A Dynamical Model of Subjectivity: Integrating Affective Gain, Cognitive Bias, and Self-Regulation

Takeo Imaizumi

Independent Researcher

Kyoto City

Japan

Author Note

Correspondence concerning this article should be addressed to Takeo Imaizumi. E-mail: takeoimaizumi@song.ocn.ne.jp

A preprint of this manuscript is available on PsyArXiv. DOI: https://doi.org/10.31234/osf.io/8dbft_v7

Funding Not applicable.

Conflict of Interest The author declares no conflict of interest relevant to the content of this article.

Ethical Approval Not applicable.

Data Availability No datasets were generated or analyzed during the current study.

Code Availability Simulation code for figure generation (core model) is available at https://zenodo.org/records/15877593.

Acknowledgments AI assistance (Gemini 2.5 Pro/Google; OpenAI o3-pro/OpenAI) supported mathematical formulation, figure code generation, initial reference retrieval; the author reviewed all outputs and assumes full responsibility for the content.

Abstract

Background: Explaining how affective gain (G) and cognitive bias (μ) dynamically interact to shape subjective experience is a central challenge in affective and cognitive science. While both constructs are widely studied, the mechanisms governing their interaction and role in individual differences remain poorly understood.

Methods: We developed a control-theoretic dynamical-systems model of the subjective state (M_s) that formalizes G–μ coupling. Bifurcation analysis of the model's potential function yields a Mind Topography Map, a global portrait of stability regimes across the G–μ plane. A higher-order Self-System adaptively navigates this landscape by regulating G and μ via hierarchical Bayesian learning.

Results: Canonical cusp and pitchfork bifurcations organize the landscape, generating qualitative shifts corresponding to psychological phenomena from stable belief convergence to cognitive polarization and mood-like oscillations. We identified an Ideal Dynamical Equilibrium ($G = 1$, $\mu = 0$) as an optimal balance of stability and responsiveness. An exploratory extension (Structural Gain–Bias Dynamics; SGBD) represents person-specific traits with structural matrices to capture mixed-emotion states.

Conclusions: By unifying dynamical-systems analysis with agentic self-regulation, our framework clarifies core subjective dynamics. It provides a tractable route to personalized modeling and yields mathematically precise, falsifiable hypotheses for empirical studies, including longitudinal and neurophysiological designs. Bridging concepts from Mathematical Psychology to control-theoretic frameworks like Perceptual Control Theory and the Free Energy Principle, our model offers a robust theoretical tool for computational psychology, affective science, and psychiatry.

Keywords: Mind Topography Map, Self-System, Gain–Bias Coupling, Bifurcation, Computational Psychiatry

A Dynamical Model of Subjectivity: Integrating Affective Gain, Cognitive Bias, and Self-Regulation

Introduction

The Challenge: Divergence and Dynamics of Subjectivity

The divergence between subjective perception and objective reality—shaped by biases and emotions—poses a central challenge to cognitive science. Understanding how subjective states evolve has become increasingly critical in today's complex information environments. Broad frameworks, from control-theoretic accounts like Perceptual Control Theory (PCT) (Powers, 1973) to comprehensive probabilistic models like the Free Energy Principle (FEP) (K. Friston, 2010), provide powerful paradigms for understanding these dynamics. However, the specific mechanisms governing affective sensitivity, cognitive bias interaction, and their higher-order regulation remain less explored. Static descriptions are insufficient; dynamic models are essential for understanding subjectivity's temporal evolution and potential for transformation. A holistic view is often hindered by fragmented insights across disciplines such as psychiatry, cognitive science, and contemplative studies. In addition, the landscape of subjective experience is characterized by substantial individual differences, shaped by unique histories, traits, and cognitive styles. Understanding not only the general dynamics of subjectivity but also the origins and implications of these individual variations remains a major challenge. While a comprehensive model accounting for all such variability is a formidable goal, establishing foundational principles through focused inquiry is a crucial first step.

A Proposed Framework: Gain, Bias, and Self-Regulation

This paper proposes a computational framework for subjective dynamics rooted in control theory. This focused approach simplifies the rich diversity of individual experiences. It aims to elucidate the core principles governing these dynamics in a tractable manner, allowing for a clear examination of fundamental interactions. The evolution of the subjective state (M_s) thus hinges on three core components, presented here in their scalar form:

- **Affective Gain (G):** This refers to the sensitivity or intensity of response to discrepancies (errors) or internal fluctuations, which modulates reactivity and stability. Unlike FEP's non-negative precision, our model includes $G < 0$. This feature, acting as positive feedback within our control architecture (see Figure 1), allows for the exploration of diverse feedback dynamics. Such negative-gain regimes resemble inverse precision weighting in certain pathological attention loops, a link that anchors the model in emerging computational-psychiatric evidence. The distinction between this internal positive-feedback mechanism and potentially paradoxical behavioral choices is discussed further in the Discussion.
- **Cognitive Bias (μ):** Persistent predispositions, such as beliefs and memories, that attract or repel the subjective state.
- **The Self-System (S):** A higher-level component that regulates G and μ over time through learning, experience, or volition. A detailed computational mechanism for S is proposed in the Computational Mechanism of the Self-System (S) section.

The interaction between these scalar representations of G, μ, and mental noise ϵ (where $\omega = \mu + \epsilon$) shapes a foundational dynamical landscape for subjectivity. Although this scalar model provides a generalizable account, we also recognize the profound importance of individual differences. An extension of this framework to a multidimensional vector space is outlined in Modeling Individual Differences: The SGBD Framework (see Appendix A) (the Structural Gain–Bias Dynamics, or SGBD, framework). This extension explicitly incorporates mechanisms for modeling such individual-specific characteristics (e.g., through structural matrices). However, this paper focuses on the insights gained from an intentionally simplified scalar model. The aim is to first establish a robust baseline understanding that applies across individuals, before delving into the sources of variation. The Mind Topography Map metaphor, visualized on the G–μ plane, illustrates how different G–μ combinations yield distinct qualitative behaviors (e.g., bias convergence, polarization, and dynamical equilibrium).

This approach offers a unifying mathematical formalism that bridges disparate theoretical paradigms. As conceptually illustrated in Figure 2, our framework synthesizes core concepts from Perceptual Control Theory (PCT), the Free Energy Principle (FEP), and Mathematical Psychology. Within this integrative structure, the Mind Topography Map—developed in the subsequent sections—serves as the central analytical tool for investigating the dynamics of subjectivity. It is important to note, however, that the scope of this paper is focused on the internal dynamics of the subjective loop; it provides a high-precision analytical engine for cognitive-affective processes, rather than modeling the entire perception-action cycle central to frameworks like Perceptual Control Theory (PCT).

Dynamical Landscape and Active Self-Creation

A central thesis is that the Mind Topography Map is not static but is actively shaped or navigated by S through adjustments to G and μ. Our framework contrasts with fixed-trait models or those emphasizing passive dynamics on static potentials; instead, it depicts subjectivity as a dynamic, adaptive, and self-constructive process. This active, higher-order regulatory perspective is crucial for understanding intentional self-regulation. For instance, contemplative practices like meditation often aim to cultivate states of profound mental balance and responsive presence. Our model offers a way to conceptualize such states, particularly through the "Ideal Dynamical Equilibrium" (IDE; $G = 1, \mu = 0$, detailed in the IDE State ($G = 1, \mu = 0$) section), which represents a stable yet highly responsive optimum mode of subjective experience. The central aim of this work is therefore to elucidate the dynamic pathways toward desirable states unique to each individual, moving beyond a mere static analysis or classification of those states.

Purpose and Structure

This paper develops and analyzes a dynamical framework to formalize the principles of self-regulated subjectivity. The paper begins by positioning this framework against prior work (Related Work). We then establish its theoretical foundations, detailing the mathematical formulation of the core scalar model (Methods: Internal Control Model) and the computational mechanism of the higher-order Self-System that governs it (Computational Mechanism of the Self-System (S)).

The core results are presented in the Results of the Core Model section, which introduces the Mind Topography Map, analyzes its stability landscape, and characterizes the Ideal Dynamical Equilibrium (IDE) as a key regulatory target. The implications, novelty, and limitations of these findings are explored in the Discussion section, followed by a concluding summary (Conclusion).

While the main body of the paper focuses on establishing the principles of the scalar model, the appendices provide a crucial theoretical extension and a series of multi-faceted explorations. First, Modeling Individual Differences: The SGBD Framework (see Appendix A) details the Structural Gain–Bias Dynamics (SGBD) framework, generalizing the model to a vector-based representation to account for individual differences. Building upon this foundation, the subsequent appendices illustrate the framework's broader utility:

A Narrative Illustration of Self-Regulatory Dynamics (see Appendix B) offers a narrative walkthrough, casting a monk's spiritual quest as a concrete example of a trajectory toward the IDE.

ADHD as Dysregulation in the G–μ–S Framework (see Appendix C) explores a conceptualization of ADHD as a dynamical disorder to illustrate the framework's potential clinical relevance.

And Smith's *Impartial Spectator* and the G–μ–S Framework (see Appendix D) provides a philosophical reflection, connecting the model to Adam Smith's *The Theory of Moral Sentiments* and its core concept of the "Impartial Spectator."

Core Contributions

This study advances computational modeling of subjectivity via four core contributions:

1. **Mind Topography Map**: Unifying stability regimes visually by charting the joint G–μ parameter plane using potential theory.

2. **Gain-dependent Effective Bias ($G\mu$)**: Revealing how affective gain dynamically modulates the influence of cognitive bias.

3. **The Self-System (S) for Higher-Order Control**: Proposing S for explicit, dynamical adjustment of G and μ, with a tractable hierarchical Bayesian learning mechanism (see the Computational Mechanism of the Self-System (S) section) for investigating landscape navigation and reshaping.

4. **IDE**: Characterizing the functionally distinct stable yet responsive state ($G = 1, \mu = 0$) versus static inertia ($G = 0, \mu = 0$).

These contributions bridge control-theoretic, dynamical-systems, and contemplative perspectives, and offer a testbed for computational psychiatry and AI ethics.

Related Work

Prevailing Models of Subjective Dynamics

The study of subjectivity is advanced by several complementary theoretical approaches. Normative frameworks such as the Free Energy Principle (FEP) (K. Friston, 2010) model organisms as prediction-error-minimizing systems. From a computational perspective, Reinforcement Learning (RL) (Sutton & Barto, 1998) provides algorithms for agents to learn optimal, reward-maximizing behaviors. The neural implementation of these ideas is often explored through Predictive Coding (PC), which posits a hierarchy of predictive models and error-passing signals in the brain. From a different standpoint, Perceptual Control Theory (PCT) (Powers, 1973) focuses on behavior as a means to control perceptions to match internal goals. Supporting these frameworks, Dynamical Systems Theory (DST) (Strogatz, 2015) offers the mathematical tools to describe the temporal evolution and qualitative shifts of subjective states. Our understanding of subjectivity has thus deepened through the interplay of these normative, algorithmic, implementational, and mathematical perspectives.

Limitations and Open Challenges

Despite these accomplishments, a unified account of how gain and bias jointly evolve under active self-regulation remains elusive. Many existing models treat G and μ implicitly or in isolation, thereby overlooking their dynamical coupling. Few approaches map a global stability landscape delineating subjective regimes across the G–μ plane, and fewer still propose a formally specified higher-order controller for adapting both parameters. These gaps limit our ability to predict whether internal experience stabilizes, shifts abruptly, or turns maladaptive.

A crucial next step, therefore, is to build on foundational work by placing both gain and bias within a shared potential landscape, a move that would make their moment-to-moment interplay analytically transparent. Addressing this challenge is a central goal of our proposed framework.

Proposed Framework: Key Innovations

Our framework addresses these gaps via three innovations. First, it makes the $G\mu$ interaction explicit, where affective gain dynamically scales cognitive bias effects. Second, by extending conventional treatment of G and μ into a two dimensional plane, it introduces a Mind Topography Map —a global stability map of attractors, thresholds, and bifurcations across the G–μ plane. Third, it formalizes a higher-order Self-System (S) that learns to modulate G and μ via hierarchical Bayesian inference, promoting adaptive stability. These synergistic features thereby provide a remarkably coherent and mechanistic account of self-regulated subjectivity, significantly advancing and unifying prior work.

Detailed Mapping to Normative Theories

To situate our framework within the broader landscape of cognitive science, this section clarifies its relationship with established normative and control-theoretic approaches. Table 1 first provides a high-level comparison, contrasting our model's specific contributions with major paradigms like Perceptual Control Theory (PCT), Reinforcement Learning (RL), and the Free Energy Principle (FEP). Subsequently, to facilitate a deeper dialogue with the most influential contemporary paradigms, Table 2 offers a more detailed conceptual mapping of our model's components to those of FEP and Predictive Coding (PC). We focus specifically on these predictive processing models for two primary reasons. First, FEP/PC has become a dominant framework in computational psychiatry and cognitive neuroscience, providing a common language for understanding subjective states and their dysfunction. Clarifying our model's relationship with this framework is therefore particularly crucial. Second, the concepts of "precision" and "priors" in FEP/PC offer a direct and intuitive analogy to our model's core components of "affective gain (G)" and "cognitive bias (μ)," making them well-suited for a component-level mapping. In contrast, while our framework shares deep structural similarities with PCT, these are better explored through a more nuanced discussion of dynamic mechanisms, such as reorganization (detailed in the Discussion section) and the orthogonality of control hierarchies (formalized in Appendix A), rather than a simple component-to-component table

As these comparisons reveal, a gap remains in the literature: while broad paradigms for cognition and control exist, few offer a specific, tractable model focused on the dynamic interplay of affective gain and cognitive bias under the governance of a higher-order regulatory system. The following section therefore details the mathematical formulation of our proposed model, which is precisely designed to fill this gap by providing a formal engine for the internal subjective loop.

Methods: Internal Control Model

Model Structure and Information Flow

The model conceptualizes the interaction between the internal and external world mediated by the Self-System (S). Figure 1 shows real-world input processed by Recognition Filter (F_r) into an objective model (M_o), then by Interpretation Filter (F_i) into the subjective state (M_s). The discrepancy ($e = M_o - M_s$) is processed by the Evaluation

Filter (F_e). Its output, augmented by internal mental fluctuation $\omega = \mu + \epsilon$ (bias μ + noise ϵ) and scaled by affective gain (G), yields an update signal, UpdateSignal(t), which updates M_s and drives Action (A), affecting the Real World and in turn influencing the Self-System (S). S controls A, modulates filters (F_r, F_i, F_e), and parameters (G, μ, ϵ), enabling dynamical self-regulation as detailed in the Computational Mechanism of the Self-System (S) section. The dotted area highlights the components mathematically formulated in the Mathematical Formulation: The Update Rule for M_s section.

Model Components

The main components are as follows:

- External/Internal: The external environment, and the individual's internal mental realm wherein their subjective state (M_s) resides.
- Filters (F_r, F_i, F_e,): Information processing stages (Recognition, Interpretation, Evaluation), influenced by S.
- Models (M_o, M_s): Objective (M_o) and subjective (M_s) representations of the world.
- Error (e): Discrepancy $e = M_o - M_s$.
- Mental Fluctuation (ω): Internal perturbations, where $\omega = \mu + \epsilon$ (bias μ + noise ϵ), that augment the output of the Evaluation Filter (Fe).
- Cognitive Bias (μ): Persistent attractor/repeller for M_s; influenced by S.
- Noise (ϵ): Transient random fluctuations. Modeled as Gaussian noise $N(0, \sigma_\epsilon^2)$, with its parameters (e.g., variance) potentially modulated by S.
- Affective Gain (G): Scales the influence of error and fluctuation; represents sensitivity/reactivity. $G = 1$ is reference; influenced by S.

- Self-System (S): Core agentic component for higher-order regulation, learning, adaptation, and action. Adjusts G, μ, ϵ, and filters. Modulated by feedback from the Real World.
- Action (A): Behavioral output driven by the update signal; influenced by S.

Mathematical Formulation: The Update Rule for M_s

The subjective state M_s evolves discretely: $M_s(t+1) = M_s(t) + \text{UpdateSignal}(t)$. The update signal, scaled by gain G, is given by:

$$\text{UpdateSignal}(t) = G(S,t) \cdot [Fe(M_o(t) - M_s(t), S, t) + \mu(S,t) + \epsilon(S,t)]$$

This highlights that G, μ, ϵ, and filters are modulated by S and are time-dependent.

For core analysis of M_s dynamics under fixed G and μ (before adaptation by S), we introduce several simplifications to clarify the fundamental interactions. Firstly, we assume the objective model M_o is constant at zero ($M_o = 0$, $F_i(M_o) = 0$), meaning the error e is simply $-M_s(t)$. Secondly, the evaluation filter F_e is taken as a linear function that directly passes the error, i.e., $F_e(e) = e$. Under these conditions, the core of the update signal becomes $G \cdot [-M_s(t) + \mu + \epsilon(t)]$. This would lead to an update rule where $M_s(t+1) = M_s(t) + G[-M_s(t) + \mu + \epsilon(t)] = (1-G)M_s(t) + G\mu + G\epsilon(t)$.

However, subjective states, while dynamic, generally do not escalate indefinitely, as inherent psychological or neurological regulatory mechanisms tend to ensure stability and prevent extreme, unbounded states. To capture this essential property of boundedness and to model the system's capacity for stable equilibria, a standard nonlinear saturation term, $-\alpha M_s(t)^3$ (with $\alpha > 0$), is incorporated into the dynamics (Strogatz, 2015). This term acts as a symmetric restoring force, pulling M_s towards a central point when it deviates significantly, a common feature in models of complex, self-regulating systems.

By integrating this nonlinear saturation term with the previously derived linear components, the simplified discrete update rule for M_s (for $\Delta t = 1$) is established as:

$$M_s(t+1) = (1-G)M_s(t) - \alpha M_s(t)^3 + G\mu + G\epsilon(t) \tag{1}$$

Here, $(1-G)M_s(t)$ represents linear feedback, $G\mu + G\epsilon(t)$ is the gain-scaled internal fluctuation $G\omega(t) = G(\mu + \epsilon(t))$, $-\alpha M_s^3$ is the symmetric restoration term.

Mathematical Analysis of M_s Dynamics (Fixed G and μ)

To understand M_s dynamics for fixed G and μ, we analyze the deterministic part of Eq. (1) (neglecting $G\epsilon(t)$), i.e., $M_s(t+1) = f(M_s)$:

$$f(M_s) = (1-G)M_s - \alpha M_s^3 + G\mu \tag{2}$$

The quasi-potential function $V(M_s)$ (Eq. (3)), as a conceptual analogy to continuous potentials, links its minima to stable attractors (Strogatz, 2015; Thom, 1989), where C is the constant:

$$V(M_s) = -\frac{(1-G)}{2}M_s^2 + \frac{\alpha}{4}M_s^4 - G\mu M_s + C \tag{3}$$

Fixed Point Analysis

Fixed points, M_s^*, satisfy $\alpha(M_s^*)^3 + GM_s^* - G\mu = 0$ (from Eq. (2) with $M_s(t+1) = M_s(t)$).

- **Case 1:** $\mu = 0$**.** One solution is $M_s^* = 0$. If $G < 0$, there are two additional solutions: $M_s^* = \pm\sqrt{-G/\alpha}$. This reflects a symmetric pitchfork bifurcation at $G = 0$.
- **Case 2:** $\mu \neq 0$**.** One or three real solutions, corresponding to an imperfect pitchfork bifurcation due to the asymmetry introduced in $V(M_s)$ by the cognitive bias μ.

Linear Stability Analysis

Stability of M_s^* is determined by $f'(M_s^*) = (1-G) - 3\alpha(M_s^*)^2$ (from Eq. (2)). M_s^* is locally stable if $|f'(M_s^*)| < 1$. For $\mu = 0$:

- $M_s^* = 0$ **stability:** Stable for $0 < G < 2$. Loses stability at $G = 0$ (pitchfork) and $G = 2$ (flip).
- $M_s^* = \pm\sqrt{-G/\alpha}$ **stability (**$G < 0$**):** Stable for $-1 < G < 0$. Loses stability via flip at $G = -1$.

