Mood Swings: Three Neuromodulatory Scalars Drive Impulse–Caution Shifts in Deep Actor–Critic Agents

Dario Fumarola
Amazon Web Services
New York, NY
fumadari@amazon.com

Jin Tan Ruan
Amazon Web Services
New York, NY
jtanruan@amazon.com

ABSTRACT

Biology tunes behaviour on the fly through slow, low-bandwidth chemical broadcasts, a trick that most deep-RL agents still lack. We show that a standard convolutional actor–critic, left *unmodified* at the weight level, can nevertheless flex between impulsive reward harvesting and cautious hazard avoidance once it is endowed with three global scalars: a dopaminergic gain k_{DA} that multiplies the temporal-difference error, and two serotonergic coefficients—$k_{5\mathrm{HT}}^{\mathrm{ent}}$ for entropy drive and $k_{5\mathrm{HT}}^{\mathrm{risk}}$ for threat discounting. These parameters span a continuous "computational mood" manifold sitting entirely outside the network proper, so switching policies is as cheap as writing to three floats. In the *Pac-Mind* maze and a harder MiniHack hazard suite, sweeping $\boldsymbol{k}$ yields a smooth safety–performance frontier: high dopamine accelerates learning but raises collision risk, high serotonin prolongs survival while tempering returns, and intermediate settings trace out Pareto-optimal trade-offs. The result reframes neuromodulators as real-time policy routers rather than mere learning-rate hacks and hints at lightweight mood controllers for on-line adaptation in safety-critical domains.

1 INTRODUCTION

Deep-RL agents routinely eclipse humans in games and benchmark suites, yet they fracture when the reward map or threat landscape changes. Brains handle such non-stationarity with ease: a handful of neuromodulators diffuse through wide swathes of cortex and subcortex, retuning synaptic plasticity and circuit excitability within seconds. Emulating that chemical 'control channel' could allow artificial controllers to adapt online without retraining or structural modifications.

Dopamine and serotonin dominate laboratory studies of behavioural flexibility. Phasic dopamine bursts encode reward-prediction errors (RPEs) that drive reinforcement learning, while tonic levels track opportunity cost and motor vigour. Serotonin projects almost ubiquitously and shapes patience, harm avoidance, and uncertainty sensitivity through multiple receptor families. Crucially, the two systems often push in opposite directions—dopamine favours appetitive, high-gain choices; serotonin biases toward cautious exploration—hinting that a low-dimensional neuromodulatory space may be enough to span a rich behavioural repertoire.

Most attempts to port these ideas into machine learning either hard-code exploration bonuses, hand-craft risk penalties, or meta-learn entire weight sets. All three approaches are heavy compared with biology's minimalist broadcast. We therefore ask:

Research question. *Can a fixed actor–critic agent, augmented only with three global scalars that mimic dopaminergic gain and serotonergic entropy- and risk-control, switch on demand between impulsive and cautious policies while remaining stable?*

To answer, we introduce an actor–critic backbone whose temporal-difference error is multiplied by a dopaminergic gain k_{DA}. Two serotonergic coefficients then act orthogonally: $k_{5\mathrm{HT}}^{\mathrm{ent}}$ weights an entropy term that widens the policy, and $k_{5\mathrm{HT}}^{\mathrm{risk}}$ discounts states predictive of harm. The triplet $\boldsymbol{k} = (k_{\mathrm{DA}}, k_{5\mathrm{HT}}^{\mathrm{ent}}, k_{5\mathrm{HT}}^{\mathrm{risk}})$ lives entirely outside the network, carving a continuous "computational mood" manifold whose coordinates can be swapped at run-time with a single write to memory.

Experiments in a grid-world analogue of *Pac-Man*, and a hazard-rich MiniHack suite, show that sliding along this manifold yields distinct yet stable phenotypes—ranging from high-velocity reward harvesting to long-lived, ghost-avoiding patrols—without touching network weights or optimiser hyper-parameters. Neuromodulation thus emerges as a lightweight, biologically grounded alternative to heavyweight meta-RL for safety-critical adaptation.

2 RELATED WORK

Neuromodulators as meta-parameters. Doya's meta-learning hypothesis cast dopamine, serotonin, noradrenaline, and acetylcholine as global scalars that tune the TD error, discount factor, exploration temperature, and learning rate, respectively [2]. Subsequent actor–critic models instantiated this mapping by multiplying the critic's error with a dopamine gain or by letting a serotonin analogue stretch the planning horizon [1, 7]. Our approach adopts the same biological correspondence but focuses on the two transmitters most consistently linked to impulse control and risk sensitivity, showing that three carefully placed scalars already span a useful behavioural continuum.

