from the parallel, interferometric processing of the two models to a serial, diagnostic mode. It "collapses" onto a specific, resolved state ŝ that minimizes a weighted model conflict. This transition models

cognitive events such as sudden disambiguation, error detection, or the instantiation of a specific hypothesis.

5. **Quantum Eraser and the Erasure Operator**. The quantum eraser experiment demonstrates that if which-path information is *erased* in a coherent manner after the particle has been detected, the interference pattern can be recovered (Walborn et al., 2002). This highlights that it is the *existence of potentially knowable distinguishing information*, not the act of observation per se, that destroys coherence. In Ze, the **Quantum Eraser Operator** $\mathcal{E}$ (Section 8) performs precisely this function. It acts by decorrelating the contextual variable e from the core state s (p(e | s) → const). This erasure of discriminative information reduces the effective dimensionality of the environment, lowers the model conflict $\Delta\mathcal{F}$, and restores the conditions for posterior compatibility ($\mathcal{I} \to 0$), thereby "recovering" the cognitive interference regime.
6. **Decoherence and Self-Induced Stabilization**. In open quantum systems, interaction with a large environment causes rapid decoherence, continuously localizing the system into a classical state (Zurek, 2003). The Ze architecture exhibits a powerful analogue: **self-induced decoherence through action**. The system's own commitment to an action policy (π_A or π_B) and the resultant sensorimotor engagement with the world generate a continuous stream of proprioceptive and exteroceptive feedback. This feedback acts as a constant "measurement," anchoring perception to a specific, actionable interpretation—the forward model's $\mathcal{M}$_A "classical" reality. This process is dominant in the high-λ wakefulness regime, ensuring perceptual stability (Clark, 2013).

The formal correspondence between the double-slit experiment and the Ze architecture suggests that the mathematical structures describing quantum coherence and measurement may be universal for any system that must manage the trade-off between maintaining multiple potential states (for robustness and prediction) and committing to a single state for decisive action. Ze provides a precise cognitive instantiation of these principles, framing perception, learning, and consciousness itself as dynamic processes of interference, measurement, and erasure played out on the stage of embodied action.

The Key Formal Conclusion

The preceding sections have meticulously constructed the Ze architecture, drawing a formal analogy with quantum mechanics to frame cognitive dynamics. This leads to a pivotal and non-trivial conclusion that reframes a foundational puzzle in both cognitive science and foundational physics. The Ze formalism demonstrates that **the phenomenon of "collapse"—the discrete transition from a state of multiple possibilities to a single, definite outcome—is not a fundamental postulate or an exogenous intervention. Rather, it is an emergent, optimization-driven phase transition within a system that maintains competing internal models of the world**.

Reframing the Problem of Collapse

In quantum mechanics, the collapse of the wave function is often treated as a primitive postulate of the Copenhagen interpretation—an unexplained, instantaneous event triggered by

measurement (von Neumann, 1932). This has long been a source of conceptual unease, prompting interpretations from many-worlds to objective collapse theories. In cognitive science, analogous phenomena—such as the sudden resolution of perceptual ambiguity (e.g., the Necker cube), the crystallization of an insight, or the commitment to a single action plan amidst uncertainty—are similarly described but often lack a unifying formal principle beyond descriptive thresholds or stochastic switches (Hohwy et al., 2008).

The Ze architecture provides a unifying formal substrate for both domains. Here, "collapse" is not a mysterious axiom but the natural, observable consequence of a continuous, resource-optimizing process. The system is perpetually engaged in minimizing variational free energies ($\mathcal{F}$_A, $\mathcal{F}$_B) that quantify the accuracy and complexity of its dual world models. The state of multiple possibilities—the "superposition"—corresponds to the **interference regime**, where the models' posteriors are compatible (($\mathcal{I} \approx 0$) and their free energies are low and comparable ($\Delta\mathcal{F} \approx 0$). In this regime, the system enjoys the benefits of a blended, robust representation.

Collapse as an Optimization-Driven Phase Transition

The transition out of this state is triggered by an optimization failure. The acquisition of "which-path" information—sensory data that reveals a previously hidden contextual variable e—fundamentally changes the structure of the inference problem. The existing models, optimized for a simpler state space (s), become severely misspecified when confronted with data generated from the expanded space s, e. Their attempts to minimize their individual free energies under this new constraint cause their solutions to diverge. One model may adjust its parameters to account for the new variable in one way, while the other model finds a different, incompatible solution.