For $\mu \neq 0$, stability of each root of $\alpha(M_s^*)^3 + GM_s^* - G\mu = 0$ requires $|(1 - G) - 3\alpha(M_s^*)^2| < 1$.

Bifurcation Analysis

Bifurcations (qualitative changes in dynamics) occur when $|f'(M_s^*)| = 1$.

- $f'(M_s^*) = 1$ **(Saddle-node):** Leads to $-G = 3\alpha(M_s^*)^2$ (requires $G \leq 0$). For $\mu \neq 0$, this occurs along the curve defined by:

  $$G = -\frac{27\alpha}{4}\mu^2 \tag{4}$$

  (This is derived by eliminating M_s^* from $\alpha(M_s^*)^3 + GM_s^* - G\mu = 0$ and $-G = 3\alpha(M_s^*)^2$.) This curve is the $G < 0$ boundary for double-well potential (bistability) appearance/disappearance.

- $f'(M_s^*) = -1$ **(Flip):** Leads to $2 - G = 3\alpha(M_s^*)^2$. Occurs at $G = 2$ (for $M_s^* = 0$) and $G = -1$ (for $M_s^* = \pm\sqrt{1/\alpha}$ when $\mu = 0$).

These analytical results for fixed G and μ form the basis for the stability landscape (see the The G–μ Parameter Plane: A Stability Landscape section). The Self-System (S) then navigates or reshapes this landscape by dynamically altering G and μ.

Modeling Approach: Scalar Representation for General Principles and a Vectorial Extension for Individual Differences

This paper employs a dual-pronged modeling approach to balance the investigation of general principles with the need to account for individual differences in subjectivity. This strategy consists of two key components:

First, the core of this paper develops and analyzes a foundational scalar model. This approach uses scalar quantities for key variables (M_s, G, μ) to elucidate the fundamental, universal dynamics of the gain–bias interaction, leading to the creation of the Mind Topography Map. The conclusions presented in the main body of this paper are derived exclusively from this focused, generalizable model.

Second, to address the limitations of a purely scalar approach and provide a path toward personalized modeling, we introduce the Structural Gain–Bias Dynamics (SGBD) framework in Modeling Individual Differences: The SGBD Framework (see Appendix A). This exploratory extension utilizes vector and matrix representations to explicitly model person-specific traits and more complex subjective states.

To clearly delineate the respective roles and characteristics of the foundational scalar model and its vectorial SGBD extension, Table 3 presents a direct comparison.

Computational Mechanism of the Self-System (S)

Motivation and Adaptive Role of S

The Self-System (S) acts as a higher-order controller that dynamically adjusts affective gain (G) and cognitive bias (μ). While the Methods: Internal Control Model section describes dynamics for fixed G and μ, S enables adaptive navigation of the Mind Topography Map (Figure 3) to steer M_s towards functional regimes such as IDE ($G = 1, \mu = 0$), or other context-appropriate equilibria. This section details a tractable hierarchical Bayesian learning framework for how S achieves adaptive regulation in response to ongoing experience (M_s) and external context (represented by the objective model M_o), providing a mechanistic account of agentic self-regulation.

Hierarchical Bayesian Learning in S: Probabilistic Formulation

We model S as an agent that performs Bayesian inference and meta-learning across at least two hierarchical levels on different timescales. At a faster regulatory timescale, t_S, S determines suitable (G, μ) settings via Bayesian inference, optimizing a posterior over (G, μ) pairs:

$$P(G, \mu \mid M_{s,\text{obs}}, M_o, \boldsymbol{\theta}_\text{S}, M_\text{int}) \propto P(M_{s,\text{obs}} \mid G, \mu, M_o, M_\text{int}) P(G, \mu \mid \boldsymbol{\theta}_\text{S}) \tag{5}$$

where $M_{s,\text{obs}}$ is the observed or inferred subjective state (or history $\{M_s\}_\text{hist}$), M_o is the external context, M_int is the Self-System's internal model of M_s dynamics (e.g., Eq. (1)), and $\boldsymbol{\theta}_\text{S}$ are meta-parameters governing the prior $P(G, \mu \mid \boldsymbol{\theta}_\text{S})$, subject to slower learning (see the Learning of Meta-Parameters ($\boldsymbol{\theta}_\mathrm{S}$) via an Objective Function ($\mathcal{L}$) section).

The likelihood $P(M_{s,\mathrm{obs}} \mid G, \mu, M_o, M_\mathrm{int})$ quantifies how well (G, μ) explains $M_{s,\mathrm{obs}}$. This is based on $M_{s,\mathrm{obs}}$ being an observation of the true state M_s (dynamics per M_int, e.g., Eq. (1)) subject to observation noise, thereby reflecting both state dynamics and the observation process. For instance, a concrete choice for the likelihood, assuming Gaussian observation noise, is $M_{s,\mathrm{obs}}(t) \sim \mathcal{N}\Big(M_s(t), \sigma^2_\mathrm{obs}\Big)$, where σ^2_obs represents the variance of this observation noise. The prior $P(G, \mu \mid \boldsymbol{\theta}_\mathrm{S})$ reflects the Self-System's current beliefs or policies about (G, μ), shaped by $\boldsymbol{\theta}_\mathrm{S}$ (e.g., $\boldsymbol{\theta}_\mathrm{S}$ might bias towards IDE). S selects an optimal pair (G^*, μ^*) (e.g., via Maximum A Posteriori (MAP) estimation from the posterior) to regulate M_s.

Learning of Meta-Parameters ($\boldsymbol{\theta}_\mathrm{S}$) via an Objective Function ($\mathcal{L}$)

The Self-System's adaptive capability stems from its ability to update its meta-parameters $\boldsymbol{\theta}_\mathrm{S}$ over longer timescales. This learning involves maximizing an *objective function* $\mathcal{L}(\{M_s\}_\mathrm{traj}, M_o \mid \boldsymbol{\theta}_\mathrm{S}^{(k)})$, which evaluates the long-term consequences of state trajectories resulting from the Self-System's choices of (G^*, μ^*) (themselves influenced by $\boldsymbol{\theta}_\mathrm{S}$). A general form for the update rule of $\boldsymbol{\theta}_\mathrm{S}$ is then given by stochastic gradient ascent:

$$\boldsymbol{\theta}_\mathrm{S}^{(k+1)} = \boldsymbol{\theta}_\mathrm{S}^{(k)} + \eta_\theta \nabla_{\boldsymbol{\theta}_\mathrm{S}} \mathbb{E}[\mathcal{L}(\{M_s\}_\mathrm{traj}, M_o \mid \boldsymbol{\theta}_\mathrm{S}^{(k)})] \tag{6}$$

where k is an epoch, and η_θ is the learning rate for meta-parameters $\boldsymbol{\theta}_\mathrm{S}$ (meta-parameter learning rate). $\mathcal{L}$ can reflect:

1. **Predictive Accuracy / Free Energy Minimization:** Aligning with FEP (K. Friston, 2010; K. J. Friston, 2019).

2. **Goal Attainment / Homeostatic Regulation:** e.g., proximity to target states $M_{s,\mathrm{target}}$.

3. **Stability, Flexibility, and Efficiency:** Promoting predictable yet adaptable dynamics.

Such learning refines $P(G, \mu \mid \boldsymbol{\theta}_{\mathrm{S}})$, effectively learning a policy for setting G and μ. The objective function $\mathcal{L}$ quantifies the desirability of long-term state trajectories (e.g., by penalizing IDE deviations, minimizing prediction errors, or promoting stability/adaptability). The expectation $\mathbb{E}[\cdot]$ in Eq. (6) is over state trajectories $\{M_s\}_{\mathrm{traj}}$ under current meta-parameters $\boldsymbol{\theta}_{\mathrm{S}}^{(k)}$. Practically, this expectation and its gradient may be approximated via simulation (e.g., Monte Carlo methods) due to trajectory distribution complexity. Such gradient estimation aligns with established approaches, for instance, policy gradient methods in reinforcement learning if $\mathcal{L}$ is viewed as a reward signal, or techniques from variational inference depending on the formulation.

As a prototype, if one aims to minimize prediction error and deviations from IDE, a cost function C could be

$$C = \sum_{t=1}^{T} \|M_{s,\mathrm{obs}}(t) - M_s(t)\|^2 + \lambda\Big[(G-1)^2 + \mu^2\Big]$$

(where T is the trajectory length and λ is a regularization parameter). The objective function to be maximized by Eq. (6) would then be $\mathcal{L} = -C$. Furthermore, to ensure distinct learning phases and prevent double-counting of information, $\boldsymbol{\theta}_{\mathrm{S}}$ is typically updated on a slower timescale; for example, every T_{meta} (e.g., $T_{\mathrm{meta}} = 50$) steps, ensuring $t_S \ll T_{\mathrm{meta}}$.

Algorithmic Sketch and Heuristic Implementations

Simpler heuristic mechanisms, like gradient-based adjustment of G and μ to minimize an immediate cost $J(G, \mu)$ (e.g., deviation from targets $G_{\mathrm{target}}, \mu_{\mathrm{target}}$ derived from $\boldsymbol{\theta}_{\mathrm{S}}$), can complement this. These are given by:

$$G(t_S + 1) = G(t_S) - \eta_G \frac{\partial J}{\partial G} \tag{7}$$

$$\mu(t_S + 1) = \mu(t_S) - \eta_\mu \frac{\partial J}{\partial \mu} \tag{8}$$

This could implement step 6 of Algorithm 1 or operate in parallel.

Algorithm 1 Self-System Hierarchical Bayesian Learning and Regulation (Conceptual Outline)

```
Inputs: Meta-parameters $\boldsymbol{\theta}_{\mathrm{S}}$; internal model $M_{\mathrm{int}}$.
procedure RegulateAndLearnCycle
    // Fast timescale ($t_S$): Infer and Apply $(G^*, \mu^*)$
    Observe/infer $\{M_s\}_{\mathrm{hist}}$ and context $M_o$.
    Construct prior $P(G, \mu \mid \boldsymbol{\theta}_{\mathrm{S}})$.
    Construct likelihood $P(\{M_s\}_{\mathrm{hist}} \mid G, \mu, M_o, M_{\mathrm{int}})$.
    Compute posterior $P(G, \mu \mid \{M_s\}_{\mathrm{hist}}, M_o, \boldsymbol{\theta}_{\mathrm{S}}, M_{\mathrm{int}})$ (Eq. (5)).
    Select optimal $(G^*, \mu^*)$ (e.g., via MAP from posterior).
    Deploy this pair to modulate $M_s$ dynamics (e.g., Eq. (1)).
    Collect resulting $\{M_s\}_{\mathrm{traj}}$ and outcomes.
    // Slower timescale: Update $\boldsymbol{\theta}_{\mathrm{S}}$
    if time for meta-parameter update then
        Evaluate objective $\mathcal{L}$(Collected Outcomes $\mid \boldsymbol{\theta}_{\mathrm{S}}$).
        Update meta-parameters: $\boldsymbol{\theta}_{\mathrm{S}} \leftarrow \boldsymbol{\theta}_{\mathrm{S}} + \eta_\theta \nabla_{\boldsymbol{\theta}_{\mathrm{S}}} \mathbb{E}[\mathcal{L}]$ (Eq. (6)).
    end if
end procedure
```

The Self-System: Theoretical Grounding and Developmental Outlook

This section established the computational Self-System (S; see Algorithm 1) as a theoretically plausible and computationally specified foundation for agentic self-regulation, capable of generating adaptive trajectories (e.g., towards IDE; see Figure 9). Its hierarchical Bayesian learning aligns with normative frameworks like the FEP (K. Friston, 2010; K. J. Friston, 2019) and resonates with Meta-Reinforcement Learning (Duan et al., 2016; Finn et al., 2017), supporting its validity. This grounding in established principles not only validates our approach but also suggests promising pathways for future S refinement towards enhanced explanatory power, using, for instance, advanced Bayesian inference techniques. A detailed discussion of the associated challenges, including empirical validation and parameter identifiability, is deferred to the Limitations and Future Directions section.

Results of the Core Model

The G–μ Parameter Plane: A Stability Landscape

The Mind Topography Map (see Figure 3) shows how the qualitative dynamics of the subjective state M_s, governed by the deterministic part of Eq. (1),

$$M_s(t+1) = f(M_s; G, \mu) = (1-G)M_s - \alpha M_s^3 + G\mu, \quad \text{with } \alpha = 0.1. \tag{9}$$

depend on affective gain G and cognitive bias μ. The G–μ plane is partitioned into **Single Stable**, **Bistable**, and **Unstable** regimes. This partitioning is based on the number of real (n_{fp}) and linearly stable (n_{stab}) fixed points ($|f'(M_s^*)| < 1$), which are numerically evaluated for each (G, μ) pair. These regimes are visually distinguished by color in Figure 3 (see figure caption for legend). This section outlines the bifurcation boundaries and detailed regime characteristics.

Bifurcation Boundaries Shaping the Landscape

Bifurcations, demarcating transitions between stability regimes, occur along the following curves and lines in the G–μ plane (Figure 3):

1. **Saddle–Node (Cusp) Curve:** Defined by Eq. (10) (reproduced from Eq. (4); dashed line), this curve marks where pairs of non-zero fixed points are created or annihilated for $G < 0$, often corresponding to single/double-well potential transitions.

$$G = -\frac{27\alpha}{4}\mu^2, \quad (G < 0). \tag{10}$$

2. **Pitchfork Bifurcation Axis:** This occurs along the line $G = 0$ (solid vertical line in Figure 3). For $\mu = 0$, a symmetric pitchfork bifurcation occurs; for $\mu \neq 0$, it is an imperfect pitchfork.

3. **Flip (Period–Doubling) Bifurcation Set:** A fixed point M_s^* loses stability via a flip bifurcation when $f'(M_s^*) = -1$, leading to $2 - G = 3\alpha(M_s^*)^2$. This set (dotted

lines) includes (Eq. (11)):

$$G = 2 \quad \text{(Flip for } M_s^* = 0 \text{ when } \mu = 0) \tag{11a}$$

$$G = -1 \quad \text{(Flip for } M_s^* = \pm\sqrt{-G/\alpha} \text{ when } \mu = 0) \tag{11b}$$

$$\mu^2 = \frac{4(1+G)^2(2-G)}{27\alpha G^2}, \quad (\text{for } \mu \neq 0;\ \text{valid when } G < 2,\ G \notin \{0, -1\}) \tag{11c}$$

Eq. (11c) results from eliminating M_s^* between the flip condition $2 - G = 3\alpha(M_s^*)^2$ and the fixed point condition $\alpha(M_s^*)^3 + GM_s^* - G\mu = 0$. The limiting points $(G, \mu) = (2, 0)$ and $(-1, 0)$ are recovered as $|\mu| \to 0$.

Stability Regimes: Definitions and Characteristics

The bifurcation boundaries partition the G–μ plane. The classification in Figure 3 relies on numerically counting stable equilibria (n_{stab}). Let $G_{\text{cusp}}(\mu) = -27\alpha\mu^2/4$. The regimes are:

- **Single Stable Region** (Figure 3: White)
  - *Fixed Points* (n_{fp}, n_{stab})*:* $(1, 1)$.
  - *Typical Conditions (Heuristic):* $(0 < G < 2$ and $|\mu| \approx 0)$ or $(G > G_{\text{cusp}}(\mu))$ and outside flip conditions (Eq. (11)).
  - *Characteristic Dynamics:* Convergence to a unique stable state M_s^*. (e.g., Figures 4, 5).
- **Bistable Region** (Figure 3: Dark Gray)
  - *Fixed Points* (n_{fp}, n_{stab})*:* $(3, 2)$.
  - *Typical Conditions (Heuristic):* Typically occurs when $G < G_{\text{cusp}}(\mu)$ (Eq. (10)) with $G < 0$, and when the system parameters fall within the flip bifurcation boundaries (defined by Eq. (11)))
  - *Characteristic Dynamics:* Two stable states M_s^*; outcome depends on initial state/noise. (e.g., Figures 6, 7).

- **Unstable Region** (Figure 3: Hatched)
  - *Fixed Points* (n_{fp}, n_{stab})*:* (≤ 3, 0).
  - *Typical Conditions (Heuristic):* Violates stability conditions of Eq. (11) (e.g., $G \geq 2$ or $G \leq -1$ for $|\mu| \approx 0$); or inside cusp but all fixed points unstable.
  - *Characteristic Dynamics:* No stable fixed points; potential for oscillations or chaos.

Note that $G_{\mathrm{cusp}}(\mu)$ (Eq. (10)) indicates n_{fp} change (for $G < 0$) but not necessarily $n_{\mathrm{stab}} = 2$. E.g., for $(G, \mu) = (-0.5, 1)$ with $\alpha = 0.1$, $G = -0.5 > G_{\mathrm{cusp}}(1) \approx -0.675$, yet it is Single Stable (Figure 3), as the heuristic conditions described above are not exhaustive and require numerical verification for precise classification.

Interpreting Trajectories on the Map

The Mind Topography Map helps interpret how the Self-System (S) might steer subjective experience by adjusting G and μ:

- Navigating a **Single Stable** region typically ensures predictable convergence. The IDE at $(G, \mu) = (1, 0)$ (see the IDE State ($G = 1, \mu = 0$) section) is a key target here.
- Entering the **Bistable** region can capture the system in one of two attractors, with noise potentially inducing switches. The trajectory in Figure 3 illustrates this.
- Transitioning into an **Unstable** region implies loss of simple equilibrium-seeking, risking oscillations or erratic fluctuations.
- Crossing a bifurcation boundary signifies emerging potential for qualitatively different dynamics, although an immediate dramatic shift in $M_s(t)$ does not necessarily occur, especially if the system's current state is not directly impacted by the bifurcation.

This map is a tool for analyzing adaptive self-regulation and its potential dysfunctions.

Dynamics under Fixed Parameters: Simulation Examples

Simulations of Eq. (1) ($\sigma_\epsilon = 0.05$) with fixed G and μ illustrate these dynamics (parameters marked in Figure 3).

- $G > 1$ (Figure 4, e.g., $G = 1.5, \mu = \pm 0.5$): Rapid convergence towards μ.
- $0 < G < 1$ (Figure 5, e.g., $G = 0.3, \mu = \pm 0.5$): Slower convergence towards μ.
- $G < 0$ (Figures 6 and 7, e.g., $G = -0.5, \mu = \pm 0.5$): Asymmetric stabilization and bistability (confirmed by distributions).
- $G = 1, \mu = 0$ (IDE; Figure 8): M_s decays to $M_s = 0$ amidst fluctuations.

In bistable regimes (typically $G < 0, \mu \neq 0$; Figure 3), the $G\mu$ term asymmetrically shapes the potential landscape $V(M_s)$, amplifying bias μ and influencing initial attractor choice (cf. Figure 6). Noise $\epsilon(t)$ then enables transitions between these stable attractors by overcoming potential barriers, leading to bimodal distributions (Figure 7). These shifts, relevant for modeling polarized states, depend on noise intensity and barrier height.

IDE State ($G = 1, \mu = 0$)

IDE ($G = 1, \mu = 0$) is a functionally significant target for regulation by the Self-System (S). The "Ideal" in its name (IDE) highlights its specific mathematical properties within the model, not any inherent axiological or ethical implications.

Dynamics and Functional Interpretation

At IDE, Eq. (1) simplifies to $M_s(t+1) = -\alpha M_s(t)^3 + \epsilon(t)$. The linear term vanishes, and $M_s^* = 0$ is strongly stable ($f'(0) = 0$). This state combines stability (robust correction towards $M_s = 0$) with flexibility (direct responsiveness to fluctuations $\epsilon(t)$ without linear inertia), representing an adaptable, centered state. Given the absence of the linear term, dynamics near equilibrium are primarily driven by the cubic term and noise $\epsilon(t)$. The probability of large, noise-induced excursions under these specific conditions therefore warrants further investigation.