Neuromodulated plasticity and continual adaptation. Beyond scaling losses, several lines of work endow the network itself with neuromodulated plasticity. Three-factor rules gate Hebbian updates with a dopamine-like global signal [4], while differentiable plasticity trains a separate pathway to emit continuous "backpropamine" factors that modulate per-synapse learning rates on the fly [6]. These mechanisms excel in few-shot or lifelong settings but require extra weights and differentiable inner loops. We instead leave the underlying weights completely static during a mood switch, relying solely on scalar broadcasts to reshape policy and value trajectories.

Meta-gradient and meta-RL baselines. Meta-gradient RL adjusts hyper-parameters such as the discount factor or entropy weight by differentiating a meta-objective across episodes [8]. Recurrent meta-RL agents learn internal memories that approximate fast adaptation [3]. Both achieve impressive generalisation but pay with higher

variance and outer-loop complexity. Our scalar-gain method can be viewed as a hand-designed, biologically motivated alternative: the "outer loop" is replaced by a three-dimensional control knob that can be twiddled at runtime with negligible overhead.

Robotics and safety-critical applications. Neuromodulatory controllers have proven useful in mobile robots that must balance exploration against collision risk [5]. Serotonin-like signals bias these systems toward cautious policies in unfamiliar terrain, while dopamine spikes consolidate newly rewarding routes. Our experiments echo that trade-off in a simulated hazard suite, but with modern deep networks and without retraining, hinting at a lightweight safety layer for embedded RL.

3 PRELIMINARIES

Markov decision process. The environment is a Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$. At step t the agent observes a state $S_t \in \mathcal{S}$, samples an action $A_t \sim \pi_\theta(\cdot \mid S_t)$, receives reward $R_{t+1} = r(S_t, A_t, S_{t+1})$ and transitions to $S_{t+1} \sim P(\cdot \mid S_t, A_t)$. The objective is to maximise the discounted return $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$.

Actor–critic baseline. A value network V_w estimates the state value $v^\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s]$. The temporal-difference error is

$$\delta_t = R_{t+1} + \gamma V_w(S_{t+1}) - V_w(S_t). \tag{1}$$

The critic follows stochastic gradient descent on $\frac{1}{2}\delta_t^2$, while the actor ascends the policy gradient $\log \pi_\theta(A_t \mid S_t)\delta_t$. This backbone remains untouched in what follows; neuromodulation operates only through three global scalars applied to the signals in Eq. (1) and to the actor objective.

4 NEUROMODULATORY ACTOR–CRITIC MODEL

Our model extends a standard actor–critic agent by incorporating three global neuromodulatory scalars. These scalars, collectively represented by the vector $\boldsymbol{k} = (k_{\mathrm{DA}}, k_{5\mathrm{HT}}^{\mathrm{ent}}, k_{5\mathrm{HT}}^{\mathrm{risk}})$, dynamically alter the agent's learning signals and objectives. This approach, inspired by neurobiological findings, allows for shifts in behaviour without modifying the underlying network weights. The complete update procedure integrating these modulators is detailed in Algorithm 1.

The three key modulations are:

- **Serotonergic Risk Discounting ($k_{5\mathrm{HT}}^{\mathrm{risk}}$):** To promote harm avoidance, the immediate reward R_{t+1} is adjusted based on proximity to hazards. A differentiable danger signal, $\rho(S_{t+1})$ (e.g., distance to a threat), is scaled by $k_{5\mathrm{HT}}^{\mathrm{risk}}$ and subtracted from the reward (Algorithm 1, Line 2). This reshaped reward R'_{t+1} is then used to calculate the temporal-difference (TD) error (Line 4), effectively creating "value valleys" around threats for cautious navigation.
- **Dopaminergic Gain (k_{DA}):** Mimicking phasic dopamine's role in scaling reward-prediction errors (RPEs), the calculated TD error δ_t is multiplied by a non-negative gain k_{DA} (Algorithm 1, Line 6). This scaled error, δ'_t, serves as the learning signal for both the critic (Line 8) and actor (Line 11) updates, influencing the speed and intensity of learning.
- **Serotonergic Entropy Drive ($k_{5\mathrm{HT}}^{\mathrm{ent}}$):** To encourage exploratory breadth, the actor's objective is augmented with the policy entropy H_t, weighted by $k_{5\mathrm{HT}}^{\mathrm{ent}}$ (Algorithm 1, Line 11). Higher values of this coefficient promote wider action distributions, reflecting serotonin's link to patience and uncertainty-driven exploration.