This divergence is quantified by the **model conflict** $\Delta\mathcal{F}$, which rises sharply. The system reaches a point where maintaining the blended, "superposed" state is no longer optimal, as it would require tolerating high and conflicting prediction errors from both models simultaneously. This is suboptimal from a variational perspective, which seeks to minimize total expected prediction error (Friston, 2010). The **localization threshold** θ represents the system's tolerance for this inefficiency.

When $\Delta\mathcal{F} > \theta$, the system undergoes a **phase transition** (Section 5). Mathematically, this is a bifurcation in the dynamics of the coupled inference processes. The stable attractor corresponding to the blended interference regime loses stability, and a new set of attractors—corresponding to resolved, model-specific interpretations—becomes dominant (Tschacher & Haken, 2007). The system "falls into" one of these new basins of attraction through the **projection operation** $q(s) \rightarrow q(s \mid \hat{s})$, where $\hat{s}$ is the state that minimizes a weighted sum of the models' free energies. This discrete jump is the **collapse**.

Agency and Measurement Without Mystery

This formulation elegantly demystifies the role of the "observer" or "measurement." In the quantum analogy, the observer is not an external, classical entity forcing a collapse. Instead, the "measurement" is the process by which one part of the system (e.g., a detector, or in cognition, a specific action policy) becomes correlated with the state of another part, acquiring "which-path" information (Riedel et al., 2016). In Ze, this is modeled by the system's own active engagement. When the system selects and enacts a specific policy π_A or π_B (Section 6), it is effectively performing a self-measurement. The action commits the system to a specific course, generating sensory consequences that are highly informative for one model and potentially disconfirming for the other. This active sampling of the environment serves as the continuous "measurement" that stabilizes the "collapsed," classical stream of perception during wakefulness.

The key insight is that **collapse is an adaptive, resource-optimizing response to unsustainable internal conflict**. It is the system's way of breaking a computational deadlock. By committing to a single interpretation (§), it can focus its resources, generate decisive actions, and pursue a coherent learning trajectory to reduce free energy under the newly clarified (though possibly simplified or provisional) model.

Implications and Unification

This conclusion has significant implications. For cognitive science, it provides a rigorous variational account of discrete perceptual and decision-making events, linking them to the continuous dynamics of predictive processing (Clark, 2013). It frames insight and ambiguity resolution not as lucky guesses but as optimal transitions in a complex adaptive system.

For the quantum cognition paradigm, it strengthens the case for viewing quantum probability not just as a useful descriptive tool for paradoxical human behavior, but as indicative of a deeper, shared computational logic between microscopic and macroscopic information-processing systems (Busemeyer & Bruza, 2012). The "collapse" in both realms can be seen as the resolution of a system struggling to maintain coherence across competing frameworks for explaining evidence.

In summary, the Ze formalism culminates in a powerful synthetic statement: **Collapse is not postulated; it is computed**. It is the inevitable, optimization-driven transition that occurs when the cost of maintaining multiple competing realities outweighs the benefit, forcing a complex system to commit, act, and thereby define its experienced world.

Why This is a Strict Theory, Not a Metaphor

The formal analogy between the Ze cognitive architecture and quantum phenomena, particularly the double-slit experiment, invites a critical distinction: Is Ze merely a suggestive metaphor, or does it constitute a **strict scientific theory**? We argue decisively for the latter. Ze is not a loose analogy that borrows quantum terminology for poetic effect. It is a rigorous, formal framework grounded in established mathematics, which provides a novel architectural explanation for

cognitive dynamics and generates falsifiable empirical predictions. Its strength lies in four pillars of theoretical rigor.

It is Built on Standard Variational Mechanics.

The entire edifice of Ze is constructed from the mathematics of variational inference and the free energy principle, which are standard tools in contemporary theoretical neuroscience and machine learning (Friston, 2010; Blei et al., 2017). The core quantities—the variational free energies $\mathcal{F}$_A and $\mathcal{F}$_B (Section 2)—are not quantum constructs but well-defined functionals from Bayesian probability theory. Their minimization is a formal optimization procedure for approximate Bayesian inference and learning (Dayan et al., 1995). The conflict measure $\Delta\mathcal{F}$ and the interference measure $\mathcal{I}$ (the Jensen-Shannon divergence) are standard information-theoretic quantities (Lin, 1991). Therefore, Ze's foundation is not speculative physics but applied mathematics with a proven track record in modeling perception, action, and learning (Buckley et al., 2017). The quantum-like phenomena emerge from the interaction dynamics of these classically defined components, not from imported quantum axioms.

It Does Not Alter Schrödinger's Equations.