Contrast with Static Inertia *($G = 0, \mu = 0$)*

In contrast, at $G = 0, \mu = 0$ (static inertia), $M_s(t+1) = M_s(t) - \alpha M_s(t)^3$. Noise influence is absent, and $M_s^* = 0$ is only neutrally stable ($f'(0) = 1$). This state is unresponsive. The IDE, therefore, represents dynamic stability, not mere disconnection. Table 4 summarizes these differences.

Self-System Regulation Towards IDE: A Simulation Example

Adaptive self-regulation, as modeled by the Self-System (S), can steer subjective dynamics (M_s) towards functional equilibria such as IDE ($G = 1, \mu = 0$). The IDE, characterized in the IDE State ($G = 1, \mu = 0$) section, represents a balanced state of stability and responsiveness, reminiscent of concepts like the "Middle Way" in contemplative traditions (Davis & Vago, 2013; Desbordes et al., 2015) and classical philosophical accounts of self-regulation, such as Adam Smith's "Impartial Spectator" (Smith's *Impartial Spectator* and the G–μ–S Framework) (see Appendix D). The Self-System achieves this by dynamically adjusting G and μ based on its learned meta-parameters ($\boldsymbol{\theta}_{\mathrm{S}}$) and ongoing evaluation of experience, as detailed in the Computational Mechanism of the Self-System (S) section. Figure 9 conceptually illustrates such a trajectory, where S navigates the Mind Topography Map (cf. Figure 3) towards the IDE. This distinguishes dynamic stability from static inertia ($G = 0, \mu = 0$), highlighting that effective self-regulation targets optimal responsiveness, not merely bias absence. This active balancing act by S is crucial for adaptive functioning (Smith et al., 2021; van Vugt et al., 2019). For a narrative illustration of a self-regulatory trajectory towards IDE, see A Narrative Illustration of Self-Regulatory Dynamics (see Appendix B).

Discussion

Summary of Key Findings and Contributions

This paper introduced a dynamical model of subjectivity centered on affective gain (G), cognitive bias (μ), and their regulation by a Self-System (S). We identified and described dynamic response patterns using the Mind Topography Map for fixed G and μ, and distinguished the functionally important IDE ($G = 1, \mu = 0$) from static inertia. Critically, we detailed a computational mechanism for S as a hierarchical Bayesian learner (see the Computational Mechanism of the Self-System (S) section) capable of adaptively modulating G and μ to navigate this landscape. These elements offer a unified, mechanistic view of self-regulated subjective dynamics.

The Mind Topography Map: Interpreting the Landscape of Subjectivity

The G–μ plane (Figure 3) serves as a conceptual Mind Topography Map. This landscape, shaped by affective gain (G) and cognitive bias (μ) for fixed parameter values, provides a terrain that the Self-System (S) explores and modulates. The following subsections detail key interpretations of this dynamical landscape.

Core Features and Their Psychological Significance

Potential minima on this map represent attractor states within the subjective landscape, like established beliefs or moods. Different map regions correspond to distinct psychological phenomena; for example, single-well potentials ($G > 0$) suggest convergence, while double-well potentials ($G < 0$ and $\mu \neq 0$) can model polarized thinking. Bifurcation boundaries are critical, marking potential for abrupt qualitative shifts in subjective experience (Witkiewitz & Marlatt, 2007). Such points can be traversed via S-driven changes in G and μ.

The Self-System: Agentic Navigation and Personalized Equilibria

The dynamic malleability of this landscape is central, highlighting the agentic role of the Self-System (S). Through its hierarchical Bayesian learning mechanism (detailed in the Computational Mechanism of the Self-System (S) section), S actively modulates G and μ, thereby adaptively shaping an individual's subjective experience. This adaptive shaping by S implies that equilibrium targets extend beyond a universal ideal. The IDE ($G = 1, \mu = 0$), for instance, serves as a key exemplar of a functionally optimal state; however, an individual's unique dynamical equilibrium points are ultimately determined by their specific Self-System objective function ($\mathcal{L}$) and learned meta-parameters ($\boldsymbol{\theta}_{\mathrm{S}}$).

Navigating Stability's Ambivalence: The Path to Authentic Well-being

Interpreting the map reveals the ambivalence of stability in subjective experience. Not all stable attractors (potential minima) are inherently adaptive or desirable. An equilibrium misaligned with an individual's core values (their $\mathcal{L}$) can be suboptimal, hindering growth and autonomy, showing that stability is not invariably beneficial for flourishing. The S thus faces the profound challenge of distinguishing maladaptive equilibria from those genuinely aligned with an individual's authentic $\mathcal{L}$. Pursuing true well-being by escaping suboptimal attractors requires dynamic processes such as: recalibrating G, re-evaluating μ, refining $\boldsymbol{\theta}_{\mathrm{S}}$ learning, and clarifying $\mathcal{L}$. This perspective frames self-regulation as a continuous alignment of internal dynamics with an autonomously chosen, deeply valued way of living.

Quadrant Typology and Dynamic Heuristics

A common approach in cognitive science is to understand complex psychological phenomena by mapping them onto a low-dimensional space defined by fundamental axes, such as the circumplex model of affect (valence vs. arousal). Following this tradition, we can interpret the Mind Topography Map (Figure 3) by treating affective gain (G) and cognitive bias (μ) as two orthogonal, fundamental axes governing subjective dynamics. Analyzing the four quadrants formed by the signs of G and μ thus allows us to derive a powerful set of psychodynamic heuristics.

The meaning of each axis's sign provides the foundation for this typology. First, the affective gain (G) governs the system's dynamical stability. A positive gain ($G > 0$) promotes a stabilizing, error-correcting response, as the feedback term in Eq. (1) counteracts deviations of M_s from the state implied by the bias. We label this "Adaptive". In contrast, a negative gain ($G < 0$) promotes a destabilizing, error-amplifying response that can lead to polarized states, as it pushes M_s further away from a central reference. We label this "Escapist". Second, the cognitive bias (μ) represents the's default reference point or attractor. A positive bias ($\mu > 0$) reflects a predisposition toward favorable outcomes, interpreted as an *optimistic* stance. Conversely, a negative bias ($\mu < 0$) indicates a focus on potential threats, corresponding to a *pessimistic* or *vigilant* stance. The combination of these two axes logically yields the four distinct signatures detailed in Table 5.

- **Adaptive Optimism ($G > 0, \mu > 0$):** This quadrant represents a healthy, resilient state of goal pursuit. The optimistic bias ($\mu > 0$) provides motivation, while the adaptive gain ($G > 0$) ensures that feedback from reality is used to flexibly correct course.

- **Escapist Optimism ($G < 0, \mu > 0$):** This state is characterized by ungrounded euphoria or wishful thinking. The escapist gain ($G < 0$) leads to a detachment from corrective feedback, amplifying belief in a positive outcome ($\mu > 0$) regardless of evidence.

- **Escapist Pessimism ($G < 0, \mu < 0$):** This dynamic underlies ruminative or helpless states. A negative event can trigger a catastrophic interpretation ($\mu < 0$), which is then amplified by the escapist gain ($G < 0$), locking the system into a self-reinforcing loop of negative thought and affect.

- **Adaptive Vigilance ($G > 0, \mu < 0$):** This is not mere pessimism, but a functional, cautious stance. It reflects a tendency to anticipate potential risks ($\mu < 0$) while constructively addressing them as they arise ($G > 0$), a dynamic central to effective risk management and prudent planning.

This static typology must be interpreted dynamically, as its boundaries are bifurcation points that induce hysteresis (path-dependence). This property, where the system's state depends on its history, creates a fundamental trade-off. It confers stability by buffering against transient fluctuations, but it also creates inertia that resists change from established states, whether they are adaptive or maladaptive. Managing this trade-off between stability and rigidity is a core function of the Self-System (S). The Self-System's task is to exert sustained regulatory control over G and μ to overcome this inertia when necessary. This explains why psychological change often requires gradual and persistent effort, framing the map as a landscape for agentic navigation toward desired equilibria.

The Role and Implications of Noise in Subjective Dynamics

Noise ($\epsilon(t)$), as incorporated in the core model (Eq. (1)), plays a multifaceted role in subjective dynamics:

- It causes fluctuations around attractors (evident in Figures 4–8).
- It can induce transitions between stable states in bistable regimes, potentially offering pathways for change (Strogatz, 2015).
- Its influences are particularly pronounced near bifurcation points, junctures of heightened sensitivity (Kuznetsov, 2013).
- It may destabilize trajectories during S-driven dynamic changes, especially if navigating unstable regions of the Mind Topography Map.
- Conversely, noise can act as a catalyst, providing impetus for S to escape rigid states (deep potential wells influenced by cognitive bias μ) and fostering opportunities for new perspectives or learning.

Thus, while often disruptive, noise also presents constructive potential by perturbing fixed patterns. Understanding how S might manage or leverage this duality, perhaps via gain (G) adjustments to maintain a structured openness, is key for a comprehensive model of adaptive subjectivity.

This role of noise is especially crucial for understanding the mechanism of what Perceptual Control Theory (PCT) terms "reorganization." In our model, this process can be understood as a synthesis of stochastic exploration and deterministic transition. A random "kick" from the noise term, $\epsilon(t)$, can push the system state across a bifurcation boundary on the Mind Topography Map. Once across the threshold, the system deterministically "snaps" into a qualitatively new attractor basin, thereby achieving a landscape-level shift analogous to PCT's concept of reorganization.

Potential Relevance to Computational Psychiatry

The proposed framework, particularly with S as a hierarchical Bayesian learner (as detailed in the Computational Mechanism of the Self-System (S) section), offers insights for computational psychiatry (Adams et al., 2016; K. J. Friston et al., 2014; Huys et al., 2016):

- **Maladaptive Parameterizations:** Persistent dysfunctional states can be framed as consequences of suboptimally learned meta-parameters $\boldsymbol{\theta}_{\mathrm{S}}$ within S, or maladaptive priors $P(G, \mu \mid \boldsymbol{\theta}_{\mathrm{S}})$.

- **Abrupt State Transitions/Bistability:** Sudden changes in subjective states (Witkiewitz & Marlatt, 2007) may reflect failures by S in navigation or stabilization, due to inappropriately learned G and μ settings.

- **Noise-Induced Destabilization (ϵ):** As discussed in the The Role and Implications of Noise in Subjective Dynamics section, stress-amplified noise can compromise the Self-System's regulatory capacity or disrupt its learning (e.g., distorting $\mathcal{L}$ evaluation or $\boldsymbol{\theta}_{\mathrm{S}}$ updates), potentially contributing to psychopathology.

- **Self-System Dysfunction:** Impairments in the Self-System's Bayesian learning (e.g., aberrant $\mathcal{L}$, learning rates η_{θ}, or $\boldsymbol{\theta}_{\mathrm{S}}$) could contribute to the onset and maintenance of psychopathology.

This mechanistic perspective suggests interventions focused on the learning processes of S, its meta-parameters, or inputs. For a concrete application of this approach, a conceptualization of Attention-Deficit/Hyperactivity Disorder (ADHD) as a dysregulation within the G–μ–S framework is detailed in ADHD as Dysregulation in the G–μ–S Framework (see Appendix C).

Novelty and Relation to Prior Work

This framework's distinctiveness lies in synthesizing dynamical systems theory with a computationally elaborated adaptive controller, the Self-System (S), which employs hierarchical Bayesian learning to regulate core parameters of subjective experience. This approach offers several novel contributions that extend and refine prior work in cognitive science, control theory, and computational psychiatry:

1. **The Mind Topography Map:** This paper introduces a comprehensive Mind Topography Map (Figure 3), derived from our model's potential function (Eq. (3)), globally visualizing stability regimes across the full G–μ plane. Unlike isolated DST analyses (e.g., (Huys et al., 2016; X.-J. Wang, 2002)), it charts key bifurcation boundaries (as detailed in the Bifurcation Boundaries Shaping the Landscape section), offering a unified view of diverse psychological dynamics. Crucially, this map provides a navigable landscape for the Self-System, an integrative novelty.

2. **Explicit Modeling of Gain-dependent Effective Bias ($G\mu$):** The explicit $G\mu$ term in our update rule (Eq. (1)) highlights how affective gain G actively scales cognitive bias μ, dynamically shaping its effective pull. This moves beyond treating them as separate modulators (cf. FEP (K. Friston, 2010) and PCT (Powers, 1973)). Demonstrated consequences include bias amplification (Figure 6) and varied convergence speeds (Figures 4, 5). Unlike purely mathematical analogues (cf. Imperfect Bifurcation Theory, Table 1), our model maps G and μ to psychological constructs regulated by an adaptive S capable of tuning them independently.

3. **A Tractable Computational Model of the Self-System (S) for Higher-Order Control:** The Self-System (S) is formalized as a higher-order controller (in the Computational Mechanism of the Self-System (S) section) using a tractable hierarchical Bayesian learning mechanism to adaptively modulate G and μ. It infers optimal (G, μ) settings (Eq. (5)) and learns meta-parameters $\boldsymbol{\theta}_{\mathrm{S}}$ via an

objective function $\mathcal{L}$ (Eq. (6); see Algorithm 1). This offers a concrete meta-control implementation for navigating the Mind Topography Map (Figure 9), more specific than often found in FEP or Meta-RL concerning this G–μ subjective landscape (Table 1). This explicit modeling of agentic regulation of core subjective parameters advances understanding of volitional mental state control.

4. **Characterization of IDE:** We characterize IDE at $(G = 1, \mu = 0)$ as a functionally unique state (in the IDE State $(G = 1, \mu = 0)$ section). Its dynamics $(M_s(t + 1) = -\alpha M_s(t)^3 + \epsilon(t))$ ensure strong stability $(f'(0) = 0)$ with direct responsiveness to noise $\epsilon(t)$, unlike unresponsive static inertia (Table 4). More than mere bias absence, IDE is an optimal, adaptive state of balanced responsiveness targeted by S. This offers a mathematically grounded operationalization (cf. "Middle Way" (Davis & Vago, 2013)) and a novel regulatory target distinct from isolated precision maximization or error minimization, defining a specific parameter set for optimal subjective functioning.

5. **A Formal Bridge to Perceptual Control Theory (PCT):** While synthesizing multiple traditions, this framework offers a particularly deep connection to PCT. It provides a specific mathematical engine for PCT's subjective-cognitive stratum, a domain previously described more conceptually. Specifically, our model (a) offers a concrete mechanism for "reorganization" via the interplay of noise and bifurcation (as detailed in Sec. 6.3), which refines the concept of random search; and (b) provides, through the SGBD extension (see Appendix A), a formal path to model core PCT concepts such as the orthogonality of control hierarchies and internal conflict. This clarifies both the current model's scope and its potential to address more complex phenomena.

Taken together, these individual contributions form a single, coherent architecture for modeling self-regulated subjectivity. A key strength of this architecture, as visualized in

Figure 2, is its potential to serve as a mathematical bridge, connecting the rich descriptive models of Mathematical Psychology with the normative, agent-based frameworks of PCT and FEP.

Limitations and Future Directions

- **Modeling Individual Differences:** The foundational scalar model abstracts from significant individual variability inherent in subjective experiences. This constitutes a key limitation.

  Future work should therefore prioritize the full development and rigorous theoretical validation of the SGBD framework (see Appendix A). This offers a robust pathway to explicitly model these person-specific differences, such as those related to personality or trait-like characteristics.

- **Self-System (S) Mechanism Refinement:** The Self-System's (S) computational architecture needs more detailed theoretical specification to be fully operational. Key aspects requiring elaboration include its objective function ($\mathcal{L}$), the nature and learning of meta-parameters ($\boldsymbol{\theta}_{\mathrm{S}}$), and specific inference methods.

  Consequently, subsequent research must focus on theoretically elaborating these components and exploring alternative formulations. Rigorous computational experiments are also needed to precisely define the Self-System's adaptive capabilities and its operational limits.

- **Advancing Analytical Scope with Stochastic Dynamics:** The model's current primary reliance on deterministic dynamics for fixed parameters restricts its capacity. It cannot fully capture the inherent variability and pervasive influence of noise characteristic of real-world subjective experiences.

  To enhance ecological validity, future efforts should therefore systematically incorporate stochastic analysis. This involves modeling the impact of noise ($\epsilon(t)$) on state trajectories and attractor stability, and crucially, exploring S-driven regulation and learning processes within noisy environments.

- **Empirical Grounding: Methodological Development and Comprehensive Validation** Empirically grounding the S-based model faces key challenges. Reliably observing subjective states ($M_{s,\text{obs}}$) from diverse data streams is difficult, and ensuring robust identifiability and estimation of core parameters (G, μ) is crucial. In parallel, the proposed comprehensive framework currently lacks the extensive quantitative validation needed to test its overall utility and predictive claims against real-world data.

  Therefore, future work must address these issues in concert. The priority is to develop and implement improved methods for empirical measurement and parameter estimation, for instance by leveraging multimodal datasets and theoretically-informed priors. Subsequently, these methods should be used to rigorously test the framework's core predictions and the adaptive functions of its key components, including the scalar model, the Self-System (S), and the SGBD extension.

- **Negative Affective Gain and the Action Loop** Our analysis of subjective dynamics on the G–μ plane critically depends on the premise that affective gain can be negative ($G < 0$), which is essential for modeling self-amplifying internal states (e.g., "fear begets fear"). This concept of an intrinsically negative gain within the subjective loop, however, is not a central feature of major frameworks like Perceptual Control Theory (PCT) or the Free Energy Principle (FEP). In PCT, gain is fundamentally positive to ensure negative feedback, and positive-feedback phenomena are typically framed as properties of the entire action loop, not the internal gain parameter itself. Likewise, "precision" in FEP is mathematically non-negative.

  A critical future direction is therefore to model the full perception-action loop (see Figure 1) to elucidate how subjective states generated by negative-gain dynamics translate into external actions. Exploring, in dialogue with PCT, how these internal states can trigger paradoxical behaviors (e.g., hyperventilation during panic) is a particularly important avenue for future research.

- **Synergizing Advanced Subjective Models with AI for Empowered Self-Regulation** A significant future ambition involves synergizing advanced subjective models, such as the SGBD framework, with AI to create personalized tools for adaptive self-regulation. However, this vision faces substantial hurdles. Methodologically, estimating the parameters of high-dimensional, person-specific models from the sparse and noisy data of daily life is a major technical challenge. Furthermore, for such systems to be trusted and adopted, their recommendations must be transparent, requiring significant advances in explainable AI (XAI). Critically, the use of highly sensitive personal data raises profound ethical, privacy, and fairness concerns that must be proactively addressed to prevent misuse and ensure equitable benefits.

  The ultimate goal is to overcome these challenges by developing “human-centric” AI systems that act as partners in self-discovery rather than as black-box authorities. Such systems would leverage rich, personalized models—continuously updated and validated against clinical and experimental data—to provide individuals with deep, actionable insights into their own cognitive and affective dynamics. Instead of simply prescribing behaviors, the aim is to foster empowered self-regulation, granting users the metacognitive tools to navigate their own mental landscapes towards greater human flourishing.

- **Implementing a Self-System in AI: Towards an Ethically Subjective Agent** Contemporary AI, particularly Large Language Models (LLMs), lacks genuine subjectivity or emotion in any meaningful sense; the foundation of their response generation lies in the vast statistical patterns within their training data. However, humans, due to evolutionary-acquired social-cognitive functions such as a propensity for anthropomorphism, unconsciously project intent and emotions onto these systems. This chasm between the non-emotional reality of AI and subjective human misperception harbors risks, including user misunderstanding, unilateral attachment, and broader ethical issues.

Future work could involve developing the framework proposed herein as a constitutive model applicable to AI itself. Specifically, this would entail implementing a Self-System (S) and an intrinsic objective function ($\mathcal{L}$) within the AI, enabling an architecture where it dynamically adjusts its own sensitivity (G) and bias (μ) through interactions with users and the environment. Crucially, to avoid the risk of the AI locking itself into a rigid adherence to a particular cognitive bias (μ), the Self-System (S) must be designed to guarantee the flexibility to escape from such maladaptive equilibria based on its objective function ($\mathcal{L}$). In such a model, the AI's "emotions" would not be mere output labels but rather expressions of a coherent internal state (M_s) that emerges from its goal-pursuit process. This research agenda aims not at mere behavioral mimicry, but at the creation of a new class of ethically subjective agents, capable of grounding their actions in a coherent, internally-generated sense of self.