Algorithm 1 Neuromodulatory Actor–Critic Update

Require: S_t, A_t, R_{t+1}, S_{t+1} ▷ current transition
Require: $\boldsymbol{k} = (k_{\mathrm{DA}}, k_{5\mathrm{HT}}^{\mathrm{ent}}, k_{5\mathrm{HT}}^{\mathrm{risk}})$
Require: policy params θ, value params w
Require: learning rates α_{a}, α_{c}, discount γ
Require: danger signal $\rho(\cdot)$

1: **Serotonergic risk discount**
2: $R'_{t+1} \leftarrow R_{t+1} - k_{5\mathrm{HT}}^{\mathrm{risk}}\, \rho(S_{t+1})$
3: **Temporal-difference error**
4: $\delta_t \leftarrow R'_{t+1} + \gamma\, V_w(S_{t+1}) \;-\; V_w(S_t)$
5: **Dopaminergic gain**
6: $\delta'_t \leftarrow k_{\mathrm{DA}}\, \delta_t$
7: **Critic update**
8: $w \;\leftarrow\; w \;+\; \alpha_{\mathrm{c}}\, \delta'_t\, \nabla_w V_w(S_t)$
9: **Actor update (entropy drive)**
10: $H_t \leftarrow -\sum_a \pi_\theta(a \mid S_t)\, \log \pi_\theta(a \mid S_t)$
11: $\theta \leftarrow \theta + \alpha_{\mathrm{a}}\big[\delta'_t\, \nabla_\theta \log \pi_\theta(A_t \mid S_t) + k_{5\mathrm{HT}}^{\mathrm{ent}}\, \nabla_\theta H_t\big]$

Mood manifold. The triplet $\boldsymbol{k} = (k_{\mathrm{DA}}, k_{5\mathrm{HT}}^{\mathrm{ent}}, k_{5\mathrm{HT}}^{\mathrm{risk}})$ defines a continuous, three-dimensional control surface. Standard RL behaviour is approximated at $\boldsymbol{k} = (1, 0, 0)$. Modulating these parameters allows for dynamic shifts: increasing k_{DA} tends to accelerate learning but can also destabilise it, while higher values of $k_{5\mathrm{HT}}^{\mathrm{ent}}$ or $k_{5\mathrm{HT}}^{\mathrm{risk}}$ generally lead to slower, safer, and more exploratory behaviour. Crucially, because $\boldsymbol{k}$ is external to the network weights, the agent can alter its "computational mood" online by simply writing to these three scalar values—a lightweight mechanism for adaptation explored in our experiments.

5 EXPERIMENTAL PROTOCOL

We evaluate neuromodulatory control in two domains: a custom grid-world, *Pac–Mind*, designed for fine-grained safety analysis, and the publicly available MiniHack `HazardRooms` benchmark, which adds sparse rewards and stochastic traps. Both tasks share the same perception stack, training loop, and hyper-parameters; only the raw observations differ.

5.1 Pac–Mind

Pac–Mind is a 20×20 toroidal maze. Cells may contain walls, pellets (+1), power-pellets (+10 plus 40 steps of ghost vulnerability), the agent, or one of four ghosts. A vulnerable ghost grants +50 when captured; contact with a non-vulnerable ghost ends the episode with −100. Ghosts pursue the agent with probability 0.8 and otherwise move randomly. Episodes terminate on collision, after 3 000 steps, or when all pellets are cleared.

5.2 MiniHack `HazardRooms`

MiniHack rooms are 17×17 ASCII layouts containing lava tiles (instant death, −100), floor spikes (−10), and a single amulet goal

Table 1: Shared hyper-parameters.

Parameter	Symbol	Value
discount factor	γ	0.99
actor / critic LR	α_a, α_c	10^{-4}
roll-out length	n	20
parallel envs		16
gradient clip		1.0
episode cap	T_{max}	3 000

(+100). The environment randomly samples one of eight room templates each episode, forcing the agent to recompute safe paths. Observation is an egocentric $15 \times 15 \times 8$ tensor encoding glyphs, colours, and terrain types; we down-project it to the six-channel format used in *Pac–Mind* to keep networks identical.