A common pitfall of quantum-inspired cognitive theories is the temptation to postulate novel quantum dynamics in the brain, a stance fraught with biophysical and scalability issues (Tegmark, 2000). Ze makes no such claim. It remains entirely agnostic about the microphysical implementation. The theory does not propose that Schrödinger's equation governs neural activity or that superposition occurs at a neuronal level. Instead, it posits that the **computational and statistical properties** of a system performing variational inference over dual generative models can exhibit formal isomorphisms with the **mathematical structure** of quantum measurement. The "wave-like" and "particle-like" behaviors are descriptions of information-processing regimes (interference vs. localization), not of physical states of matter. Thus, Ze is compatible with all known neurophysiology while offering a higher-level functional explanation (Brette, 2022).

It Adds a Novel Architectural Level of Explanation.

Ze transcends metaphor by proposing a specific, testable architectural hypothesis about cognitive organization. It is not merely saying "cognition is like quantum mechanics." It is proposing that a necessary feature of advanced intelligence is the maintenance of two distinct, asymmetric generative models—a causal/forward model ($\mathcal{M}$_A) and a counterfactual/inverse model ($\mathcal{M}$_B)—whose interaction is governed by the minimization of their combined free energies (Sections 1 & 2). This duality and the resulting conflict dynamics ($\Delta\mathcal{F}$) provide a formal, architectural explanation for phenomena that are otherwise described separately: perceptual multistability, metacognitive confidence, epistemic foraging, and sleep-related memory reorganization (Shea et al., 2014; Findling et al., 2023). The theory makes concrete claims: that neural representations corresponding to q_A and q_B should be dissociable, that their relative precision should modulate behavioral interference, and that the global conflict signal $\Delta\mathcal{F}$ should correlate with neural markers of surprise and with the triggering of exploratory behaviors.

It Yields New, Falsifiable Experimental Predictions.

The ultimate criterion for a strict theory is its ability to generate novel, testable predictions. Ze generates a rich set of such predictions across levels of analysis:

- **Neurophysiological**: The theory predicts specific neural signatures. We should observe two distinct but interacting neural populations or networks whose activity patterns correspond to the evolving posteriors q_A and q_B. The global conflict signal $\Delta\mathcal{F}$ should be encoded in neuromodulatory systems (e.g., norepinephrine or acetylcholine) or large-scale synchronization measures (e.g., frontal theta power), correlating with pupil dilation and behavioral markers of uncertainty (Aston-Jones & Cohen, 2005). The application of the "quantum eraser" operator $\mathcal{E}$ (e.g., during sleep or after task mastery) should be observable as a decoupling of functional connectivity between networks encoding specific contextual details and those encoding general schemata (Tomov et al., 2021).
- **Behavioral/Cognitive**: The model makes quantitative predictions about reaction times and error rates. Transitions from the interference regime (($\mathcal{I} \approx 0$) to the localization regime ($\Delta\mathcal{F} > \theta$) should be marked by increased response time variability and a higher probability of exploratory actions, as the system searches for ŝ. The parameter θ (localization threshold) should be manipulable; for instance, stress or cognitive load should lower θ, making individuals more prone to premature perceptual "collapse" or decision-making. The theory also predicts that during the low-λ sleep regime, learning paradigms should show enhanced generalization and schema formation, as the erasure operator $\mathcal{E}$ is more active (Lewis & Durrant, 2011).
- **Computational**: The architecture can be implemented as an active inference agent in simulated or robotic environments. We can test whether an agent equipped with the dual $\mathcal{M}$_A / $\mathcal{M}$_B structure and the Ze conflict-resolution dynamics outperforms a single-model agent in environments requiring the discovery of hidden contexts (which-path information) and the flexible switching between exploratory and exploitative policies.

In conclusion, Ze qualifies as a strict theory because it is formally grounded, non-contradictory with underlying physics, architecturally specific, and empirically vulnerable. It uses the mathematical isomorphisms with quantum formalism not as a metaphysical claim but as a powerful guiding principle to discover a previously undescribed level of cognitive organization. The theory does not reduce mind to quantum physics; instead, it suggests that certain deep principles of inference, information, and measurement manifest in both domains, providing a unified formal language to describe how systems—from particles to persons—navigate a world of hidden possibilities.