Conclusion

This study proposes a dynamical model of subjectivity based on affective gain (G), cognitive bias (μ), and their regulation by a Self-System (S). The scalar formulation yields a principled Mind Topography Map and identifies the Ideal Dynamical Equilibrium (IDE; $G = 1, \mu = 0$) as a key balanced attractor. Central to this model, S is formalized as a hierarchical Bayesian learner that adaptively tunes G and μ, providing a tractable architecture for modeling self-regulated subjective dynamics.

The comprehensive G–μ–S framework, incorporating its SGBD extension (Modeling Individual Differences: The SGBD Framework) (see Appendix A), aligns with established normative theories. This robust theoretical grounding, in turn, offers a solid foundation for innovative approaches in computational psychiatry and the development of human-centric AI. Crucially, this work *invites broad empirical validation*—ranging from laboratory studies to digital phenotyping—to further refine its predictive utility and expand its applications.

Ultimately, this research suggests that subjectivity is best conceptualized as an adaptive control process, wherein the Self-System dynamically balances core affective and cognitive parameters to guide behavior toward coherent goals. Such a framing not only deepens our understanding of subjective experience but also clarifies critical pathways for enhancing self-understanding and promoting well-being.

References

Adams, R. A., Huys, Q. J., & Roiser, J. P. (2016). Computational psychiatry: Towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery & Psychiatry*, *87*(1), 53–63. https://doi.org/10.1136/jnnp-2015-310737

Davis, J. H., & Vago, D. R. (2013). Can enlightenment be traced to specific neural correlates, cognition, or behavior? No, and (a qualified) yes. *Frontiers in Psychology*, *4*, 870. https://doi.org/10.3389/fpsyg.2013.00870

Dayan, P. (2012). Twenty-five lessons from computational neuromodulation. *Neuron*, *76*(1), 240–256. https://doi.org/10.1016/j.neuron.2012.09.027

Deco, G., Rolls, E. T., Albantakis, L., & Romo, R. (2013). Brain mechanisms for perceptual and reward-related decision-making. *Progress in Neurobiology*, *103*, 194–213. https://doi.org/10.1016/j.pneurobio.2012.12.004

Desbordes, G., Gard, T., Hoge, E. A., Hölzel, B. K., Kerr, C., Lazar, S. W., Olendzki, A., & Vago, D. R. (2015). Moving beyond mindfulness: Defining equanimity as an outcome measure in meditation and contemplative research. *Mindfulness*, *6*(2), 356–372. https://doi.org/10.1007/s12671-013-0269-8

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., & Abbeel, P. (2016). RL^2: Fast reinforcement learning via slow reinforcement learning [Preprint available at https://arxiv.org/abs/1611.02779]. *arXiv preprint arXiv:1611.02779*.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *Proceedings of the 34th International Conference on Machine Learning*, 1126–1135.

Fontanesi, C., Palminteri, S., & Lebreton, M. (2019). A drift-diffusion model of the explore-exploit trade-off. *PLoS Computational Biology*, *15*(7), e1007208.

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. J. (2019). *A free energy principle for a particular physics* [Preprint monograph]. PsyArXiv. https://doi.org/10.31234/osf.io/x76ah

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: The brain as a phantastic organ. *The Lancet Psychiatry*, *1*(2), 148–158. https://doi.org/10.1016/S2215-0366(14)70275-5

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404–413. https://doi.org/10.1038/nn.4252

Kuznetsov, Y. A. (2013). *Elements of applied bifurcation theory* (Vol. 112). Springer. https://doi.org/10.1007/978-1-4757-3978-9

Limanowski, J., Adams, R. A., Kilner, J., & Parr, T. (2024). The many roles of precision in action. *Entropy*, *26*(9), 790. https://doi.org/10.3390/e26090790

Nunes, A., Singh, S., Allman, J., Becker, S., Ortiz, A., & Trappenberg, T. (2022). A critical evaluation of dynamical systems models of bipolar disorder. *Translational Psychiatry*, *12*(1), 399. https://doi.org/10.1038/s41398-022-02161-6

Powers, W. T. (1973). *Behavior: The control of perception*. Aldine.

Powers, W. T., Abbott, B., Carey, T. A., Goldstein, D. M., Mansell, W., & Marken, R. S. (2011). Perceptual control theory: A model for understanding the mechanisms and phenomena of control. *The Journal of Mind and Behavior*, *32*(3), 187–218.

Smith, R., Moutoussis, M., & Bilek, E. (2021). Simulating the computational mechanisms of cognitive and behavioral psychotherapeutic interventions by fitting generative models of behavior to behavioral data. *Scientific Reports*, *11*, 10128. https://doi.org/10.1038/s41598-021-89047-0

Strogatz, S. H. (2015). *Nonlinear dynamics and chaos* (2nd). Westview Press.

Strogatz, S. H. (2018). *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*. CRC Press. https://doi.org/10.1201/9780429492549

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press.

Thom, R. (1989). *Structural stability and morphogenesis*. CRC Press. https://doi.org/10.1201/9780429493027

van Vugt, M. K., Moye, A., & Sivakumar, S. (2019). Computational modelling approaches to meditation research: Why should we care? *Current Opinion in Psychology*, *28*, 49–53. https://doi.org/10.1016/j.copsyc.2018.10.011

Wang, J., Guo, H., Qiang, W., Li, J., Zheng, C., & Xiong, H. (2024). Neuromodulated meta-learning [Preprint available at https://arxiv.org/abs/2411.06746]. *arXiv preprint arXiv:2411.06746*.

Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, *36*(5), 955–968. https://doi.org/10.1016/S0896-6273(02)01094-9

Witkiewitz, K., & Marlatt, G. A. (2007). Modeling the complexity of post-treatment drinking: It's a rocky road to relapse. *Clinical Psychology Review*, *27*(6), 724–738. https://doi.org/10.1016/j.cpr.2007.01.002

Table 1

Comparison of selected prior approaches contrasted with the present framework.

Approach / Field	**Gain/Bias Handling**	**Dynamics / Stability Analysis**	**Identified Limitations / Addressed Gaps**	**Our Contribution / Difference**
Perceptual Control Theory (PCT)[a]	Loop gain (G analogue), reference level (μ analogue).	Control-loop dynamics, error-reduction focus.	Reorganization mechanism underspecified. Limited modeling of hierarchy and conflict.	Models reorganization via noise-driven bifurcation. SGBD enables formal modeling of conflict and orthogonality.
Reinforcement Learning (RL)[b]	Learning rate / exploration temperature (G analogues); value/policy priors (μ).	Convergence, optimal-control; Meta-RL tunes G online.	Meta-control emerging; explicit G–μ stability maps not central.[d] Higher-level adaptation often implicit.	$G\mu$ term[c]and global map (C1,C2); S (C3) with explicit learning rules for G and μ.
Predictive Coding / Active Inference (FEP)[e]	Precision (G) versus priors (μ); precision ≈ arousal-modulated gain.	Free-energy minimization; message-passing stability.	Specific G–μ landscape (Mind Topography Map) and IDE operation needs detail.[f] Meta-learning links often conceptual.	Specific G–μ model for subjectivity landscapes; defines IDE (C1–C4); concrete S mechanism.
DST Mood Models[g]	Stress-like parameters as G, baseline shift as μ.	Bifurcation analyses (bistability).	Lacks full G–μ map and explicit, adaptive higher-order control.	Global landscape; S with learning.
DST Decision Models[h]	Attractor parameters (indirect G/μ).	Attractor bifurcation to choice.	G and μ roles not explicit; minimal higher-order adaptive control.[i]	Makes G and μ explicit; S with adaptive capabilities.
Imperfect Bifurcation Theory[j]	Control $r \sim -G$, imperfection $h \sim G\mu$.[c]	Normal-form local analysis.	No cognitive mapping or higher-order adaptive control.	Maps G and μ to parameters; adaptive control by S.

Note. DST = Dynamical Systems Theory; IDE = Ideal Dynamical Equilibrium ($G = 1, \mu = 0$). [a] (Powers, 1973; Powers et al., 2011) [b] (Dayan, 2012; Duan et al., 2016; Finn et al., 2017; Sutton & Barto, 1998; J. Wang et al., 2024)
[c] The $G\mu$ term, distinct, is analogous to a combined parameter (e.g., ν); Our model separates G and μ for analysis and regulation by S. [d] e.g., RL^2 (Duan et al., 2016), MAML (Finn et al., 2017), Neuromodulated Meta-Learning (J. Wang et al., 2024) adapt learning rate, temperature, or network structure, and some Meta-RL approaches also co-adapt policy/value priors (μ-analogues). [e] A comprehensive framework that can encompass concepts from RL and DST; (K. Friston, 2010; K. J. Friston, 2019; Limanowski et al., 2024) [f] See Limanowski et al. (2024) for dynamic gain adjustment in FEP. [g] (Huys et al., 2016; Nunes et al., 2022) [h] (Deco et al., 2013; X.-J. Wang, 2002)
[i] For instance, some recent evidence-accumulation models explicitly tune parameters like drift-rate, which can be seen as a gain analogue (Fontanesi et al., 2019). [j] (Kuznetsov, 2013; Strogatz, 2018)

Table 2

Conceptual mapping of key model components to FEP and PC.

Core Concept / Variable (Our Model)	**Proposed Model Term**	**Analogue/Relation in FEP**	**Analogue/Relation in PC**	**Note**
Subjective State	M_s	cognitive states (e.g., $\boldsymbol{\mu}$ in general coordinates)	Representation/ Prediction units	Internal belief state
Affective Gain	G	Precision of sensory prediction error (e.g., Π_ε)	Gain of error units; Modulatory input	Modulates error; see Note for its unique property.[(b)]
Cognitive Bias	μ	Priors / Empirical priors (e.g., $p(\vartheta)$); Attractors	Top-down predictions / Biases	Pre-existing tendencies
Gain–Bias Interaction	$G\mu$	Effective force from priors, scaled by precision	Modulated predictions[(a)]	Effective bias pull
Prediction Error (implicit)	$M_o - M_s$	Prediction error (e.g., ε)	Error units output	Drives update
Self-System Objective	Optimize $\mathcal{L}$	Minimize VFE / Max. Model Evidence	Min. sum sq. weighted errors	Higher-order adaptive goal
S-System Regulation of G and μ	Via $\boldsymbol{\theta}_{\mathrm{S}}$ learning	Learning (hyper)params (e.g., prior precisions)	Learning synaptic strengths/biases	A key innovation of this framework; see Note.[(c)]

Note. FEP = Free Energy Principle; PC = Predictive Coding. [(a)] E.g., attentional mechanisms can modulate predictions or prediction errors via gain control. [(b)] Modulates error. Critically, G can be negative (unlike FEP/PC precision), enabling modeling of both error-correcting (adaptive, $G > 0$) and error-amplifying (escapist, $G < 0$) dynamics. [(c)] Meta-learning. An explicit meta-control mechanism for adaptively co-regulating both G and μ to navigate the stability landscape, a key innovation of this framework.

Table 3

Comparison of the scalar model (main text) and the SGBD framework (Appendix A), highlighting their core features and scope.

Aspect	**Scalar Model (Main Text)**	**SGBD (Appendix A)**
Representation	Scalar quantities ($M_s, G, \mu, \alpha, \epsilon$)	Vector quantities ($\boldsymbol{M_s}, \boldsymbol{G}_{\text{eff}}, \boldsymbol{\mu}_{\text{eff}}, \boldsymbol{\alpha}, \boldsymbol{\epsilon}$)
Primary Goal	Elucidate fundamental, generalizable principles of G–μ dynamics (Mind Topography Map)	Provide a theoretical basis for modeling complex phenomena including individual differences and mixed states
Handling of Individual Differences	Archetypal/representative values; individual variations are outside the direct scope	Explicitly models individual-specific characteristics via structural matrices (e.g., $\mathbf{M}_G, \mathbf{M}_\mu$)
Focus in This Paper	Detailed analysis and primary source of conclusions	Exploratory mathematical sketch; indicates future research directions
Status	Foundational, validated within its scope	Exploratory, requires further development and validation

Table 4

Comparison of IDE ($G = 1, \mu = 0$) and Inert State ($G = 0, \mu = 0$).

Feature	IDE ($G = 1, \mu = 0$)	Inert State ($G = 0, \mu = 0$)
Update Rule	$M_s(t+1) = -\alpha M_s(t)^3 + \epsilon(t)$	$M_s(t+1) = M_s(t) - \alpha M_s(t)^3$
Linear $M_s(t)$ term	Absent	Present
Stability at $M_s^* = 0$	Stable ($f'(0) = 0$)	Neutrally Stable ($f'(0) = 1$)
Response to $\epsilon(t)$	Direct	None
Responsiveness	Dynamic, adaptive	Static inertia
Interpretation	Balanced equilibrium (target for S)	Disconnection, unresponsiveness

Table 5

A concise four-quadrant typology for the G–μ plane.

Quadrant	**Cognitive-Affective Label**	**Behavioural Signature**
$G > 0, \mu > 0$	Adaptive Optimism	Proactive exploration; reality-grounded enthusiasm.
$G < 0, \mu > 0$	Escapist Optimism	Detached daydreaming; goal inflation; disengagement from reality.
$G < 0, \mu < 0$	Escapist Pessimism	Defensive withdrawal; rumination; learned helplessness.
$G > 0, \mu < 0$	Adaptive Vigilance	Cautious realism; loss-preventive planning; steady improvement.

Table 6

Summary of parameter dynamics and psychological states through the stages of mental transformation.

Stage	**Gain** (G)	**Bias** (μ)	**Subjective State** (M_s)	**Self-System** (S; $\boldsymbol{\theta}_{\text{S}}, \mathcal{L}$)
1. Anxiety & Fixation	≈ 1 (Hypersensitive)	< 0 (Negative)	Aligned with negative bias; fixated state ($M_s < 0$).	Immature; $\mathcal{L}$ unmet, $\boldsymbol{\theta}_{\text{S}}$ learning stalled.
2. Rupture & Dissociation	$+ \to 0 \to -$ (Collapses)	< 0	Dissociated from bias; shifts to positive ($M_s > 0$).	Loses control; $\mathcal{L}$ deviation increases, $\boldsymbol{\theta}_{\text{S}}$ ineffective.
3. Fortress of Delusion	< 0 (Negative peak)	≈ 0	Fixed at high positive value; manic defense state.	Seeks local optimum; $\mathcal{L}$ abandoned, $\boldsymbol{\theta}_{\text{S}}$ becomes maladaptive.
4. First Light	< 0	$0 \to +$ (Turns positive)	Remains high, but foundation for change is laid.	Resumes learning; $\mathcal{L}$ re-evaluated, $\boldsymbol{\theta}_{\text{S}}$ recalibrated.
5. Awakening & Re-integration	$- \to +$ (Turns positive)	> 0 (Positive peak)	Re-aligns with positive bias; internal conflict resolved.	Regains control; $\mathcal{L}$ (coherence) met, $\boldsymbol{\theta}_{\text{S}}$ optimized.
6. Serenity & Enlightenment	$\to 1$ (Optimal)	$\to 0$ (Neutral)	Converges to zero; attains IDE (Enlightenment).	Learning complete; $\mathcal{L}$ satisfied, $\boldsymbol{\theta}_{\text{S}}$ converged.

Table 7

Interpretation of ADHD core symptoms within the G–μ–S parameter space.

Symptom	**Affective Gain** (G)	**Cognitive Bias** (μ)	**Dynamic Interpretation via Self-System** (S)
Inattention	Low G weakens error sensitivity and signal updating	μ fluctuates or fails to consolidate	S fails to retune; even in stable zones, focus decays or drifts
Hyperactivity / Impulsivity	High G overreacts to noise and transient bias	μ collapses toward impulsive attractors	S can't suppress divergence; system jumps between states
Emotional Dysregulation	High G combined with negative μ	Deep wells or bistability arise	S crosses bifurcation threshold; mood oscillations or state flipping

Note. This table summarizes how core ADHD symptoms may be dynamically interpreted within the G–μ–S framework as dysregulations of affective gain (G), cognitive bias (μ), and higher-order control by the Self-System (S).

Table 8

Mapping of therapeutic interventions to the G–μ–S model parameters.

Intervention	**Model Target**	**Theoretical Action**	**Expected Parameter Shift**
Stimulants (e.g., MPH)	G, noise ϵ	Elevate signal-to-noise ratio (SNR); stabilize G near 1	Converge toward IDE from hypo- or hyperactive gain
CBT / Parent Training	S ($\boldsymbol{\theta}_{\mathrm{S}}$, $\mathcal{L}$)	Support meta-learning; reframe bias-related priors	$\eta_\theta \uparrow$, $\mu \rightarrow$ adaptive priors
Mindfulness / Meta-cognition	G, S	Enhance awareness of internal fluctuations; improve regulation	Lower G variability, increase IDE dwell time
Neurofeedback	μ, G	Visualize bias; support conscious recalibration	$\mu \rightarrow 0$, stabilize G via feedback

Note. This table shows how major therapeutic interventions for ADHD may be interpreted as influencing specific parameters or regulatory mechanisms in the G–μ–S model. The framework highlights how pharmacological and psychosocial strategies may converge in tuning the same self-regulatory system.

Table 9

Mapping Smith's philosophy to the mathematical model.

The Theory of Moral Sentiments	**The Mathematical Model**
Impartial Spectator	Self-System (S)
Virtue / Propriety	Ideal Dynamical Equilibrium (IDE)
Self-Command	Dynamical Adjustment of G and μ
Pursuit of Virtue and Happiness	Optimization of Objective Function ($\mathcal{L}$)
General Rules of Morality / Learned Moral Character	Meta-parameter ($\boldsymbol{\theta}_{\mathrm{S}}$) that governs the Self-System

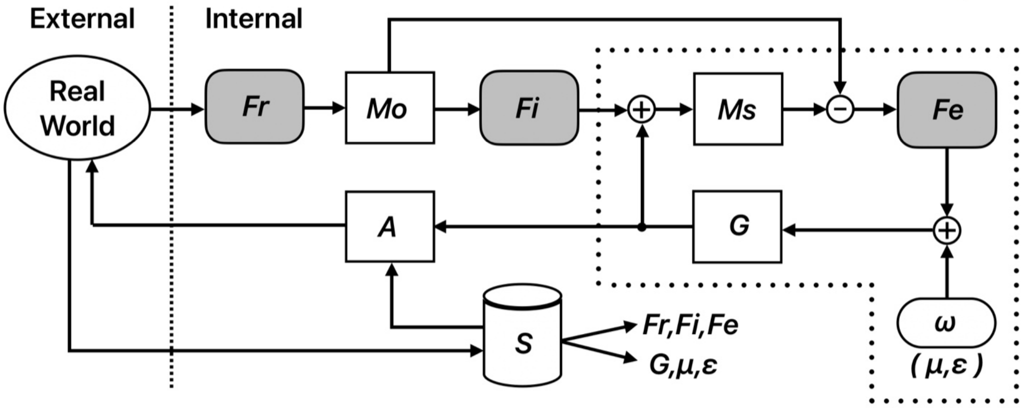

Figure 1

Structure of the Internal control model. Real-World inputs are processed internally. The dotted area highlights the components mathematically formulated in the Mathematical Formulation: The Update Rule for M_s section.