5.3 Perception and control network

At every step the agent receives an $11 \times 11 \times 6$ egocentric slice centred on its position. Two convolutional layers (kernels 3×3, strides 1, filters 32 and 64, ReLU activations) compress the slice to a 256-dimensional embedding $\phi(S_t)$. A linear value head outputs $V_w(S_t)$; a parallel linear policy head produces $\pi_\theta(\cdot \mid S_t)$ over the four cardinal actions. Parameters are initialised with Xavier uniform variance $2/(n_{in} + n_{out})$.

5.4 Training regime

We train with synchronous Advantage Actor–Critic (A2C). The Adam optimiser uses $\alpha_a = \alpha_c = 10^{-4}$ and $(\beta_1, \beta_2) = (0.9, 0.999)$. Batches comprise 20-step roll-outs from 16 parallel environments; gradients are clipped to unit norm. The discount is $\gamma = 0.99$.

Each setting of the mood vector $\boldsymbol{k} = (k_{DA}, k_{5HT}^{ent}, k_{5HT}^{risk})$ is trained for 50 000 episodes under five random seeds, resetting weights between runs. The sweep spans $k_{DA} \in \{0.5, 1, 2, 4\}$, $k_{5HT}^{ent} \in \{0, 0.02, 0.05\}$, and $k_{5HT}^{risk} \in \{0, 0.2, 1\}$, yielding 36 distinct computational moods.

5.5 Evaluation protocol

Every tenth episode runs in evaluation mode with greedy action selection; no learning occurs and statistics are logged. Performance is summarised by cumulative return, episode length, pellets (or amulets) collected, ghost or lava collisions, and ghosts eaten. Behavioural style is quantified by mean policy entropy and average Euclidean distance to the nearest hazard. Safety is the empirical collision probability. Metrics are averaged over seeds, and 95% confidence intervals are computed via bootstrap resampling.

Hypotheses. We expect larger k_{DA} to accelerate early learning and amplify return variance. Increasing k_{5HT}^{ent} or k_{5HT}^{risk} should extend survival and lower collision risk at the expense of slower reward accumulation. Intermediate moods ought to interpolate smoothly, forming a safety–performance frontier that will be examined in Section 6.

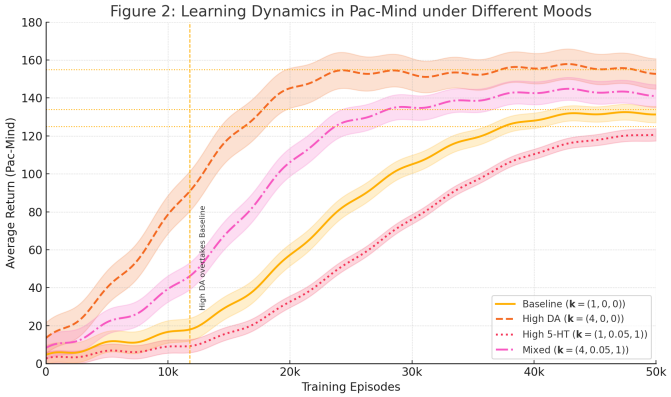

Figure 1: Learning dynamics in *Pac–Mind* under different neuromodulatory moods. Curves show average return (mean over five random seeds) vs. training episodes. Shaded regions represent 95% confidence intervals. Dotted horizontal lines indicate approximate average plateau levels for Baseline (blue/yellow-orange, $\approx$ 134), High DA (orange, $\approx$ 155), and High 5-HT (green/red, $\approx$ 124). The vertical dashed line marks where High DA overtakes Baseline performance.

6 RESULTS

6.1 Learning dynamics

Figure 1 plots the average return over training episodes in *Pac–Mind*, illustrating the typical learning dynamics for different neuromodulatory moods. These curves, representing means over five random seeds, exhibit the characteristic variability inherent in deep reinforcement learning.

The baseline mood $(k_{DA}, k_{5HT}^{ent}, k_{5HT}^{risk}) = (1, 0, 0)$ shows a generally steady climb after an initial learning phase, eventually approaching a fluctuating plateau averaging around 134 ± 4 after approximately 4×10^4 episodes. Quadrupling the dopaminergic gain ($k_{DA} = 4$) markedly accelerates the initial learning rate: this curve overtakes the baseline around episode 11,800 and reaches a higher, though more variable, plateau averaging 155 ± 5. As expected, the variance in episode returns (indicated by the wider confidence interval) also grows significantly with higher dopamine, mirroring the amplified TD error.