The Minimal Testable Prediction

A rigorous theory must ultimately be evaluated against empirical evidence. While the Ze architecture generates a broad range of predictions, a single, minimal, and critical prediction can serve as a decisive test to distinguish it from alternative frameworks, particularly from standard

models of quantum decoherence applied to cognition. This core prediction concerns the role of **action** in resolving perceptual or cognitive ambiguity. Specifically, **Ze predicts that in scenarios of high model conflict ($\Delta\mathcal{F} > \theta$), the active alternation of policies based on competing models (π_A and π_B) will lead to faster localization (i.e., a collapse to a resolved state $\hat{s}$) compared to passive observation or measurement**. This prediction directly contradicts the expected dynamics of a system undergoing standard environmental decoherence and highlights the active, interventionist nature of the Ze agent.

The Prediction in Context

Recall that localization in Ze is the phase transition triggered when conflict exceeds a threshold: $\Delta\mathcal{F} > \theta$. The subsequent projection to a resolved state $\hat{s}$ is the cognitive analogue of wave function collapse. The key question is: What factors influence the *rate* or *latency* of this transition once conflict is high?

In standard decoherence theory, as applied to open quantum systems, the transition from a coherent superposition to a classical mixture (the emergence of "pointer states") is driven by the uncontrolled interaction of the system with a large, noisy environment (Zurek, 2003). The process is *passive*. Information about the system becomes encoded in the environment through entanglement, and the rate of decoherence depends on environmental parameters (e.g., temperature, interaction strength) and the system's own susceptibility. An observer measuring the system does not fundamentally alter this passive process; they merely access the already-decohered information. In a cognitive metaphor of passive decoherence, one would expect perceptual resolution to occur at a speed determined by the inherent noise in neural processing and the accumulation of passive sensory evidence (Kvam et al., 2015).

Ze proposes a fundamentally different mechanism. Localization is not a passive environmental washout but an active inference process driven by the imperative to minimize expected free energy. When conflict is high, the system is not a passive observer; it is an agent with two competing action policies, π_A and π_B, each derived from a different world model (Section 6). The meta-control rule selects actions from the policy expected to minimize total future free energy ($\mathcal{F}_A + \mathcal{F}_B$). Critically, **active alternation between these policies is a form of *interventionist exploration***. An action from π_A tests the predictions of the forward causal model, while an action from π_B tests the counterfactual narrative. Each action generates highly informative, disambiguating sensory feedback that is specifically tailored to reduce the uncertainty of one model relative to the other (Gottlieb & Oudeyer, 2018).

The Mechanism for Accelerated Localization

This active, alternating intervention accelerates localization through two synergistic mechanisms:

1. Rapid Information Gain: Passive observation provides data that is ambiguous with respect to the competing models. In contrast, an action chosen by π_A is designed to create an outcome that $\mathcal{M}_A$ predicts confidently but that would be surprising under $\mathcal{M}_B$

(and vice-versa). This targeted sampling yields *high diagnostic power*, sharply increasing the evidence in favor of one model over the other. This rapidly widens the free energy gap ($\Delta\mathcal{F}$), reinforcing the dominance of the winning model and solidifying the projection to its preferred state $\hat{s}$ (Pezzulo et al., 2013).

2. Precision Amplification: In active inference frameworks, the selection and execution of an action are accompanied by an increase in the *precision* (inverse variance) assigned to the sensory consequences expected under the chosen policy (Friston et al., 2012). This precision weighting amplifies the impact of the resulting prediction error. Therefore, when an action from π_A yields the predicted outcome, the ensuing prediction error for $\mathcal{M}_B$ is not only large but is also assigned high precision, causing a dramatic increase in $\mathcal{F}_B$. This precision-modulated signal acts as a powerful accelerator for the localization transition, a mechanism absent in passive observation.

Therefore, an agent following the Ze meta-control rule will actively seek out interventions that are maximally informative for resolving its internal conflict. This strategic exploration should lead to a significantly **shorter latency between the onset of high conflict ($\Delta\mathcal{F} > \theta$) and the completion of the localization transition (stable commitment to $\hat{s}$)**, compared to an agent or system that is only allowed to passively view an unfolding, ambiguous scenario.

Experimental Design and Distinction from Alternatives

This prediction can be tested in behavioral experiments. A paradigm could involve an ambiguous perceptual decision-making task or a volatile bandit task where the underlying rewarding context (the "which-path" variable e) changes unpredictably (Findling et al., 2023). The key manipulation is the availability of *interventionist actions* versus *passive observation* following a change-point that induces high conflict.