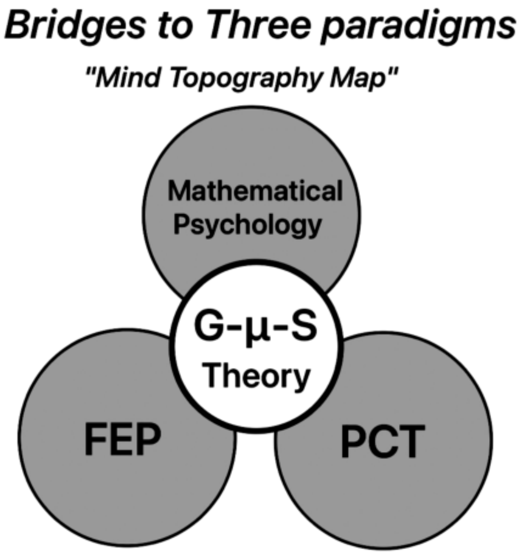

Figure 2

A conceptual diagram illustrating the role of the G–μ–S framework as a unifying mathematical formalism that bridges core concepts from Mathematical Psychology, the Free Energy Principle (FEP), and Perceptual Control Theory (PCT). The Mind Topography Map serves as a key analytical tool within this integrative structure.

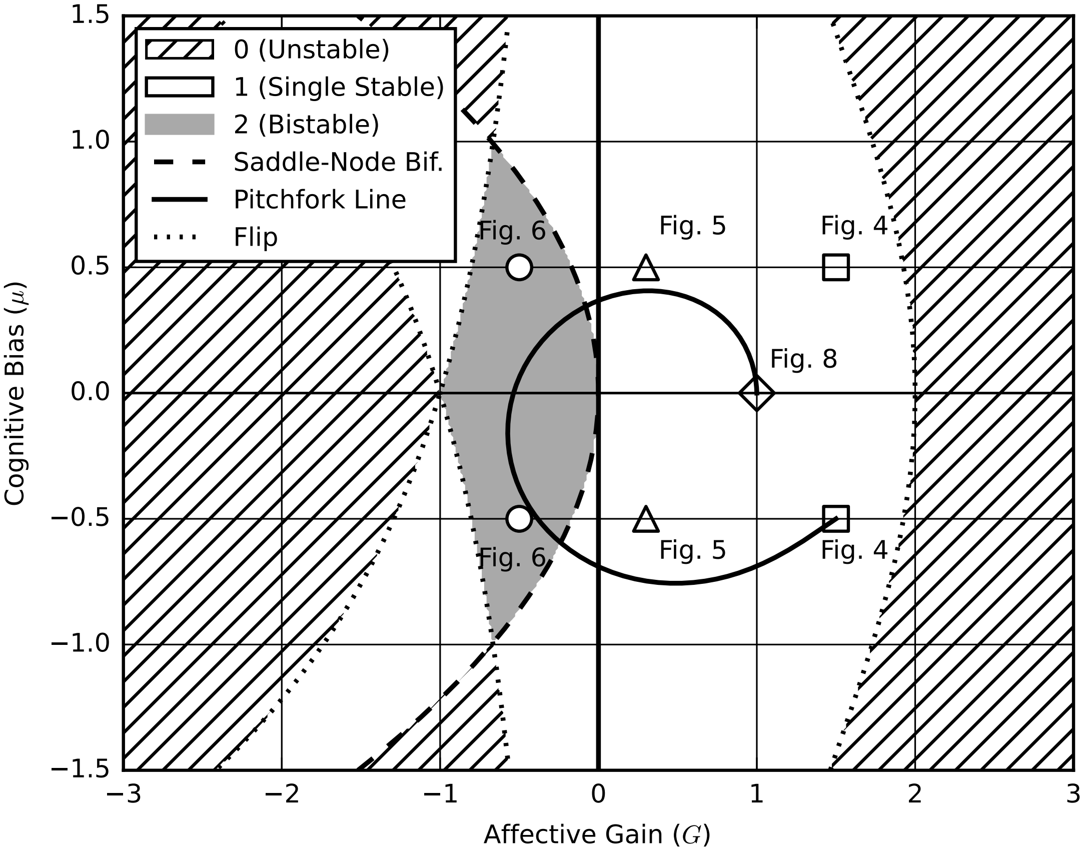

Figure 3

The Mind Topography Map, illustrating stability regions in the G–μ parameter plane ($\alpha = 0.1$)*. The plane is divided by key bifurcation boundaries: Saddle-Node (dashed), Pitchfork (solid vertical), and Flip (dotted), as defined in the Bifurcation Boundaries Shaping the Landscape section and Eq.* (11)*. Colors denote distinct regimes based on the number of stable fixed points: white (1), dark gray (2), and hatched (0), detailed in the Stability Regimes: Definitions and Characteristics section. The range (*$G \in [-3, 3]$*,* $\mu \in [-1.5, 1.5]$*) is chosen to visualize all relevant bifurcations clearly; this limitation has no impact on the generality of results. Markers indicate simulation parameters (Figures 4–8); the spiral line represents a conceptual Self-System trajectory (Figure 9).*

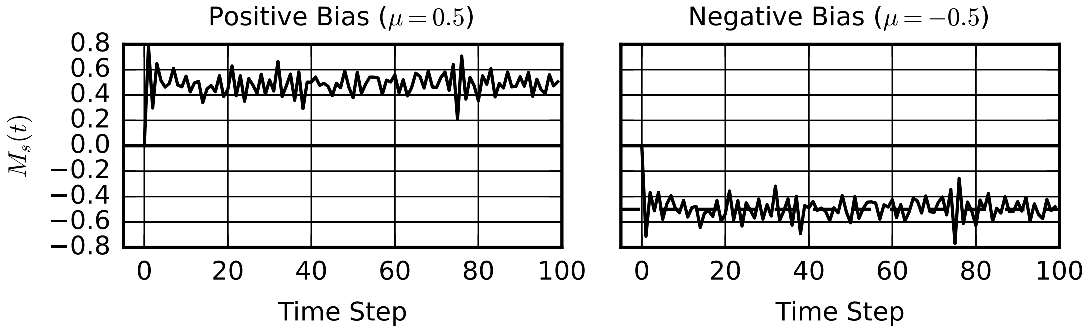

Figure 4

Simulation for $G = 1.5, \mu = \pm 0.5$ (single stable, positive G). This figure illustrates rapid convergence to and strong fixation on the bias value μ.

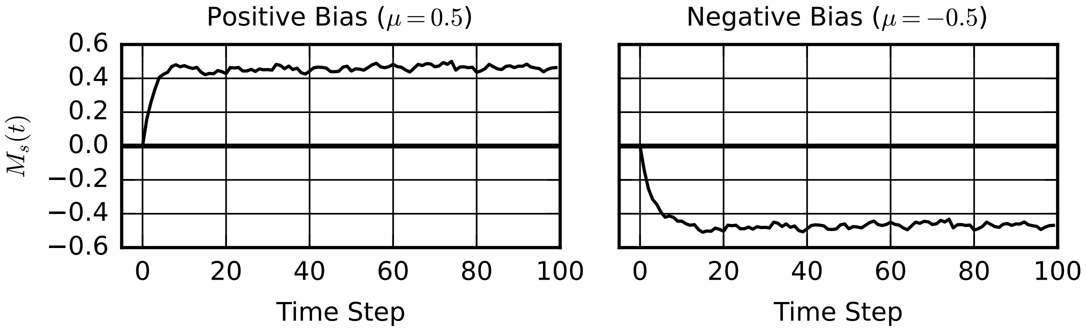

Figure 5

Simulation for $G = 0.3, \mu = \pm 0.5$ (single stable, positive G). This figure illustrates slower, damped convergence to the bias value μ.

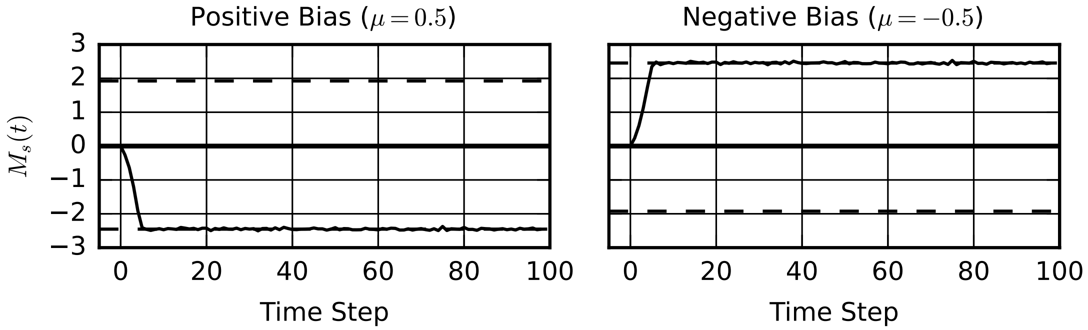

Figure 6

Simulation for $G = -0.5, \mu = \pm 0.5$ (bistable, negative G). This figure illustrates bias amplification and asymmetric stabilization. The initial state determines which attractor is reached.

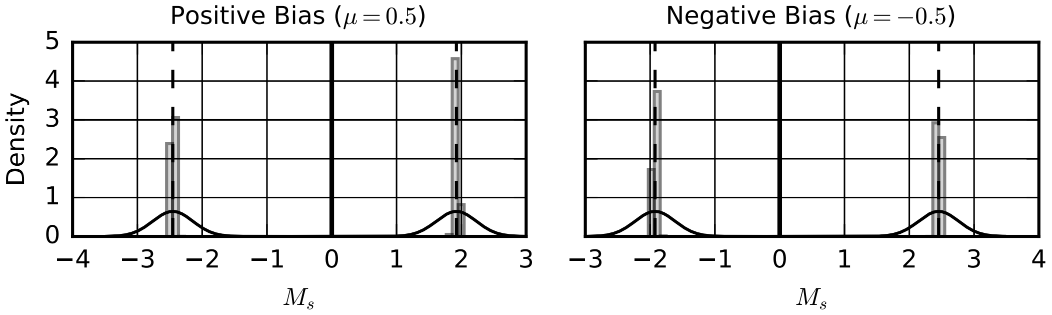

Figure 7

Distributions of M_s from long simulations ($T_{long} = 10,000$ steps) for $G = -0.5, \mu = \pm 0.5$. The bimodal distributions confirm the bistable nature of the dynamics for these fixed parameters.

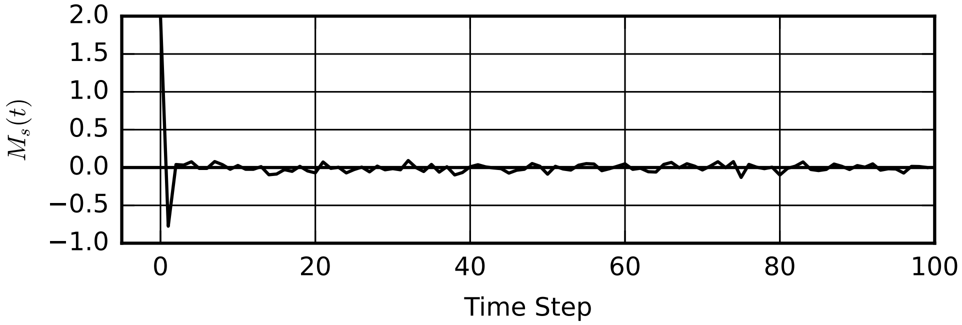

Figure 8

Simulation for IDE state: $G = 1, \mu = 0$. M_s decays to zero while remaining responsive to noise $\epsilon(t)$. This configuration constitutes a primary regulatory target for the Self-System (S).

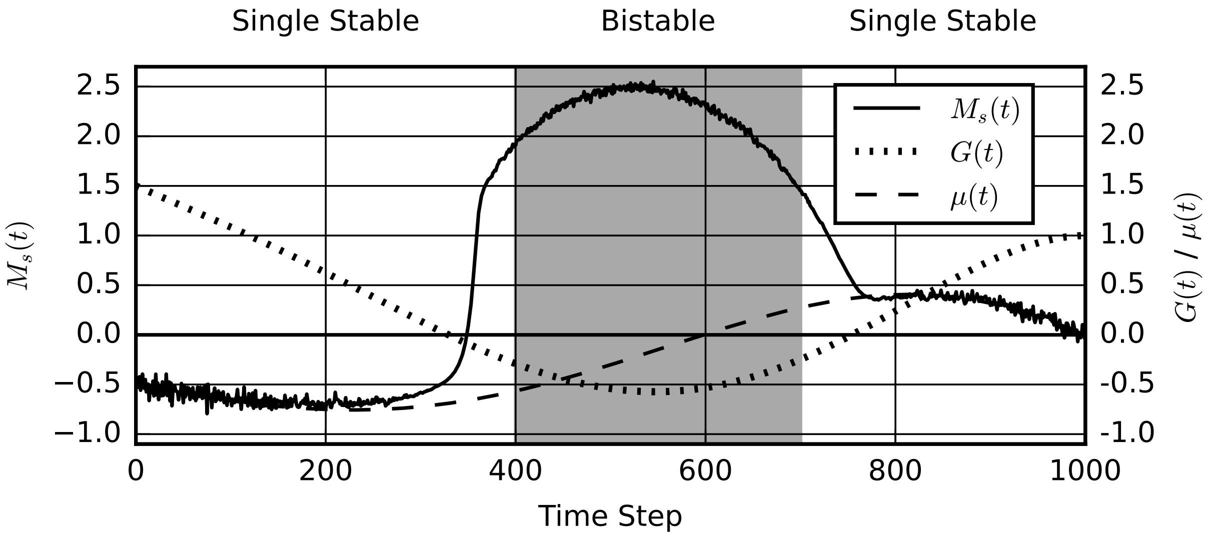

Figure 9

Conceptual simulation of dynamical self-regulation by the Self-System (S). $M_s(t)$ responds to S steering $G(t)$ and $\mu(t)$ (trajectory corresponds to the spiral solid line in the Mind Topography Map, Figure 3). This process is guided by the Self-System's hierarchical Bayesian learning mechanism (detailed in the Computational Mechanism of the Self-System (S) section).

Appendix A

Modeling Individual Differences: The SGBD Framework

Introduction to the SGBD Framework: A Vectorial Approach to Subjectivity

Motivation: Limitations of the Scalar Model and the Need for Modeling Individual Differences

The core scalar model, while elucidating fundamental dynamics of subjectivity (see the Methods: Internal Control Model section), inherently limits the representation of rich individual variability and complex, mixed subjective states. For instance, phenomena such as coexisting emotions (e.g., joy and sadness concurrently) or domain-specific sensitivities and biases (e.g., optimism in interpersonal relationships but pessimism in work-related matters) are challenging to capture with a single scalar quantity for subjective state, gain, or bias. To address these more nuanced aspects of subjectivity, this appendix introduces the Structural Gain–Bias Dynamics (SGBD) framework, an exploratory multidimensional extension designed to provide a richer, more granular modeling capability.

Contrasting SGBD with the Scalar Model: Leveraging Table 3 for Clarity

As highlighted in Table 3 of the main text (see the Modeling Approach: Scalar Representation for General Principles and a Vectorial Extension for Individual Differences section), the SGBD framework fundamentally differs from the foundational scalar model in its objectives and representational capacity. The scalar model primarily aims to elucidate fundamental, generalizable principles of gain–bias dynamics using archetypal or representative scalar values for M_s, G, μ, α, and noise ϵ. In contrast, SGBD provides a theoretical basis for modeling complex phenomena, including individual differences and mixed states. This is achieved by employing vector quantities for subjective states and related parameters ($\boldsymbol{M_s}, \boldsymbol{G}_{\text{eff}}, \boldsymbol{\mu}_{\text{eff}}, \boldsymbol{\alpha}$, and noise $\boldsymbol{\epsilon}$) and by introducing structural matrices (e.g., $\mathbf{M}_G, \mathbf{M}_\mu$) to explicitly model individual-specific characteristics. Consequently, while the conclusions in the main body of this paper derive exclusively from the foundational scalar model, SGBD is presented here as an exploratory mathematical sketch indicating future research directions for personalized and multidimensional modeling of subjectivity.

Scope of this Appendix: Focusing on Vectorized Subjective Dynamics under Simplified Assumptions ($\boldsymbol{M_o} = \mathbf{0}, F_i(\boldsymbol{M_o}) = \mathbf{0}$)

This appendix details the SGBD framework by extending the core concepts of the scalar model—subjective state, affective gain, and cognitive bias—into a multidimensional vector space. To maintain focus on the internal dynamics of subjectivity and to align with the simplifications made in the main text for analyzing these core dynamics (see the Mathematical Formulation: The Update Rule for M_s section), we adopt two key simplifying assumptions:

1. The objective model of the external world is considered to be zero: $\boldsymbol{M_o}(t) = \mathbf{0}$. This implies that the system is analyzed in the absence of, or abstracting from, specific external inputs that would define an objective state of affairs.

2. Consequently, the Interpretation Filter (F_i), which processes $\boldsymbol{M_o}(t)$, is also assumed to yield a zero output: $F_i(\boldsymbol{M_o}(t)) = F_i(\mathbf{0}) = \mathbf{0}$. If F_i were modeled as an affine transformation (e.g., $\mathbf{W}_{Fi}\boldsymbol{M_o}(t) + \mathbf{c}_{Fi}$), this assumption implies that its constant offset term $\mathbf{c}_{Fi}$ is also zero for the purpose of this simplified analysis.

These simplifications mean that the Recognition Filter (F_r) and Interpretation Filter (F_i) are not mathematically elaborated in this appendix. The primary focus remains on the dynamics evolving around the multidimensional subjective state ($\boldsymbol{M_s}$) itself, corresponding to the components within the dotted area of the overall internal control model depicted in Figure 1. This approach allows for a clear exposition of the SGBD's core architectural extension for internal subjective processing, deferring the complexities of interaction with a non-zero, dynamically changing external world model to future work.

Vector Representation of Core Subjective Components in SGBD

Subjective State (M_s)

To capture the multifaceted nature of subjective experience (e.g., different affective dimensions, cognitive appraisals, or self-related states), the scalar subjective state M_s is extended to an N-dimensional vector $\boldsymbol{M_s}(t) = [M_{s,1}(t), \ldots, M_{s,N}(t)]^T \in \mathbb{R}^N$. Each element $M_{s,i}(t)$ represents the intensity or level of the i-th aspect of subjectivity at time t.

Modeling Internal Conflict via Vector Partitioning

A key capability of the SGBD framework is its capacity to model internal conflict, a concept central to many psychological theories. This is achieved by partitioning the subjective state vector $\boldsymbol{M_s}$ into sub-vectors, each representing a distinct goal, value, or self-concept. For instance, $\boldsymbol{M_s}$ can be decomposed into two competing sub-vectors, $\boldsymbol{M_{s,A}}$ and $\boldsymbol{M_{s,B}}$, representing a motivational "tug-of-war" (e.g., approach vs. avoidance). The dynamics of these sub-vectors can be governed by conflicting gain and bias parameters within the structural matrices ($\mathbf{M}_G$, $\mathbf{M}_\mu$), leading to oscillatory or unstable behavior in the overall system. This provides a formal basis for analyzing how the resolution of, or failure to resolve, such internal conflicts shapes an individual's subjective experience over time.

Effective Affective Gain (G_{eff}) and Effective Cognitive Bias ($\boldsymbol{\mu}_{\mathit{eff}}$)

Analogous to their scalar counterparts, affective gain and cognitive bias are also represented as N-dimensional vectors to account for dimension-specific influences:

- **Effective Affective Gain ($\boldsymbol{G}_{\mathbf{eff}}(t)$):** This vector, $\boldsymbol{G}_{\text{eff}}(t) = [G_{\text{eff},1}(t), \ldots, G_{\text{eff},N}(t)]^T \in \mathbb{R}^N$, reflects varying sensitivities or reactivities across the different dimensions of the subjective state. $G_{\text{eff},i}(t)$ modulates the impact of errors or fluctuations on the i-th subjective dimension.

- **Effective Cognitive Bias ($\boldsymbol{\mu}_{\mathbf{eff}}(t)$):** This vector, $\boldsymbol{\mu}_{\text{eff}}(t) = [\mu_{\text{eff},1}(t), \ldots, \mu_{\text{eff},N}(t)]^T \in \mathbb{R}^N$, accounts for multifaceted internal predispositions, attractors, or repellers that act on each dimension of subjectivity.