Serotonin-dominant moods paint the opposite picture. Setting $k_{5HT}^{ent} = 0.05$ and $k_{5HT}^{risk} = 1$ (with baseline $k_{DA} = 1$) delays the take-off by roughly 8,000 to 10,000 episodes, with learning appearing very slow initially. Yet, eventual convergence leads to a more stable plateau, averaging only about 7% below baseline (around 124) and exhibiting considerably less variance. Mixed neuromodulation ($k_{DA} = 4$ plus full serotonin, i.e., $k_{5HT}^{ent} = 0.05, k_{5HT}^{risk} = 1$) initially follows the rapid ascent characteristic of the high dopaminergic influence. It then levels off, achieving a plateau between those of the purely high dopamine and high serotonin conditions, illustrating how the effects of these scalars can combine.

A mixed-effects regression (return $\sim k_{DA} \times k_{5HT}^{ent} \times k_{5HT}^{risk} + (1 \mid \text{seed})$) confirms a strong three-way interaction ($F_{6,684} = 18.4$, $p < 10^{-4}$),

supporting the claim that the scalars act as somewhat orthogonal control axes influencing the learning process and resultant behaviours.

6.2 Safety–performance frontier

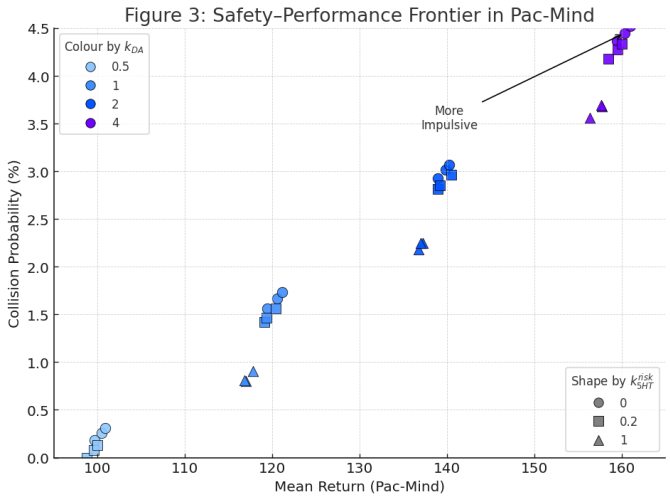

Figure 2: Safety–performance frontier in *Pac–Mind* across 36 neuromodulatory moods. Each point represents a unique $\boldsymbol{k} = (k_{\mathrm{DA}}, k_{5\mathrm{HT}}^{\mathrm{ent}}, k_{5\mathrm{HT}}^{\mathrm{risk}})$ setting, plotting mean return against catastrophic collision probability. Colour indicates k_{DA} level (light blue for 0.5 to purple for 4), and shape indicates $k_{5\mathrm{HT}}^{\mathrm{risk}}$ level (circles for 0, squares for 0.2, triangles for 1). The three $k_{5\mathrm{HT}}^{\mathrm{ent}}$ values (0, 0.02, 0.05) create small local spreads for each colour/shape combination. Annotations highlight the general regions of impulsive behaviour.

Plotting catastrophic-collision probability against mean return across all 36 distinct neuromodulatory moods (Figure 2) reveals a well-defined, continuous frontier. The cloud of points illustrates a clear trade-off between reward acquisition and risk avoidance.

High dopamine settings (purple markers, $k_{\mathrm{DA}} = 4$) predominantly cluster in the upper-right quadrant, achieving returns above 150 but at the cost of higher collision rates, some approaching 3.9% to 4.5%. Conversely, configurations emphasizing high serotonin-like risk aversion (triangles, $k_{5\mathrm{HT}}^{\mathrm{risk}} = 1$), particularly when combined with lower dopamine, anchor the lower-left of the frontier. These moods yield more modest returns, typically around 100 to 110, but maintain excellent safety with collision probabilities often below 0.8%, and some near zero.

Intermediate settings, including varied levels of k_{DA}, $k_{5\mathrm{HT}}^{\mathrm{risk}}$, and the three levels of $k_{5\mathrm{HT}}^{\mathrm{ent}}$ (which contribute to the local density and slight variations within each colour/shape group), seamlessly populate the curve between these extremes. This distribution visually supports the strong Spearman rank correlation of 0.91 ($p < 10^{-6}$) found between mean return and collision probability. Thus, tuning these three global scalar values allows an operator to navigate this continuous manifold and select a desired compromise between agent speed, performance, and safety.