- **Active Condition (Ze-predicted)**: Participants have access to two distinct types of actions. One action type (e.g., a "probe" button) provides information specifically diagnostic of the physical contingencies (testing $\mathcal{M}_A$). The other (e.g., a "hypothesis-test" button) provides information diagnostic of the rule or context (testing $\mathcal{M}_B$). Ze predicts that participants who actively alternate between these action types following a change-point will identify the new correct state $\hat{s}$ faster and with fewer total observations than those in a passive condition.
- **Passive Observation Condition (Decoherence baseline)**: Participants see the outcomes generated by a pre-programmed sequence of actions or by a yoked control, receiving identical sensory information but without the ability to choose interventions. Standard models of evidence accumulation or passive decoherence predict that resolution time will depend only on the information rate, not on its active or passive nature.

A confirmation of faster localization in the active condition would provide strong support for Ze's core tenet that cognitive collapse is an optimization process driven by active, model-based intervention. It would falsify models that treat perception as a purely passive accumulation

process or as a decoherence-like washout by unstructured noise. This minimal testable prediction thus serves as a crucial empirical linchpin for the entire Ze formalism.

References

Ahissar, E., & Assa, E. (2016). Perception as a closed-loop convergence process. eLife, 5, e12830. https://doi.org/10.7554/eLife.12830

Angela, J. Y., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. Neuron, 46(4), 681–692. https://doi.org/10.1016/j.neuron.2005.04.026

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. Annual Review of Neuroscience, 28, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Atmanspacher, H., & beim Graben, P. (2007). Contextual emergence of mental states from neurodynamics. Chaos and Complexity Letters, 2(2/3), 151–168.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. Nature Human Behaviour, 1(4), 0064. https://doi.org/10.1038/s41562-017-0064

Beal, M. J. (2003). Variational algorithms for approximate Bayesian inference [Doctoral dissertation, University College London].

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. Journal of the American Statistical Association, 112(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Botvinick, M., & Toussaint, M. (2012). Planning as inference. Trends in Cognitive Sciences, 16(10), 485–488. https://doi.org/10.1016/j.tics.2012.08.006

Brette, R. (2022). Brains as computers: metaphor, analogy, theory or fact? Frontiers in Ecology and Evolution, 10, 878729. https://doi.org/10.3389/fevo.2022.878729

Bruza, P. D., Wang, Z., & Busemeyer, J. R. (2015). Quantum cognition: a new theoretical approach to psychology. Trends in Cognitive Sciences, 19(7), 383-393. https://doi.org/10.1016/j.tics.2015.05.001

Bruza, P. D., Wang, Z., & Busemeyer, J. R. (2015). Quantum cognition: a new theoretical approach to psychology. Trends in Cognitive Sciences, 19(7), 383–393. https://doi.org/10.1016/j.tics.2015.05.001

Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. Journal of Mathematical Psychology, 81, 55–79. https://doi.org/10.1016/j.jmp.2017.09.004

Buschman, T. J., & Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. Science, 315(5820), 1860–1862. https://doi.org/10.1126/science.1138071

Busemeyer, J. R., & Bruza, P. D. (2012). Quantum models of cognition and decision. Cambridge University Press.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., & Bengio, Y. (2015). A recurrent latent variable model for sequential data. Advances in Neural Information Processing Systems, 28.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Clark, J. J., & Yuille, A. L. (1990). Data fusion for sensory information processing systems. Kluwer Academic Publishers.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. Trends in Cognitive Sciences, 10(7), 294–300. https://doi.org/10.1016/j.tics.2006.05.004

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. Neural Computation, 7(5), 889–904. https://doi.org/10.1162/neco.1995.7.5.889

De Berker, A. O., Rutledge, R. B., Mathys, C., Marshall, L., Cross, G. F., Dolan, R. J., & Bestmann, S. (2016). Computations of uncertainty mediate acute stress responses in humans. Nature Communications, 7, 10996. https://doi.org/10.1038/ncomms10996

Diekelmann, S., & Born, J. (2010). The memory function of sleep. Nature Reviews Neuroscience, 11(2), 114–126. https://doi.org/10.1038/nrn2762

Endres, D. M., & Schindelin, J. E. (2003). A new metric for probability distributions. IEEE Transactions on Information Theory, 49(7), 1858–1860. https://doi.org/10.1109/TIT.2003.813506

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 415(6870), 429–433. https://doi.org/10.1038/415429a

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. Frontiers in Human Neuroscience, 4, 215. https://doi.org/10.3389/fnhum.2010.00215

Feynman, R. P., Leighton, R. B., & Sands, M. (1965). The Feynman lectures on physics, Vol. 3: Quantum mechanics. Addison-Wesley.