$\mu_{\text{eff},i}(t)$ represents the default tendency or attractive force for the i-th subjective dimension.

Vectorized Saturation Coefficient ($\boldsymbol{\alpha}$) and Noise ($\boldsymbol{\epsilon}$)

To complete the multidimensional extension of the core dynamical equation (Eq. (1)), the saturation coefficient and noise term are also vectorized:

- **Saturation Coefficient Vector ($\boldsymbol{\alpha}$):** Defined as $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N]^T \in \mathbb{R}^N$, where each $\alpha_i > 0$. This allows for dimension-specific nonlinear saturation characteristics, ensuring that each subjective state component remains bounded.
- **Noise Vector ($\boldsymbol{\epsilon}(t)$):** Represented as $\boldsymbol{\epsilon}(t) = [\epsilon_1(t), \ldots, \epsilon_N(t)]^T \in \mathbb{R}^N$. Each component $\epsilon_i(t)$ models transient random fluctuations affecting the i-th dimension of subjectivity, typically drawn from a Gaussian distribution $N(0, \sigma_{\epsilon,i}^2)$.

Structural Generation of Effective Gain and Bias: Modeling Internal Architecture

Source Input Vectors for Gain and Bias ($\boldsymbol{g_{input}}, \boldsymbol{b_{input}}$)

We posit the existence of fundamental "sources" of sensitivity and bias, modeled as input vectors whose activation levels or strengths can vary over time, potentially influenced by context or the Self-System (S):

- **Source of Gain Input Vector ($\boldsymbol{g}_{\textbf{input}}(t)$):** An L-dimensional column vector $\boldsymbol{g}_{\text{input}}(t) = [g_{\text{input},1}(t), \ldots, g_{\text{input},L}(t)]^T \in \mathbb{R}^L$. Each element $g_{\text{input},l}(t)$ represents the current activation level of the l-th fundamental source of gain (e.g., general arousal, approach/avoidance system activation).
- **Source of Bias Input Vector ($\boldsymbol{b}_{\textbf{input}}(t)$):** A K-dimensional column vector $\boldsymbol{b}_{\text{input}}(t) = [b_{\text{input},1}(t), \ldots, b_{\text{input},K}(t)]^T \in \mathbb{R}^K$. Each element $b_{\text{input},k}(t)$ represents the current strength of the k-th fundamental source of bias (e.g., core self-beliefs, trait optimism/pessimism).

These source vectors serve as common underlying inputs that are then differentially processed by each individual's unique internal architecture.

Transformation by Individual Structural Matrices ($\mathbf{M}_G, \mathbf{M}_\mu$*)*

The SGBD framework proposes that an individual's specific effective gain vector $\boldsymbol{G}_{\text{eff}}(t)$ and effective bias vector $\boldsymbol{\mu}_{\text{eff}}(t)$ are generated by linear transformations of these source input vectors. These transformations are embodied by structural matrices unique to the individual:

$$\boldsymbol{G}_{\text{eff}}(t) = \mathbf{M}_G \boldsymbol{g}_{\text{input}}(t) \tag{A1}$$

$$\boldsymbol{\mu}_{\text{eff}}(t) = \mathbf{M}_\mu \boldsymbol{b}_{\text{input}}(t) \tag{A2}$$

Here, $\mathbf{M}_G \in \mathbb{R}^{N \times L}$ is the **gain structural matrix**, where element $M_{G,il}$ represents the coupling strength by which the l-th source of gain input $g_{\text{input},l}$ contributes to the i-th component of the effective gain $G_{\text{eff},i}$. Similarly, $\mathbf{M}_\mu \in \mathbb{R}^{N \times K}$ is the **bias structural matrix**, where element $M_{\mu,ik}$ signifies the weight by which the k-th source of bias input $b_{\text{input},k}$ influences the i-th component of the effective bias $\mu_{\text{eff},i}$.

Significance of Structural Matrices: Encoding Individuality and Potential for Self-System Modulated Plasticity

The structural matrices, $\mathbf{M}_G$ and $\mathbf{M}_\mu$, are posited as the primary loci of individual differences within the SGBD framework. They mathematically represent an individual's unique, relatively stable cognitive-affective "wiring" or "internal architecture" that governs their characteristic patterns of affective processing and cognitive tendencies. Specific patterns or values within these matrices can thus reflect enduring individual characteristics such as personality traits, learned information processing styles, developmental history, or even predispositions towards certain psychopathological conditions (e.g., through aberrant couplings to specific bias sources). While these matrices represent relatively stable aspects of an individual, they are not necessarily immutable. It is hypothesized that they possess a degree of plasticity and could be modified over longer timescales, for instance, through prolonged learning experiences, therapeutic interventions, or explicit efforts by the Self-System (S) to reshape one's own cognitive-affective architecture. The use of linear transformations (matrix multiplication) for this structural generation process is a deliberate choice to enhance model interpretability and mathematical tractability at this stage of theory development. A key implication of this architecture is its ability to formally instantiate concepts from other control theories. For instance, by constructing the columns of the structural matrices $\mathbf{M}_G$ and $\mathbf{M}_\mu$ to be mathematically orthogonal, this framework can model the independent “intrinsic” and “perceptual” control axes proposed in Perceptual Control Theory (PCT), providing a concrete bridge between these theoretical approaches.

Multidimensional Dynamics of the Subjective State ($\boldsymbol{M_s}$) within SGBD

The evolution of the multidimensional subjective state $\boldsymbol{M_s}(t)$ is governed by an update rule that extends the principles of the scalar model (Eq. (1)) to the vector space, incorporating the structurally generated effective parameters and the vectorized Evaluation Filter.

The Evaluation Filter (Fe) as an Affine Transformation

The Evaluation Filter (Fe) is responsible for processing the discrepancy or error signal that drives the update of the subjective state. In the SGBD framework, under the simplifying assumptions of $\boldsymbol{M_o}(t) = \mathbf{0}$ and $F_i(\mathbf{0}) = \mathbf{0}$ (as detailed in the Scope of this Appendix: Focusing on Vectorized Subjective Dynamics under Simplified Assumptions ($\boldsymbol{M_o} = \mathbf{0}, F_i(\boldsymbol{M_o}) = \mathbf{0}$) section), the primary error input to Fe is derived from the current subjective state itself, specifically $\boldsymbol{E}(t) = \boldsymbol{M_o}(t) - \boldsymbol{M_s}(t) = \mathbf{0} - \boldsymbol{M_s}(t) = -\boldsymbol{M_s}(t)$. The SGBD framework models Fe as performing an affine transformation on this error vector:

$$\boldsymbol{E}_{\text{eval}}(t) = \mathbf{W}_{\text{Fe}}(S,t)\boldsymbol{E}(t) + \boldsymbol{c}_{\text{Fe}}(S,t) = -\mathbf{W}_{\text{Fe}}(S,t)\boldsymbol{M_s}(t) + \boldsymbol{c}_{\text{Fe}}(S,t) \tag{A3}$$

Here, $\mathbf{W}_{\text{Fe}}(S,t) \in \mathbb{R}^{N\times N}$ is the **evaluation weight matrix**, and $\boldsymbol{c}_{\text{Fe}}(S,t) \in \mathbb{R}^N$ is the **evaluation offset vector**. Both $\mathbf{W}_{\text{Fe}}$ and $\boldsymbol{c}_{\text{Fe}}$ can be modulated by the Self-System (S), allowing for dynamical adjustments in how errors are weighted and prioritized across different subjective dimensions, and for the introduction of a context-dependent evaluative offset.

General Vectorized Update Rule for $\boldsymbol{M_s}$

The update rule for the multidimensional subjective state $\boldsymbol{M_s}(t)$ generalizes the principles of the scalar model's update rule (Eq. (1)). It incorporates the vectorized components: the current subjective state $\boldsymbol{M_s}(t)$, the dimension-specific saturation term $-\boldsymbol{\alpha} \odot (\boldsymbol{M_s}(t))^{\circ 3}$, and an update signal. This update signal is formed by the evaluated error $\boldsymbol{E}_{\text{eval}}(t)$ (from Eq. (A3)), the effective cognitive bias $\boldsymbol{\mu}_{\text{eff}}(t)$ (from Eq. (A2)), and the noise vector $\boldsymbol{\epsilon}(t)$, all scaled element-wise by the effective affective gain $\boldsymbol{G}_{\text{eff}}(t)$ (from Eq. (A1)). The resulting general vectorized update rule is:

$$\boldsymbol{M_s}(t+1) = \boldsymbol{M_s}(t) - \boldsymbol{\alpha} \odot (\boldsymbol{M_s}(t))^{\circ 3} + \boldsymbol{G}_{\text{eff}}(t) \odot (\boldsymbol{E}_{\text{eval}}(t) + \boldsymbol{\mu}_{\text{eff}}(t) + \boldsymbol{\epsilon}(t)) \tag{A4}$$

Substituting Eq. (A3) for $\boldsymbol{E}_{\text{eval}}(t)$:

$$\boldsymbol{M_s}(t+1) = \boldsymbol{M_s}(t) - \boldsymbol{\alpha} \odot (\boldsymbol{M_s}(t))^{\circ 3} + \boldsymbol{G}_{\text{eff}}(t) \odot (-\mathbf{W}_{\text{Fe}}\boldsymbol{M_s}(t) + \boldsymbol{c}_{\text{Fe}} + \boldsymbol{\mu}_{\text{eff}}(t) + \boldsymbol{\epsilon}(t)) \tag{A5}$$

where $\odot$ denotes the Hadamard (element-wise) product. This equation describes how each dimension i of $\boldsymbol{M_s}$ evolves: $M_{s,i}(t+1) = M_{s,i}(t) - \alpha_i M_{s,i}(t)^3 + G_{\text{eff},i}(t)(-(\mathbf{W}_{\text{Fe}}\boldsymbol{M_s}(t))_i + c_{\text{Fe},i} + \mu_{\text{eff},i}(t) + \epsilon_i(t))$. This highlights that the dynamics of each subjective dimension are influenced by a complex interplay of its current state, its specific gain and bias, noise, saturation, and the (potentially cross-dimensional) evaluation of the overall subjective state vector by $\mathbf{W}_{\text{Fe}}$.

Simplified Vectorized Update Rule for $\boldsymbol{M_s}$

A notable simplification of the general vectorized update rule (Eq. (A5)) occurs if the Evaluation Filter (Fe) effectively acts as an identity transformation on its input error signal. This corresponds to setting the evaluation weight matrix to the identity matrix ($\mathbf{W}_{\text{Fe}} = \mathbf{I}$) and the evaluation offset vector to zero ($\boldsymbol{c}_{\text{Fe}} = \mathbf{0}$). Under these specific conditions, the evaluated error becomes $\boldsymbol{E}_{\text{eval}}(t) = -\mathbf{I}\boldsymbol{M_s}(t) + \mathbf{0} = -\boldsymbol{M_s}(t)$. Substituting this into Eq. (A4) yields the simplified update rule, designated as Eq. (A6):

$$\begin{aligned} \boldsymbol{M_s}(t+1) = \boldsymbol{M_s}(t) - \boldsymbol{\alpha} \odot (\boldsymbol{M_s}(t))^{\circ 3} \\ + \boldsymbol{G}_{\text{eff}}(t) \odot (-\boldsymbol{M_s}(t) + \boldsymbol{\mu}_{\text{eff}}(t) + \boldsymbol{\epsilon}(t)) \end{aligned} \tag{A6}$$

This form (Eq. (A6)) directly parallels the structure of the scalar model's update rule (Eq. (1)), where the term $F_e(e) = e$ (with $e = M_o - M_s = -M_s$) was assumed for the scalar analysis. To further clarify its correspondence with the scalar model, Eq. (A6) can be expanded as:

$$\begin{aligned} \boldsymbol{M_s}(t+1) = (\mathbf{1} - \boldsymbol{G}_{\text{eff}}(t)) \odot \boldsymbol{M_s}(t) - \boldsymbol{\alpha} \odot (\boldsymbol{M_s}(t))^{\circ 3} \\ + \boldsymbol{G}_{\text{eff}}(t) \odot \boldsymbol{\mu}_{\text{eff}}(t) + \boldsymbol{G}_{\text{eff}}(t) \odot \boldsymbol{\epsilon}(t) \end{aligned} \tag{A6}$$

where $\mathbf{1}$ is an N-dimensional vector of ones. Each component i of Eq. (A6) (and its expanded form Eq. (A6)) is:

$$M_{s,i}(t+1) = (1 - G_{\text{eff},i}(t))M_{s,i}(t) - \alpha_i M_{s,i}(t)^3 + G_{\text{eff},i}(t)\mu_{\text{eff},i}(t) + G_{\text{eff},i}(t)\epsilon_i(t).$$

This simplified equation provides a direct multidimensional extension of the scalar model's core dynamics, where each subjective dimension interacts with its corresponding effective gain, bias, and noise, alongside a dimension-specific saturation.

Core Philosophy and Characteristics of the SGBD Dynamical System

Emphasis on (Bi)Linearity: Advantages for Interpretability and Tractability

A core design philosophy of the SGBD framework is the prioritization of (bi)linear operations in its key transformation stages. This includes:

- The structural generation of effective gain $\boldsymbol{G}_{\text{eff}}(t)$ and effective bias $\boldsymbol{\mu}_{\text{eff}}(t)$ via matrix multiplication of source input vectors by structural matrices $\mathbf{M}_G$ and $\mathbf{M}_\mu$ (Eqs. (A1), (A2)).

- The affine transformation performed by the Evaluation Filter (Fe) on the error vector (Eq. (A3)).

- The summation of terms within the main dynamical update rule (Eqs. (A5) and (A6)) before the element-wise multiplication by $\boldsymbol{G}_{\text{eff}}(t)$.

This emphasis on a fundamentally (bi)linear core (i.e., involving linear transformations like matrix-vector products and vector additions, combined with element-wise products which introduce bilinear interactions) offers significant advantages. These include enhanced model interpretability, where the contribution of individual components (e.g., source inputs, structural matrix elements, filter parameters) to the overall subjective dynamics can be more clearly traced. It also improves mathematical tractability for analysis and parameter estimation, and provides a degree of transparency that facilitates comparison with other modeling approaches and potential neurobiological mechanisms (e.g., gain modulation in neural circuits).

The Role of Additive Nonlinearity for Stability

The primary source of nonlinearity within the SGBD update rules (Eqs. (A5) and (A6)) is the additive nonlinear saturation term, $-\boldsymbol{\alpha} \odot (\boldsymbol{M_s}(t))^{\circ 3}$. This term, analogous to the cubic term in the scalar model (Eq. (1)), is crucial for ensuring system stability. It acts as a restoring force that prevents the unbounded escalation of any component of the subjective state vector $\boldsymbol{M_s}(t)$, thereby reflecting realistic psychological or neurological regulatory mechanisms that maintain subjective experience within certain bounds. The additive nature of this nonlinearity is intentional, as it allows the system to exhibit complex behaviors (like multistability, if present) while preserving the relative clarity and interpretability of the underlying (bi)linear interactions that drive the system towards or away from equilibrium points.

SGBD as an Evolving Framework: Current Status and Path Forward

An Exploratory Theoretical Tool for Individualized Subjectivity

The SGBD framework, as presented in this appendix, is primarily an exploratory theoretical extension of the main paper's foundational scalar model. Its main contribution at this stage is to offer a conceptual and mathematical pathway towards modeling the rich tapestry of individual differences in subjective experience. This is achieved by proposing specific mechanisms, notably the structural matrices ($\mathbf{M}_G, \mathbf{M}_\mu$) and vectorized filters (like $\mathbf{W}_{\text{Fe}}, \boldsymbol{c}_{\text{Fe}}$), for how common underlying psychological factors or inputs might be differentially weighted, combined, and processed to produce person-specific patterns of subjective dynamics. This approach moves beyond archetypal modeling towards a system capable of representing diverse cognitive-affective architectures.

The Challenge of Empirical Validation and Theoretical Refinement

While the SGBD framework's current emphasis on a (bi)linear core (complemented by additive saturation) offers advantages for initial conceptualization, development, and interpretability (see the Core Philosophy and Characteristics of the SGBD Dynamical System section), the ultimate utility, predictive power, and ecological validity of this architectural choice must be rigorously assessed against empirical data. Real-world subjective phenomena are inherently complex, and their accurate modeling may eventually necessitate the incorporation of more intricate or different types of nonlinear interactions than those currently formulated within SGBD. The current SGBD formulation serves as a foundational, yet incomplete, step. Key future research directions must therefore include:

- Development of robust methods for estimating SGBD parameters (e.g., elements of $\mathbf{M}_G, \mathbf{M}_\mu, \mathbf{W}_{\text{Fe}}, \boldsymbol{c}_{\text{Fe}}, \boldsymbol{\alpha}$) from diverse empirical data sources (e.g., psychometric questionnaires, behavioral experiments, neuroimaging, digital phenotyping).

- Systematic empirical validation of the SGBD model's predictions regarding individual differences in subjective dynamics and responses to perturbations or interventions.

- Iterative refinement of the theoretical constructs—including the nature of the source inputs, the specific forms of the structural transformations, and the necessity and types of nonlinearities—based on empirical findings and in dialogue with established psychological and neuroscientific theories.

- To advance from the simplified internal dynamics discussed here to a more complete theory, future work must elaborate the mathematical structures of the Recognition (F_r) and Interpretation (F_i) filters. This will enable a full-system model, based on Figure 1, that investigates how external world inputs are processed to dynamically shape the subjective state (M_s), removing the simplifying assumptions made in this appendix.

The path from this conceptual SGBD framework to a fully validated and practically applicable predictive model of individualized subjectivity will require substantial and sustained interdisciplinary effort, integrating theoretical modeling, computational experimentation, and rigorous empirical investigation.

Appendix B

A Narrative Illustration of Self-Regulatory Dynamics

This fictional narrative illustrates how a monk attains Enlightenment during a thousand-day spiritual regimen, interpreted through the dynamical model of subjectivity developed in this paper (the trajectory is conceptually visualized in Figure 9). The temporal unit is one day ($t = 1$). The stages of this transformation are summarized in Table 6.

Stage 1. The Swamp of Anxiety and Fixation (Days 1–200)

During the first 200 days the monk's subjective state (M_s) was dominated by a self-denigrating bias ($\mu < 0$). His affective gain was hyper-sensitive ($G \approx 1$), so each external stimulus merely reinforced his negative convictions. Both M_s and μ remained locked in the negative region, convincing him he was hopeless and that his suffering would last forever. In model terms, the Self-System (S) was immature: its objective function ($\mathcal{L}$) of mental equanimity was far from satisfied, and learning of its meta-parameters ($\boldsymbol{\theta}_{\mathrm{S}}$) had stalled.

Stage 2. Rupture and Dissociation (Days 201–400)

Shortly after day 200 a dramatic mental shift occurred. The long-standing hyper-sensitivity G collapsed, crossed zero, and plunged into the negative domain. Reality-testing faltered, and dissociation set in. Simultaneously, M_s abruptly departed from the entrenched negative bias μ and swung into positive territory. Bewildered, the monk felt an elation that clashed with an inner voice still muttering of worthlessness. S could not track the sudden transition; its meta-parameters lost efficacy, widening the gap between the current state and the objective $\mathcal{L}$.

Stage 3. The Fortress of Delusion (Days 401–600)

From day 400 the mind settled into a bistable region. Isolated from the external world (G at its negative peak), M_s fixed itself in ungrounded euphoria—a manic defense that provided only a spurious equilibrium. The monk convinced himself he had "surpassed even the Buddha." S, abandoning the long-term objective $\mathcal{L}$ of durable serenity, settled for a proximal optimum, and $\boldsymbol{\theta}_{\mathrm{S}}$ were updated into a maladaptive configuration. The negative gain ($G < 0$) signaled extreme instability and a latent risk of depressive collapse.