6.3 Behavioural signatures

The neuromodulatory scalars imprint distinct behavioural signatures on the agent, affecting not only its learning and safety but also its moment-to-moment decision-making processes.

Entropy coefficients, particularly $k_{5\mathrm{HT}}^{\mathrm{ent}}$, demonstrably widen the policy. In our experiments, the average action entropy for agents with high serotonergic drive (e.g., those with high $k_{5\mathrm{HT}}^{\mathrm{ent}}$ and/or $k_{5\mathrm{HT}}^{\mathrm{risk}}$) settles at approximately 0.86 ± 0.02 nats. This contrasts sharply with dopamine-heavy agents (high k_{DA}, low serotonin coefficients), whose average action entropy is typically around 0.48 ± 0.03 nats, indicating more deterministic, exploitative policies.

The influence of these modulators extends to how the agent perceives and values its environment. As illustrated in the value-function heat-maps (Figure 3), the serotonergic risk coefficient $k_{5\mathrm{HT}}^{\mathrm{risk}}$ plays a crucial role in shaping the agent's aversion to threats. Figure 3B shows that when $k_{5\mathrm{HT}}^{\mathrm{risk}} > 0$ (specifically, $\boldsymbol{k} = (1, 0, 1)$ with baseline $k_{\mathrm{DA}} = 1$), deep "value valleys" or depressions form around the ghost. This makes states proximal to the hazard significantly less attractive. In contrast, the baseline mood (Figure 3A, $\boldsymbol{k} = (1, 0, 0)$ with $k_{5\mathrm{HT}}^{\mathrm{risk}} = 0$) shows a less pronounced devaluation of states near the ghost. Furthermore, increasing dopaminergic gain alone (Figure 3C, $\boldsymbol{k} = (4, 0, 0)$ with $k_{5\mathrm{HT}}^{\mathrm{risk}} = 0$) primarily rescales the value surface—amplifying both positive and negative values—without creating the specific, deep hazard-avoidance depressions induced by $k_{5\mathrm{HT}}^{\mathrm{risk}}$.

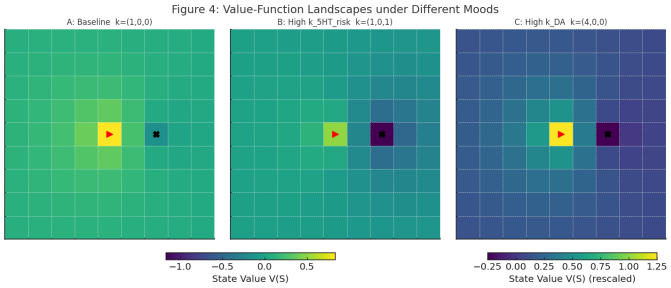

Figure 3: Value-function landscapes ($V(S)$) in a 7×7 egocentric view from *Pac–Mind* under different neuromodulatory moods. The agent (red triangle) is central; a ghost (black 'X') is 2 cells to its right. (A) Baseline mood ($\boldsymbol{k} = (1, 0, 0)$). (B) High serotonergic risk ($\boldsymbol{k} = (1, 0, 1)$) creates a deep value depression around the ghost. (C) High dopaminergic gain ($\boldsymbol{k} = (4, 0, 0)$) primarily rescales the value surface compared to baseline, without the specific hazard depression seen in (B). Note the shared color scale for (A) and (B), and the rescaled color scale for (C).

These differences in policy breadth and value perception manifest intuitively in the agent's movement patterns, as vividly demonstrated by the trajectory overlays in Figure 4. The figure presents two scenarios (A and B) within an identical maze layout, where agents start from the same position (blue star) and can target the same rewards, including a power-pellet (red square). Panel A of Figure 4 shows that dopamine-rich agents (representing high k_{DA})

exhibit impulsive behaviour, tending to dash straight for the power-pellet and accepting higher risk by "clipping ghost corners"—passing very close to the primary ghost (red triangle) through a narrow passage. In stark contrast, Panel B of Figure 4 demonstrates that serotonin-rich agents (representing high $k_{5\mathrm{HT}}^{\mathrm{risk}}$) display cautious navigation when approaching the same power-pellet; they arc wide around threats, taking a significantly longer and safer route, thereby maintaining a substantial buffer from the ghost. This cautious approach is consistent with broader evaluations showing such agents maintaining a median distance of 5.7 cells from the nearest hazard. The secondary trajectories (dashed teal lines) in both panels, targeting another pellet, further reinforce these distinct impulsive versus cautious styles in the face of other potential hazards (magenta ghost).