Findling, C., Chopin, N., & Koechlin, E. (2023). Imprecise neural computations as a source of adaptive behavioural variability. Nature Communications, 14, 3686. https://doi.org/10.1038/s41467-023-39380-x

FitzGibbon, L., Lau, J. K., & Murayama, K. (2020). The seductive lure of curiosity: information as a motivationally salient reward. Current Opinion in Behavioral Sciences, 35, 21–27. https://doi.org/10.1016/j.cobeha.2020.05.014

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. Psychological Review, 124(1), 91–114. https://doi.org/10.1037/rev0000045

Friston, K. (2005). A theory of cortical responses. Philosophical Transactions of the Royal Society B: Biological Sciences, 360(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. Neural Computation, 29(1), 1–49. https://doi.org/10.1162/NECO_a_00912

Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. Frontiers in Psychology, 3, 130. https://doi.org/10.3389/fpsyg.2012.00130

Gallego, J. A., Perich, M. G., Chowdhury, R. H., Solla, S. A., & Miller, L. E. (2020). Long-term stability of cortical population dynamics underlying consistent behavior. Nature Neuroscience, 23(2), 260–270. https://doi.org/10.1038/s41593-019-0555-4

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. Current Opinion in Neurobiology, 20(2), 251–256. https://doi.org/10.1016/j.conb.2010.02.008

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. Science, 349(6245), 273-278. https://doi.org/10.1126/science.aac6076

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. Psychological Review, 128(5), 936–975. https://doi.org/10.1037/rev0000281

Gomez-Ramirez, J., & Sanz, R. (2013). A model of how the brain discovers and manipulates relational structures. Frontiers in Psychology, 4, 963. https://doi.org/10.3389/fpsyg.2013.00963

Gottlieb, J., & Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. Nature Reviews Neuroscience, 19(12), 758–770. https://doi.org/10.1038/s41583-018-0078-0

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. Trends in Cognitive Sciences, 17(11), 585–593. https://doi.org/10.1016/j.tics.2013.09.001

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. Topics in Cognitive Science, 7(2), 217–229. https://doi.org/10.1111/tops.12142

Haken, H. (1983). Synergetics: An introduction. Springer-Verlag.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. Neuron, 95(2), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

Hasson, U., Chen, J., & Honey, C. J. (2015). Hierarchical process memory: Memory as an integral component of information processing. Trends in Cognitive Sciences, 19(6), 304–313. https://doi.org/10.1016/j.tics.2015.04.006

Heeger, D. J. (2017). Theory of cortical function. Proceedings of the National Academy of Sciences, 114(8), 1773–1782. https://doi.org/10.1073/pnas.1619788114

Hobson, J. A., & Friston, K. J. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. Progress in Neurobiology, 98(1), 82–98. https://doi.org/10.1016/j.pneurobio.2012.05.003

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. Cognition, 108(3), 687–701. https://doi.org/10.1016/j.cognition.2008.05.010

Jaba, T. (2022). Dasatinib and quercetin: short-term simultaneous administration yields senolytic effect in humans. Issues and Developments in Medicine and Medical Research Vol. 2, 22-31.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. Machine Learning, 37(2), 183–233. https://doi.org/10.1023/A:1007665907178

Kaplan, R., & Friston, K. J. (2018). Planning and navigation as active inference. Biological Cybernetics, 112(4), 323–343. https://doi.org/10.1007/s00422-018-0753-2

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. Annual Review of Psychology, 55, 271–304. https://doi.org/10.1146/annurev.psych.55.090902.142005

Khrennikov, A. (2010). Ubiquitous quantum structure: From psychology to finance. Springer.

Kim, Y. H., Yu, R., Kulik, S. P., Shih, Y., & Scully, M. O. (2000). Delayed "choice" quantum eraser. Physical Review Letters, 84(1), 1–5. https://doi.org/10.1103/PhysRevLett.84.1

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114. https://arxiv.org/abs/1312.6114

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. Trends in Neurosciences, 27(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007

Kondo, H. M., Van Ee, R., Nojiri, K., Kitagawa, N., & Kashino, M. (2022). Multiple timescales of the dynamics in bistable perception. Scientific Reports, 12(1), 2045. https://doi.org/10.1038/s41598-022-06014-z

Kounios, J., & Beeman, M. (2014). The cognitive neuroscience of insight. Annual Review of Psychology, 65, 71–93. https://doi.org/10.1146/annurev-psych-010213-115154

Krishnan, R. G., Shalit, U., & Sontag, D. (2017). Structured inference networks for nonlinear state space models. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 31, No. 1).