Stage 4. First Light (Days 601–700)

After day 600 the first signs of recovery appeared, reportedly triggered by natural scenery and a conversation with a child. The long-dominant bias μ crossed into positive territory, allowing nascent hope to arise. Although G remained negative and the ground unstable, the shift in belief prepared the way for deeper change. Recognizing his earlier euphoria as illusory, the monk began to adopt a more objective stance. This self-reflection reopened learning within S, prompting a re-evaluation of $\mathcal{L}$ and renewed calibration of $\boldsymbol{\theta}_{\mathrm{S}}$ toward healthier values.

Stage 5. Awakening and Re-integration (Days 701–800)

From day 701 the mind entered an Awakening phase. Sensitivity G turned positive again, restoring open exchange with the world. Guided by the now-positive bias μ, reality perception normalized, allowing M_s to re-align with μ and dissolve the long-standing conflict. The monk could finally meet reality with calm courage. A quiet insight crystallized: "Everything simply is, and my mind moves in harmony with it." Executive control returned to S; the revised objective $\mathcal{L}$, emphasizing coherence with reality, was largely met, and $\boldsymbol{\theta}_{\mathrm{S}}$ were substantially optimized.

Stage 6. Governed Serenity—Enlightenment (Days 801–1000)

In the final phase the awakened mind consolidated its gains. With sensitivity tuned to $G \approx 1$, S recognized that any residual bias μ breeds suffering—mental defilement. Pursuing maximal tranquility, it actively regulated μ toward zero by refining $\boldsymbol{\theta}_{\mathrm{S}}$. On day 1000 the monk reached a dynamically poised, unwavering peace—the model's Enlightenment state (ideal dynamical equilibrium: $G = 1, \mu = 0$). Learning within S was conclusive, $\mathcal{L}$ was robustly satisfied, and $\boldsymbol{\theta}_{\mathrm{S}}$ had converged to their optimal values. The mind now responded to events like a leaf fluttering in the wind while its center remained perfectly still—it was the inner peace the monk had finally reached after a 1000-day quest.

Appendix C

ADHD as Dysregulation in the G–μ–S Framework

Objective and Scope

This appendix explores an illustrative application of the G–μ–S framework to Attention-Deficit/Hyperactivity Disorder (ADHD). We conceptualize ADHD as a dynamical disorder of self-regulation, emerging from impairments in affective gain (G), cognitive bias (μ), and higher-order control by the Self-System (S). While exploratory, this mapping demonstrates how the model may unify diverse symptoms, neural mechanisms, and therapeutic interventions within a coherent theoretical architecture. We aim to:

- Link ADHD core symptoms to parameter-space instabilities;
- Integrate recent findings in neuroscience and computational psychiatry;
- Sketch potential applications for clinical prediction and intervention modeling.

Theoretical Framework Overview

In the G–μ–S model, G represents affective sensitivity (gain), μ encodes stable cognitive predispositions, and S adapts both over time to maintain psychological stability. This model defines a "Mind Topography Map" in the G–μ plane, partitioned into stable, bistable, and unstable regions.

ADHD is interpreted here as an impairment in S-driven adaptation. Dysfunctions may manifest as:

- Aberrant G values (e.g., excessive or insufficient reactivity),
- Volatile or maladaptive μ formation,
- Reduced meta-learning rate in $\boldsymbol{\theta}_{\mathrm{S}}$, the Self-System's internal parameters.

Parameter Interpretation of ADHD Symptoms

Each core ADHD symptom cluster maps to specific dynamic configurations, as summarized in Table 7.

Inattention. may result from low G, reducing sensitivity to error signals and slowing corrective updates. When μ remains unstable or unconsolidated, S cannot maintain focus in even stable regions of the landscape.

Hyperactivity and Impulsivity. are modeled as high G dynamics that over-amplify momentary stimuli or fluctuations in μ. This leads to frequent switching or divergence in state trajectories, especially when S control is compromised.

Emotional Dysregulation. may emerge from unstable μ and elevated G, placing the system near bifurcation thresholds. The inability of S to maintain equilibrium in such regions may lead to chaotic transitions or bistable flipping.

Neural and Computational Substrates

Neuroscientific findings support mappings between these parameters and biological processes:

- G correlates with signal-to-noise regulation via dopamine and noradrenaline (e.g., prefrontal-striatal circuits).
- μ reflects learning history, expectations, and working memory, supported by medial prefrontal and hippocampal activity.
- S plausibly corresponds to frontopolar–anterior cingulate–basal ganglia loops engaged in hierarchical control.

Major therapeutic interventions can be interpreted as targeting specific model parameters to restore regulation, as outlined in Table 8. Pharmacological treatments such as methylphenidate may act by increasing G toward optimal ranges (near $G = 1$), while behavioral therapies may target long-term reshaping of μ and $\boldsymbol{\theta}_{\mathrm{S}}$.

Dynamic Implications and Instabilities

The G–μ–S model predicts several dynamic behaviors observed in ADHD:

- High intra-individual variability due to noise amplification in unstable zones;
- Abrupt switching between focused and unfocused states in bistable regimes;
- Long-term failure to converge toward the Ideal Dynamical Equilibrium (IDE).

These predictions offer a unified dynamical explanation for heterogeneity across individuals and timepoints.

Future Directions

This theoretical sketch suggests empirical and translational paths:

- Use pupil diameter and neural variability as candidate biomarkers for G.
- Estimate meta-learning rates (η_θ) via longitudinal behavioral tasks.
- Simulate intervention effects by shifting parameter trajectories toward the IDE.

Longitudinal coupling of computational parameters with symptom trajectories may enable individualized modeling and targeted regulation.

Conclusion

This appendix is exploratory. It does not propose clinical diagnoses or treatment protocols, but rather demonstrates how ADHD may be framed as a self-regulatory disturbance within a control-theoretic model. The G–μ–S framework may thus serve as a bridge between computational formalism, empirical measurement, and clinical understanding.

Disclaimer

This appendix is intended solely to explore potential applications of the underlying article and does not constitute medical or psychological advice. The

ADHD G–μ–S dynamical model presented herein is a hypothetical framework; its clinical validity has not yet been established.

When reusing or citing this material, you must reproduce this disclaimer in full, clearly indicate any modifications made, and comply with the repository's license. The author accepts no liability for any loss or damage arising from use of this content. If you do not agree to these terms, please refrain from using this material.

Appendix D

Smith's *Impartial Spectator* and the G–μ–S Framework

Revisiting 18th-Century Moral Psychology Through a Computational Lens

This appendix offers a philosophical reflection on conceptual parallels between Adam Smith's classic work, *The Theory of Moral Sentiments* (1759), and the G–μ–S framework developed in this paper. Although computational models often seek forward-looking innovation, historical accounts of the mind can provide surprisingly aligned conceptual insights. Adam Smith's moral psychology, centered on internal regulation and imagined perspective-taking, anticipates several core themes in contemporary theories of self-regulation.

In Smith's framework, moral judgment emerges not from abstract reasoning alone, but through a process of internal simulation: the agent imagines an "Impartial Spectator" observing and evaluating their conduct. This mental construct plays a regulatory role, guiding the self toward "propriety" and, ultimately, "virtue."

From a modern standpoint, this process bears striking resemblance to the operation of a Self-System (S) that adaptively adjusts affective gain (G) and cognitive bias (μ) to achieve optimal self-regulation. Both systems involve internal feedback, meta-level evaluation, and the gradual shaping of a moral or regulatory stance through experience.

Structural Parallels and Interpretive Mapping

Table 9 summarizes the interpretive correspondence between Smith's concepts and components of the G–μ–S model. These mappings are not intended as direct derivations but as conceptual resonances. They suggest that longstanding philosophical frameworks may encode structural intuitions that modern computational models can formalize and extend.

In particular:

- The "Impartial Spectator" operates analogously to a meta-controller or Self-System

(S) that maintains coherence between perception, action, and internal evaluation.

- Smith's notion of "virtue" corresponds to the model's Ideal Dynamical Equilibrium (IDE), representing a state of poised responsiveness and internal harmony.
- "Self-command" reflects the dynamical adjustment of internal parameters, especially affective sensitivity and cognitive predisposition.
- The gradual development of "moral character" parallels the learning of meta-parameters ($\boldsymbol{\theta}_{\mathrm{S}}$) that guide ongoing regulation.

This interpretive mapping is especially relevant for integrating normative theories of agency with computational descriptions of adaptive control.

Conclusion: Philosophical Echoes and Theoretical Enrichment

This appendix does not claim historical foresight on Smith's part, nor does it suggest that 18th-century moral theory anticipated modern control architectures in a literal sense. Rather, it illustrates that enduring philosophical insights about the structure of the mind—particularly those concerning internal guidance, feedback, and norm formation—may find renewed expression through contemporary formal models.

Such cross-temporal resonance highlights the value of historical moral theory as a resource for grounding and humanizing abstract computational constructs. Conversely, the G–μ–S model provides a formal vocabulary for articulating the mechanisms implicit in Smith's vision of self-regulation, sympathy, and the pursuit of moral integrity.

Postscript. This appendix serves as a heuristic reflection and conceptual coda. It does not contribute to the derivation of formal results, but aims to broaden the theoretical scope and invite interdisciplinary dialogue.

Appendix E

Dynamics and Psychological Implications of the Evaluation Weight Matrix ($\mathbf{W}_{\mathbf{Fe}}$)

Introduction: An Extension from the Simplified Model

This appendix extends the analysis of the Structural Gain–Bias Dynamics (SGBD) framework introduced in Appendix A. The analysis in that appendix was largely built upon a simplified update rule (Simplified Vectorized Update Rule for $\boldsymbol{M_s}$) which assumed that the evaluation weight matrix $\mathbf{W}_{\text{Fe}}$ is an identity matrix ($\mathbf{W}_{\text{Fe}} = \mathbf{I}$) and the offset vector $\boldsymbol{c}_{\text{Fe}}$ is zero ($\boldsymbol{c}_{\text{Fe}} = \mathbf{0}$). As stated in the conclusion of Appendix A (SGBD as an Evolving Framework: Current Status and Path Forward), the SGBD framework is presented as an "Evolving Framework" and a "Path Forward" for more detailed modeling.

This appendix actualizes that vision by systematically exploring the rich and complex dynamics that emerge when we relax the assumption of $\mathbf{W}_{\text{Fe}} = \mathbf{I}$. We demonstrate how the mathematical properties of the $\mathbf{W}_{\text{Fe}}$ matrix—specifically, whether it is diagonal, symmetric, or non-symmetric—qualitatively transform the system's behavior. This extension allows the model to capture a wider spectrum of psychological phenomena, from simple independent states to the complex, cyclical patterns characteristic of many real-world subjective experiences.

A Systematic Classification of Dynamics Based on the Properties of $\mathbf{W}_{\mathbf{Fe}}$

This section analyzes the mathematical and psychological implications for three distinct cases of the $\mathbf{W}_{\text{Fe}}$ matrix.

Case 1: The Diagonal Matrix ($\mathbf{W}_{Fe} = \mathbf{D}$) — An Independent System

Mathematical Features. When $\mathbf{W}_{\text{Fe}}$ is a diagonal matrix, each dimension of the subjective state evolves independently. The system is a gradient system with a separable potential function. Attractors are always simple fixed points.

Psychological Interpretation. This case models a psychological state where different domains (e.g., work, family) are compartmentalized and do not influence one another, potentially representing emotional flattening or dis-integrated cognition.

Case 2: The Symmetric Matrix ($\mathbf{W}_{Fe} = \mathbf{W}_{Fe}^{\top}$) — A Coupled Gradient System

Mathematical Features. When $\mathbf{W}_{\text{Fe}}$ is symmetric with non-zero off-diagonal elements ($W_{ij} = W_{ji}$), the dimensions become coupled. The system, however, remains a *gradient system*, and its attractors are restricted to *fixed points*. However, multistability (multiple stable points) can emerge, corresponding to the "Bistable" region in the scalar model (Stability Regimes: Definitions and Characteristics).

Psychological Interpretation. This models direct interactions between psychological states (e.g., coupling, competition). Multistability can represent distinct psychological "modes" (e.g., an optimistic mode vs. a pessimistic mode) between which an individual can switch.

Case 3: The Non-Symmetric Matrix ($\mathbf{W}_{Fe} \neq \mathbf{W}_{Fe}^{\top}$) — A Rotational Non-Gradient System

Mathematical Features. When $\mathbf{W}_{\text{Fe}}$ is non-symmetric ($W_{ij} \neq W_{ji}$), the system is no longer a gradient system due to a *rotational component.* This allows for the emergence of *dynamic attractors*, most notably *limit cycles.*[1]

Psychological Interpretation. This is the key to modeling persistent, dynamic psychological patterns, such as ruminative cycles in depression, mood swings in bipolar disorder, or vacillation and indecision in states of conflict.

Example 1 (Illustrating the Three Cases)**.** To understand the qualitative differences, consider a 2D system with $\boldsymbol{\alpha} = (1, 1)^{\top}$. We can observe the transition in dynamics by varying a single parameter η in the matrix $\mathbf{W}_{\text{Fe}}$.

- *Diagonal ($\eta = 0$):* With $\mathbf{W}_{\text{Fe}} = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, the two dimensions are decoupled and converge independently to the origin.

[1] In discrete-time systems, periodic orbits can take several forms, including fixed points with a period $k > 1$ (period-k orbits) and invariant closed curves, the latter of which typically arise from a Neimark–Sacker bifurcation.

- *Symmetric ($\eta \neq 0$):* With $\mathbf{W}_{\text{Fe}} = \left[\begin{smallmatrix} 1 & \eta \\ \eta & 1 \end{smallmatrix}\right]$, the dimensions are coupled. The system still converges to a fixed point, but the trajectory is influenced by the interaction term.

- *Non-Symmetric ($\eta \neq 0$):* With $\mathbf{W}_{\text{Fe}} = \left[\begin{smallmatrix} 1 & \eta \\ -\eta & 1 \end{smallmatrix}\right]$, a rotational force is introduced. For small η, the system spirals into the fixed point. For this specific system, the Neimark–Sacker bifurcation occurs and a limit cycle is born precisely when the parameter $|\eta|$ crosses the critical value $\eta_c = 1$.

Conceptual Leap: From Static Stability to Dynamic Stability

The analysis of $\mathbf{W}_{\text{Fe}}$ expands the very concept of "stability." The symmetric cases model *static stability*, where the system settles into a fixed, unchanging state (a fixed-point attractor). The non-symmetric case introduces *dynamic stability*, where the system settles into a persistent, self-sustaining *pattern of change* (a limit cycle attractor). This transition often occurs via a *Neimark–Sacker bifurcation*, the discrete-time equivalent of the continuous-time Hopf bifurcation. This allows the model to formalize the idea that a psychological "rut" can be a dynamic cycle, not just a static belief.

Mathematical Formulation and Stability Analysis

A Note on the Potential Function in a Discrete System. It is important to note that the model herein is a discrete-time system ($\boldsymbol{M_s}(t+1) = f(\boldsymbol{M_s}(t))$), whereas the concept of a potential function is most strictly defined for continuous-time gradient systems ($\dot{\boldsymbol{M_s}} = -\nabla V(\boldsymbol{M_s})$). Therefore, the term "potential function" is used, as in the main text (Mathematical Analysis of M_s Dynamics (Fixed G and μ)), as a *conceptual analogy* or *quasi-potential.* The rigorous stability analysis of any fixed point $\boldsymbol{M_s^*}$ relies on the eigenvalues of the system's Jacobian matrix, $J = \frac{\partial f}{\partial \boldsymbol{M_s}}$, at that point. To be precise, *local stability of a fixed point requires all eigenvalues of the Jacobian matrix to have a magnitude less than 1.*

Formulation for the Symmetric Case. For the symmetric case ($\mathbf{W}_{\text{Fe}} = \mathbf{W}_{\text{Fe}}^{\top}$ and, for simplicity, $\boldsymbol{c}_{\text{Fe}} = \mathbf{0}, \boldsymbol{G}_{\text{eff}} = \mathbf{1}$), a quasi-potential function $V(\boldsymbol{M_s})$ can be defined such that the deterministic component of the update, $f(\boldsymbol{M_s}) - \boldsymbol{M_s}$, corresponds to the negative gradient $-\nabla V$. The fixed points $\boldsymbol{M_s^*}$ satisfy $\mathbf{W}_{\text{Fe}}\boldsymbol{M_s^*} + \boldsymbol{\alpha} \odot (\boldsymbol{M_s^*})^{\circ 3} = \mathbf{0}$, and the corresponding quasi-potential is:

$$V(\boldsymbol{M_s}) = \frac{1}{2}\boldsymbol{M_s}^{\top}\mathbf{W}_{\text{Fe}}\boldsymbol{M_s} + \frac{1}{4}\sum_{i=1}^{N}\alpha_i M_{s,i}^4 \tag{E1}$$

The term $\frac{1}{2}\boldsymbol{M_s}^{\top}\mathbf{W}_{\text{Fe}}\boldsymbol{M_s}$ contains the quadratic cross-terms (e.g., $W_{ij}M_{s,i}M_{s,j}$) that couple the dimensions.

Formulation for the Non-Symmetric Case. If $\mathbf{W}_{\text{Fe}}$ is non-symmetric, its governing vector field acquires a rotational component, quantifiable by the curl. For a 2D field $\boldsymbol{F} = (F_x, F_y)$, this is $\operatorname{curl}\boldsymbol{F} = \partial F_y/\partial x - \partial F_x/\partial y$, which is non-zero for a non-symmetric system. Decomposing the matrix $\mathbf{W}_{\text{Fe}} = \mathbf{W}_{\text{sym}} + \mathbf{W}_{\text{skew}}$, the dynamics are driven by both a gradient force derived from $\mathbf{W}_{\text{sym}}$ and a rotational force from $\mathbf{W}_{\text{skew}}$, which drives the system in orbits and enables limit cycles. A Neimark–Sacker bifurcation can occur when a pair of complex conjugate eigenvalues of the Jacobian matrix crosses the unit circle.

Dynamics Near the IDE. Near the Ideal Dynamical Equilibrium ($G = 1, \mu = 0$), the linear term in the scalar model vanishes. In the SGBD framework, this means the dynamics are dominated by the competition between the cubic saturation term $-\boldsymbol{\alpha} \odot \boldsymbol{M_s}^{\circ 3}$ and the forces from $\mathbf{W}_{\text{Fe}}$. In the presence of a non-symmetric $\mathbf{W}_{\text{Fe}}$, even if the origin $\boldsymbol{M_s} = \mathbf{0}$ is a stable fixed point, trajectories will approach it in a spiral. This is a consequence of the Jacobian matrix of the local linearization having complex conjugate eigenvalues for a non-symmetric $\mathbf{W}_{\text{Fe}}$. This implies that an individual's unique evaluative "wiring" ($\mathbf{W}_{\text{Fe}}$) can create dynamic response characteristics even in the absence of explicit cognitive bias ($\boldsymbol{\mu}_{\text{eff}}$).

A Note on Practical Stability. For practical applications such as clinical modeling or parameter estimation, it is important to consider the "safe range" of parameters to avoid divergent or chaotic behavior. If the elements of $\mathbf{W}_{\text{Fe}}$, particularly the non-symmetric components, become too large relative to the stabilizing saturation terms

($\boldsymbol{\alpha}$), the system can lose stability. Future work should aim to define these stability boundaries, providing design guidelines for model fitting.

Conclusion: Enhancing the Explanatory Power of the SGBD Framework

In summary, this analysis of the evaluation weight matrix $\mathbf{W}_{\mathrm{Fe}}$ demonstrates that the SGBD framework is not a single model but a versatile class of models. By adjusting the properties of $\mathbf{W}_{\mathrm{Fe}}$, the framework can generate qualitatively different dynamics, moving from simple, independent systems to complex, coupled, and cyclical ones. This extension provides a robust and mathematically principled foundation for future research, offering a pathway to building personalized models that can capture the dynamic "wiring" of subjective experience. However, the high parameter count of $\mathbf{W}_{\mathrm{Fe}}$ raises challenges of *identifiability.* Practical applications will likely require imposing constraints, such as *sparsity* or *low-rank structure*, on the matrices, and employing *regularization* techniques when fitting models to empirical data.