Kvam, P. D., Pleskac, T. J., Yu, S., & Busemeyer, J. R. (2015). Interference effects of choice on confidence: Quantum characteristics of evidence accumulation. Proceedings of the National Academy of Sciences, 112(34), 10645–10650. https://doi.org/10.1073/pnas.1500688112

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. Behavioral and Brain Sciences, 40, e253. https://doi.org/10.1017/S0140525X16001837

Lewis, P. A., & Durrant, S. J. (2011). Overlapping memory replay during sleep builds cognitive schemata. Trends in Cognitive Sciences, 15(8), 343–351. https://doi.org/10.1016/j.tics.2011.06.004

Liang, Y., Li, Y., Khanna, S., Liu, Y., & Li, J. (2018). Monitoring and diagnosing the causes of anomalies in distributed systems. In Proceedings of the 2018 ACM Symposium on Cloud Computing (pp. 476–488).

Lieder, F., Griffiths, T. L., & Goodman, N. D. (2018). Strategy selection as rational metareasoning. Psychological Review, 125(6), 852–889. https://doi.org/10.1037/rev0000135

Lin, J. (1991). Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory, 37(1), 145–151. https://doi.org/10.1109/18.61115

Linson, A., & Friston, K. (2019). Reframing PTSD for computational psychiatry with the active inference framework. Cognitive, Affective, & Behavioral Neuroscience, 19(3), 651–669. https://doi.org/10.3758/s13415-019-00723-1

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. Nature Neuroscience, 9(11), 1432–1438. https://doi.org/10.1038/nn1790

MacKay, D. J. (2003). Information theory, inference and learning algorithms. Cambridge University Press.

Mackay, D. J. C. (2021). Information theory, inference and learning algorithms (New ed.). Cambridge University Press.

Merali, Z. (2015). The quantum source of space-time. Nature, 527(7578), 290–293. https://doi.org/10.1038/527290a

Meyniel, F., Schlunegger, D., & Dehaene, S. (2015). The sense of confidence during probabilistic learning: A normative account. PLoS Computational Biology, 11(6), e1004305. https://doi.org/10.1371/journal.pcbi.1004305

Mirza, M. B., Adams, R. A., Mathys, C., & Friston, K. J. (2016). Scene construction, visual foraging, and active inference. Frontiers in Computational Neuroscience, 10, 56. https://doi.org/10.3389/fncom.2016.00056

Parr, T., & Friston, K. J. (2018). The discrete and continuous brain: From decisions to movement—And back again. Neural Computation, 30(9), 2319–2347. https://doi.org/10.1162/neco_a_01102

Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press.

Pezzulo, G., Rigoli, F., & Friston, K. (2013). Active inference, homeostatic regulation and adaptive behavioural control. Progress in Neurobiology, 134, 17–35. https://doi.org/10.1016/j.pneurobio.2015.09.001

Pfeifer, R., & Bongard, J. (2006). How the body shapes the way we think: A new view of intelligence. MIT Press.

Pothos, E. M., & Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? Behavioral and Brain Sciences, 36(3), 255–274. https://doi.org/10.1017/S0140525X12001525

Pouget, A., Dayan, P., & Zemel, R. (2003). Inference and computation with population codes. Annual Review of Neuroscience, 26, 381–410. https://doi.org/10.1146/annurev.neuro.26.041002.131112

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. Nature Neuroscience, 19(3), 366–374. https://doi.org/10.1038/nn.4240

Rabinovich, M. I., Friston, K. J., & Varona, P. (Eds.). (2012). Principles of brain dynamics: Global state interactions. MIT Press.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience, 2(1), 79–87. https://doi.org/10.1038/4580

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In Proceedings of the 31st International Conference on Machine Learning (pp. 1278–1286). PMLR.

Riedel, C. J., Zurek, W. H., & Zwolak, M. (2016). The rise and fall of redundancy in decoherence and quantum Darwinism. New Journal of Physics, 18(2), 023010. https://doi.org/10.1088/1367-2630/18/2/023010

Sajid, N., Ball, P. J., Parr, T., & Friston, K. J. (2021). Active inference: Demystified and compared. Neural Computation, 33(3), 674–712. https://doi.org/10.1162/neco_a_01357

Sandkühler, S., & Bhattacharya, J. (2008). Deconstructing insight: EEG correlates of insightful problem solving. PLoS ONE, 3(1), e1459. https://doi.org/10.1371/journal.pone.0001459

Schapiro, A. C., & Turk-Browne, N. B. (2015). Statistical learning. Brain Mapping, 3, 501–506. https://doi.org/10.1016/B978-0-12-397025-1.00326-8

Schwartenbeck, P., FitzGerald, T., Dolan, R. J., & Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. Frontiers in Psychology, 4, 710. https://doi.org/10.3389/fpsyg.2013.00710

Scully, M. O., & Drühl, K. (1982). Quantum eraser: A proposed photon correlation experiment concerning observation and "delayed choice" in quantum mechanics. Physical Review A, 25(4), 2208–2213. https://doi.org/10.1103/PhysRevA.25.2208

Scully, M. O., Englert, B. G., & Walther, H. (1991). Quantum optical tests of complementarity. Nature, 351(6322), 111–116. https://doi.org/10.1038/351111a0

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. Trends in Cognitive Sciences, 18(4), 186–193. https://doi.org/10.1016/j.tics.2014.01.006

Siegler, R. S. (2005). Children's learning. American Psychologist, 60(8), 769–778. https://doi.org/10.1037/0003-066X.60.8.769

Sohn, K., Lee, H., & Yan, X. (2015). Learning structured output representation using deep conditional generative models. Advances in Neural Information Processing Systems, 28.

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: A computational framework and potential neural correlates. Psychological Review, 119(1), 120–154. https://doi.org/10.1037/a0026435

Stickgold, R., & Walker, M. P. (2013). Sleep-dependent memory triage: evolving generalization through selective processing. Nature Neuroscience, 16(2), 139–145. https://doi.org/10.1038/nn.3303

Tegmark, M. (2000). Importance of quantum decoherence in brain processes. Physical Review E, 61(4), 4194–4206. https://doi.org/10.1103/PhysRevE.61.4194

Tervo, D. G. R., Tenenbaum, J. B., & Gershman, S. J. (2016). Toward the neural implementation of structure learning. Current Opinion in Neurobiology, 37, 99–105. https://doi.org/10.1016/j.conb.2016.01.014

Tkemaladze, J. (2023). Reduction, proliferation, and differentiation defects of stem cells over time: a consequence of selective accumulation of old centrioles in the stem cells?. Molecular Biology Reports, 50(3), 2751-2761. DOI : https://pubmed.ncbi.nlm.nih.gov/36583780/

Tkemaladze, J. (2024). Editorial: Molecular mechanism of ageing and therapeutic advances through targeting glycative and oxidative stress. Front Pharmacol. 2024 Mar 6;14:1324446. DOI : 10.3389/fphar.2023.1324446. PMID: 38510429; PMCID: PMC10953819.

Tkemaladze, J. (2026). Old Centrioles Make Old Bodies. Annals of Rejuvenation Science, 1(1). DOI : https://doi.org/10.65649/yx9sn772

Tkemaladze, J. (2026). Visions of the Future. Longevity Horizon, 2(1). DOI : https://doi.org/10.65649/8be27s21

Tognoli, E., & Kelso, J. A. S. (2014). The metastable brain. Neuron, 81(1), 35–48. https://doi.org/10.1016/j.neuron.2013.12.022

Tomov, M. S., Yagati, S., Kumar, A., Yang, W., & Gershman, S. J. (2021). Discovery of hierarchical representations for efficient planning. PLoS Computational Biology, 16(4), e1007594. https://doi.org/10.1371/journal.pcbi.1007594

Tononi, G., & Cirelli, C. (2014). Sleep and the price of plasticity: from synaptic and cellular homeostasis to memory consolidation and integration. Neuron, 81(1), 12–34. https://doi.org/10.1016/j.neuron.2013.12.025

Tschacher, W., & Haken, H. (2007). Intentionality in non-equilibrium systems? The functional aspects of self-organized pattern formation. New Ideas in Psychology, 25(1), 1–15. https://doi.org/10.1016/j.newideapsych.2006.09.002

von Neumann, J. (1932). Mathematical foundations of quantum mechanics. Princeton University Press.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. Cognitive Science, 38(4), 599–637. https://doi.org/10.1111/cogs.12101

Walborn, S. P., Terra Cunha, M. O., Pádua, S., & Monken, C. H. (2002). Double-slit quantum eraser. Physical Review A, 65(3), 033818. https://doi.org/10.1103/PhysRevA.65.033818

Wilson, R. C., & Niv, Y. (2011). Inferring relevance in a changing world. Frontiers in Human Neuroscience, 5, 189. https://doi.org/10.3389/fnhum.2011.00189

Wilson, R. C., Nassar, M. R., & Gold, J. I. (2010). Bayesian online learning of the hazard rate in change-point problems. Neural Computation, 22(9), 2452–2476. https://doi.org/10.1162/NECO_a_00007

Zurek, W. H. (2003). Decoherence, einselection, and the quantum origins of the classical. Reviews of Modern Physics, 75(3), 715–775. https://doi.org/10.1103/RevModPhys.75.715