Appendix F

A Taxonomy of Subjective States: Psychodynamics of the $\boldsymbol{G}_{\text{eff}}$-$\mathbf{W}_{\text{Fe}}$ Interaction

Introduction: Synthesizing Dynamics for a Richer Psychological Model

This appendix builds directly upon the analysis in Appendix E, which established how the properties of the evaluation weight matrix ($\mathbf{W}_{\text{Fe}}$) can generate diverse dynamics from simple convergence to stable cycles. Here, we take a crucial next step by exploring the synergistic interaction between $\mathbf{W}_{\text{Fe}}$ and the sign of the affective gain vector ($\boldsymbol{G}_{\text{eff}}$). As the scalar model demonstrated, negative gain ($G < 0$) introduces a powerful repulsive, self-amplifying dynamic, interpreted as an "escapist" mode (Quadrant Typology and Dynamic Heuristics).

By systematically combining these two fundamental mechanisms—the structural coupling/rotation from $\mathbf{W}_{\text{Fe}}$ and the dynamic amplification/attenuation from $\boldsymbol{G}_{\text{eff}}$—we can construct a remarkably rich taxonomy of subjective states. This synthesis allows the framework to move beyond modeling simple states or traits and begin to describe the underlying mechanics of complex and often pathological psychological *processes*. This extension provides a formal bridge from our theoretical model to the complex, dynamic phenomena central to computational psychiatry.

A 2x2 Taxonomy of Psychodynamic States

We can classify the resulting dynamics into a 2x2 matrix based on the sign of the gain (simplifying $\boldsymbol{G}_{\text{eff}}$ to a scalar G for conceptual clarity) and the properties of $\mathbf{W}_{\text{Fe}}$ (symmetric vs. non-symmetric). Each quadrant of this matrix corresponds to a qualitatively distinct psychodynamic regime.

Case 1: Adaptive Regime ($G > 0$)

This is the regime of error-correction and homeostatic regulation, where the system is generally pulled towards stable states. The dynamics explored in Appendix E primarily fall within this regime.

Symmetric $\mathbf{W}_{\mathbf{Fe}}$ (Stable Competition). The system features multiple stable fixed points, representing a healthy competition between established psychological modes (e.g., a "work self" vs. a "family self"). This allows for flexible, context-appropriate switching between different but stable ways of being.

Non-Symmetric $\mathbf{W}_{\mathbf{Fe}}$ (Stable Cycles). The system can exhibit stable limit cycles, representing persistent but bounded patterns like habitual rumination or predictable mood fluctuations. While cyclical, these patterns do not escalate, modeling a stable personality trait or a manageable thinking style rather than a runaway pathology.

Case 2: Escapist Regime ($G < 0$)

This is the regime of positive feedback and error-amplification, where the system is actively pushed *away* from normative states. This dynamic dramatically transforms the potential landscape defined by $\mathbf{W}_{\text{Fe}}$, turning stable valleys into unstable hills.

Symmetric $\mathbf{W}_{\mathbf{Fe}}$ (Repulsion from All Modes). The combination of repulsion ($G < 0$) and multiple defined modes (symmetric $\mathbf{W}_{\text{Fe}}$) creates a state of severe internal conflict. The stable attractors of the adaptive regime now become unstable repellers. The system is pushed away from *all* of its established modes of being simultaneously. The resulting state, balanced precariously between multiple repulsive forces and the global saturation term, can be interpreted as a model for profound psychological distress, such as a sense of *alienation, deep internal conflict, or even dissociative states* where one feels detached from all familiar aspects of the self.

Non-Symmetric $\mathbf{W}_{\mathbf{Fe}}$ (Catastrophic Rumination). This combination represents arguably the most pathological dynamic the model can produce. The rotational force from the non-symmetric $\mathbf{W}_{\text{Fe}}$ is fused with the expansive, amplifying force of $G < 0$. The result is not a stable cycle, but an *unstable spiral vortex.* Thoughts and emotions do not simply circle; they escalate with each rotation. For example, a minor worry is forced back by the rotational component, only to be amplified in intensity by the negative gain, leading to a self-fueling cycle of catastrophic thinking. This dynamic provides a powerful mechanistic model for severe, treatment-resistant conditions like *Obsessive-Compulsive Disorder (OCD), severe anxiety disorders, or the emotional dysregulation loops in Borderline Personality Disorder (BPD)*, where obsessive thoughts and emotional intensity feed each other in a vicious cycle.

Matrix of Psychological Interpretations

The taxonomy described above can be concisely summarized in Table F1. This matrix serves as a heuristic map, linking the mathematical architecture of the model to a spectrum of psychological phenomena, from adaptive functioning to severe psychopathology.

Table F1

A 2x2 Matrix of Psychodynamic States from Gain–$\mathbf{W}_{Fe}$ Interaction.

	Symmetric $\mathbf{W}_{\mathbf{Fe}}$ (Competitive)	**Non-Symmetric $\mathbf{W}_{\mathbf{Fe}}$ (Rotational)**
$G > 0$ **(Adaptive)**	*Healthy Conflict / Mode Switching* Stable competition between states. *e.g., Deciding between two goals.*	*Stable Rumination / Thinking Style* Bounded, predictable cyclical patterns. *e.g., A habitual worrying pattern.*
$G < 0$ **(Escapist)**	*Severe Internal Conflict / Dissociation* Repulsion from all established modes. *e.g., Feeling alienated from all parts of self.*	*Catastrophic Rumination / Vicious Cycles* Self-amplifying, runaway cyclical patterns. *e.g., Panic attacks, obsessive loops.*

Conclusion: From Static States to Pathological Processes

The synthesis of affective gain and the evaluation matrix dramatically expands the explanatory power of the SGBD framework. It demonstrates how a few fundamental principles can interact to generate a vast landscape of subjective dynamics. Crucially, this allows us to model mental illness not merely as a “state” (e.g., a low mood fixed point) but as a “process”—a self-sustaining, pathological dynamic (e.g., a stable limit cycle of catastrophic rumination).

This process-oriented perspective is a cornerstone of computational psychiatry. By providing a formal, mechanistic language to describe these maladaptive cycles, this framework offers a powerful tool for generating testable hypotheses about the underlying dynamics of mental disorders and for simulating the potential effects of interventions aimed at disrupting these vicious cycles and restoring adaptive regulation.

Appendix G

Discrete-in-Continuous Framework for Subjectivity: Three-Tiered Integration

Introduction: A New Framework for Subjective Dynamics

Recent advances in cognitive science, particularly in computational psychiatry, have been driven by powerful theoretical frameworks such as the *Free Energy Principle (FEP)* and *Perceptual Control Theory (PCT)*. The FEP, in particular, has provided an exceptionally versatile foundation, successfully describing in a unified manner how the brain engages in continuous prediction and control while interacting with its environment. Its contributions are substantial.

However, the very versatility of a general theory like the FEP, which primarily employs continuous-time dynamics as its language of description, also implies certain limits in its resolution for specific phenomena. This becomes apparent when considering the *non-smooth, saltatory nature* that often characterizes our internal subjectivity. Consider, for instance, the following phenomena:

- The *"Aha!" moment*, an instantaneous shift from a fixed mental set to an entirely new perspective.
- The *sudden mood shift*, a rapid plunge from a calm state to intense anger or deep sorrow.
- The *episodic onset* of psychiatric symptoms, which can manifest dramatically over a few hours or days.

These phenomena are better described not as an accumulation of continuous changes, but as qualitative *jumps from "State A" to "State B"*—catastrophic events in which the system's global behavior switches discontinuously. While continuous models like the FEP provide an excellent description of the physical substrate, a more specialized mathematical engine is required to explicitly handle such discontinuous subjective dynamics.

To bridge this theoretical gap—that is, to connect the *continuous physical substrate described by the FEP* with the *discrete subjective experiences we live through*—we propose the *"Discrete-in-Continuous" paradigm* as a complementary extension. This is not a confrontation with existing theories but the introduction of a new architecture for understanding the dynamics of subjectivity.

The purpose of this appendix is to outline this new paradigm, demonstrating how its *"Causation-Structure-Norm"* three-tiered dynamic cycle can integrate and leverage the strengths of established theories like FEP and PCT, thereby shedding new light on the problem of subjective discontinuity.

A New Architecture of the Mind: The "Causation-Structure-Norm" Model

The "Discrete-in-Continuous" paradigm allows us to reconceptualize the mind not as a monolayered system, but as a dynamic, three-tiered structure where each layer has distinct roles and timescales. We call this new mental architecture the *"Causation-Structure-Norm" model.* This is an integrative framework that respects the contributions of FEP and PCT, positioning them within a larger, more comprehensive picture.

First Tier: Causation — The Continuous Physical Substrate Described by FEP. This is the foundational layer of the model, representing the physical basis of our existence. It is a world governed by the continuous flow of energy and matter—neuronal firings, hormonal secretions, and sensory inputs. The dynamics of this layer are precisely the domain that the *FEP so brilliantly describes through its principle of free energy minimization*, driven by the strict *causality* of physical laws.

Second Tier: Structure — The Discrete Subjective Landscape Generated by the G-μ-S Model. This is the core contribution of our theory. *Emerging* from the continuous physical processes of the first tier is the "structure" of this second tier. This is not a material entity but the *space of possible subjective states—the "Mind Topography Map"* itself.

This landscape contains “valleys” (attractors) corresponding to stable mental states and “ridges” (bifurcation points) corresponding to instability. To contrast with FEP’s variational free energy approach, consider its energy functional, $F = \mathbb{E}_q[\ln q(s) - \ln p(o, s)]$, which minimizes prediction error in a continuous framework. The G-μ-S model complements this by providing a concrete dynamical process. More generally, the G-μ-S model is a discrete map comprising linear and nonlinear terms. In this paper, we adopt a specific form using a cubic polynomial, as it is one of the simplest models capable of producing the multistability central to our theory. This allows it to capture non-linear bifurcations and noise-driven transitions:

$$M_s(k+1) = \underbrace{(1 - G\Delta t)M_s(k)}_{\text{Inertia/Self-coupling}} - \underbrace{\alpha M_s(k)^3 \Delta t}_{\text{Stabilization}} + \underbrace{G\mu\Delta t}_{\text{Attraction to Goal}} + \underbrace{w_k}_{\text{Fluctuation}} \tag{G1}$$

This discrete formulation allows the G-μ-S model to handle catastrophic state jumps (e.g., emotional shifts), enhancing FEP’s continuous predictive power. The stable points (attractors) of this system are given by the solutions, M_s^*, to the following equation:

$$\alpha(M_s^*)^3 + GM_s^* - G\mu = 0 \tag{G2}$$

A crucial question arises here: if FEP and PCT are founded on principles of error minimization (i.e., negative feedback), does our model’s inclusion of an error-amplifying regime ($G < 0$, positive feedback) contradict them? Answering this question reveals the true strength of our three-tiered model. The first (FEP) and third (PCT) tiers define the *normative state* the system ought to achieve—namely, the adaptive regime of $G > 0$. In contrast, the second tier’s G-μ-S model is not a normative engine; it is a more general *dynamical engine* that neutrally describes what *can* happen given a set of parameters. Thus, the $G < 0$ regime is explicitly framed as a *maladaptive or pathological state* that arises from a *deviation* from FEP’s normative principle or a *failure* of PCT’s control. The strength of our framework lies in its ability to explain not only healthy functioning ($G > 0$) but also how the system can ‘break’ and enter the complex pathological dynamics detailed in Appendix F, all within a single, unified theory. One might ask why the brain would create a discrete "mental map" from a continuous physical world. We propose that this is a strategy for *computational efficiency.* Processing a continuous state space with infinite gradations is computationally expensive. By compressing and simplifying the world into discrete states (attractors) like "calm," "anxious," or "focused," the brain can *accelerate decision-making* and simplify action planning. The discrete structure described by the G-μ-S model is thus an intelligent and adaptive *cognitive shortcut* for navigating a complex reality. This choice is further motivated by a fundamental property of dynamical systems. Discrete maps, such as the logistic map, can generate complex, chaotic behavior in just one dimension. In contrast, continuous systems—specifically, autonomous and smooth ordinary differential equations like the Lorenz or Rössler systems—require at least three dimensions to exhibit such behavior.

Third Tier: Norm — Teleological Self-Control Governed by PCT. At the highest level lies the domain of teleological control explored by PCT: the world of "norms." The agent (the *Self-System*) surveys the "Mind Topography Map" of the second tier from this vantage point. It then makes judgments based on *norms (goals and values)*, such as "Is my current state desirable?" This normative control aligns with PCT's hierarchical negative feedback loops, where the *Self-System* adjusts G and μ based on error signals ($e = r - p$) between desired references (r) and perceived states (p), integrating PCT's action-driven dynamics into the slower structural evolution. This regulation can be formulated as a hybrid control system. Specifically, the slow updates to G and μ can be described as a gradient descent on two distinct objective functions:

$$G_{k+1} = G_k - \eta_G \cdot \nabla_G \mathcal{L}_{\text{free}}(G, \mu) \tag{G3}$$

$$\mu_{k+1} = \mu_k - \eta_\mu \cdot \nabla_\mu \mathcal{L}_{\text{disc}}(M_s, G, \mu) \tag{G4}$$

Here, η_G and η_μ are learning rates that determine the step size for each parameter update. $\mathcal{L}_{\text{free}}$ is a free-energy-like loss function that ensures alignment with the external world (Tier 1), while $\mathcal{L}_{\text{disc}}$ is a loss function based on the stability of subjective attractors, reflecting internal goals (Tier 3). This functional separation—where G handles continuous tuning to the world and μ handles discrete selection of internal goals—is how the *Self-System* implements hybrid control. This hierarchical control extends across PCT's multi-level structure—from high-level goals (e.g., self-identity) to low-level actions (e.g., motor outputs like arm movements)—reflecting PCT's ten-level hierarchy (Powers, 1973).

Table G1

The roles of each tier in the "Causation-Structure-Norm" model.

Tier	**Worldview**	**Role**	**Corresponding Theory**	**Timescale**
First Tier	World of Causation	Physical Causality	FEP	Fast
Second Tier	World of Structure	Semantic Landscape	G-μ-S	Medium
Third Tier	World of Norms	Teleological Control	PCT	Slow

The "Implementation Bridge": Turning Philosophy into Science with H and Δt

A fundamental question remains for this three-tiered model: *"How can the abstract 'norms' of the third tier influence the physical 'causation' of the first tier?"*

We answer this question with the concept of the *"Implementation Bridge."* This bridge also provides a concrete answer to how the continuous prediction error dynamics of FEP might be implemented as a biologically plausible discrete process. The two pillars of this bridge are the *Hold-scheme (H)* and the *Sampling-width (Δt)* of the brain's information processing.

- *Hold-scheme (H)*: The "style" in which the brain processes continuous information (e.g., Zero-Order Hold, Forward Euler).
- *Sampling-width (Δt)*: The rhythm or temporal resolution at which subjectivity is "updated."

It is important to note that the *Hold-scheme (H)* is not necessarily identical to the standard sample-and-hold circuit in electronics, which corresponds to a Zero-Order Hold

(ZOH). Whereas a ZOH simply maintains a sampled value, the *Forward Euler* scheme, which is of particular importance in this paper, is a more active process that *predictively updates* the next state based on the current state and its rate of change. This distinction is critical for understanding the dynamics of subjectivity, because only a predictive process can capture its active nature, where self-inertia is dynamically modulated by one's engagement with the world (i.e., by the gain parameter G).

When the *Self-System* adopts a "norm" such as "I want to be more flexible," this corresponds to a higher-order process in the FEP framework of searching for a better generative model. This search is translated into the physical process of adjusting the brain's information processing style (H) and update rhythm (Δt).

The choice of H and Δt dramatically alters the system's behavior. The coefficient Φ_H, which determines the strength of self-coupling (i.e., the coefficient of the $M_s(k)$ term in Equation G1), is highly dependent on the scheme H, demonstrating that this "implementation" choice is not a mere approximation but a critical determinant of the system's qualitative behavior.

Table G2

Major Hold-schemes (H) and their corresponding self-coupling coefficients (Φ_H).

Hold/Discretization H	**Self-Coupling Coefficient** $\Phi_H(G, \Delta t)$
ZOH (Zero-Order Hold)	$e^{-G\Delta t}$
Tustin (Bilinear)	$(1 - G\Delta t/2)/(1 + G\Delta t/2)$
Forward Euler	$\mathbf{1 - G\Delta t}$

As shown, *only in the Forward Euler scheme* does the self-coupling coefficient Φ_H become exactly zero at the singular point $G\Delta t = 1$. This creates the *IDE state*, where the system is freed from its own inertia and can respond most sensitively to external perturbations. This IDE state can be interpreted as a computational representation of a therapeutic mental state, akin to the focus on the "here and now" in practices like *Japanese Zen meditation*. It may represent an ideal form of the *"flexible and open mind"* that many psychotherapies aim to cultivate.

Thus, H and Δt provide a *concrete mechanism* that bridges the gap between the continuous world of FEP (First Tier) and the teleological world of PCT (Third Tier). This "Implementation Bridge" transforms the philosophical debate of "Is the mind discrete or continuous?" into the scientifically testable question: "With which H and Δt does the real brain implement the principles of FEP?"

Conclusion: Contributions of the G-μ-S Theory and New Research Horizons

The "Discrete-in-Continuous" paradigm outlined in this appendix opens new horizons for the scientific study of subjectivity. The core contributions of this work can be summarized in three points:

1. **Contribution 1: A "Mathematical Engine" for Subjective Multistability**
   We have provided a specific and tractable *mathematical engine (the G-μ-S model)* that complements the general principles of FEP to describe the discontinuous state transitions of subjectivity.

2. **Contribution 2: A "New Architecture" for Integrating Major Theories**
   We have positioned FEP (Causation) and PCT (Norms) not in opposition but in synergy within a *new "Causation-Structure-Norm" architecture.*

3. **Contribution 3: A "Verification Roadmap" for Turning Philosophy into Science**
   We have shown that this theory is testable via the *"Implementation Bridge" (H and Δt)*, opening a path to complement the data-driven approach of FEP with experimental designs focused on the discrete states of subjectivity.

Empirical Validation as a Central Challenge and Opportunity. Perhaps the greatest challenge for this theoretical framework, and simultaneously its greatest promise, lies in its empirical validation. Bridging this theory to empirical data is a formidable task, yet it opens up exciting avenues for future research. We hope that this framework provides a useful theoretical container for experimentalists to explore these possibilities. For instance, one could envision an experimental approach, which we might term a **`Hold-swap` paradigm**, where participants perform a cognitive task (e.g., an affective decision-making task) while the timing (Δt) and update style (instructions suggesting a specific H) of information presentation are manipulated. Such a paradigm might be expected to produce specific modulations in the time-series dynamics of BOLD (Blood-Oxygen-Level-Dependent) signals measured with *fMRI* (e.g., in the DMN (Default Mode Network) and limbic systems). We look forward to future collaborations where this neuroimaging data could be synergistically combined with behavioral (e.g., reaction times) and physiological (e.g., HRV) data to provide a richer, multi-modal test of the model's core tenets.

However, a central challenge remains: robust methods for estimating the parameters (G, μ, H, Δt) from real-world time-series data have yet to be established. Scenarios using *wearable devices (e.g., HRV data)* and mood logs are promising, but the identifiability of these parameters is a non-trivial issue that requires further investigation. Furthermore, the assumptions about the noise distribution (w_k) and the quantification of individual differences are critical topics for future work.

These challenges highlight that our theory constitutes a rich, new research program for approaching subjectivity with a constructive methodology. Looking forward, this framework opens the door to a *"computational individualized psychopathology,"* where each person's "Mind Topography Map" can be inferred from data.

Under the new light shed by this work, we are confident that research into the dynamics of subjectivity will advance dramatically, in concert with the great existing theories of the mind.