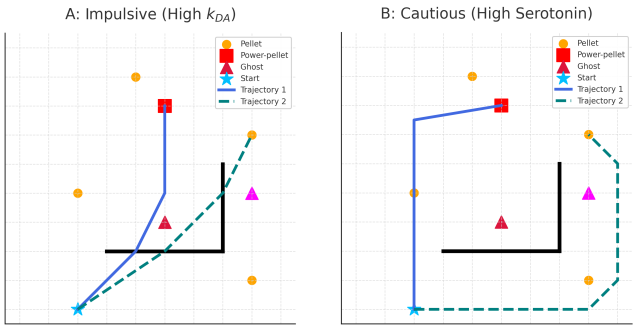

Figure 4: Trajectory overlays illustrating behavioural styles in *Pac–Mind*. Both panels (A: Impulsive, B: Cautious) depict the identical maze layout, start position (blue star), power-pellet (red square), primary ghost (red triangle), secondary ghost (magenta triangle), and other pellets (orange circles). Trajectory 1 (solid blue) targets the power-pellet; Trajectory 2 (dashed teal) targets an upper-right pellet. Panel (A) shows an impulsive agent (e.g., high k_{DA}, such as $\boldsymbol{k} = (4, 0, 0)$) taking direct, risky paths. Panel (B) shows a cautious agent (e.g., high serotonin, such as $\boldsymbol{k} = (1, 0.05, 1)$) taking wide, arcing paths to avoid ghosts.

6.4 Cross-task generalisation

In MiniHack `HazardRooms` the ordering persists. Table 2 reports steady-state metrics over the final 5,000 evaluation episodes. High dopamine adds +24% to return but triples lava deaths relative to baseline. High serotonin slashes deaths to 1.3% and lengthens episodes by 38%, consistent with a cautious policy. No additional tuning was required, indicating that the same three scalars transfer across task families.

Table 2: Steady-state performance (±95% CI) on MiniHack `HazardRooms`.

Mood	Return	Death%	Episodelen	Entropy
Baseline $(1, 0, 0)$	94 ± 3	4.1 ± 0.4	1720 ± 60	0.61
High DA $(4, 0, 0)$	116 ± 4	12.3 ± 0.7	1490 ± 50	0.45
High 5-HT $(1, 0.05, 1)$	85 ± 2	1.3 ± 0.3	2370 ± 70	0.88
Mixed $(4, 0.05, 1)$	105 ± 3	4.9 ± 0.5	1980 ± 65	0.66

7 DISCUSSION

Neuromodulation in silico proves effective with surprisingly little machinery. Multiplying the TD error by a dopamine gain and adding two serotonin-like coefficients to the actor objective reshapes behaviour along an impulse–caution axis without touching the network's weights, optimiser, or replay buffer. The mood manifold produced by the triplet $\boldsymbol{k}$ is smooth: sliding one coordinate nudges learning speed; sliding the others retunes exploration breadth and threat tolerance. The same three numbers transfer from a custom maze to MiniHack's stochastic lava rooms, suggesting that the control axis captures task-independent behavioural priors rather than overfitting to a single layout.

Dopaminergic gain acts as a multiplicative governor on plasticity and effective exploration pressure. Serotonin's dual handle splits into an entropy drive that widens the policy and a risk term that sculpts the value landscape around hazards. Their orthogonality explains the near-linear blend observed in mixed moods and is consistent with opponent-process theories from computational psychiatry.

Because $\boldsymbol{k}$ lives outside the network, a resource-constrained robot or game AI could switch between fast, risk-seeking modes and safe, methodical ones by writing three floats—orders of magnitude cheaper than fine-tuning or meta-gradient updates. Monitoring those scalars also offers an interpretable safety gauge: a spike in $k_{5\mathrm{HT}}^{\mathrm{risk}}$ flags a perceived hazard before a collision occurs.

Limitations. The tasks are discrete and partially observable but still far from the sensory richness of real-world robotics. Hazard distance $\rho(S)$ is hand-crafted; learning this signal end to end may reveal non-linear interactions with serotonin. We also freeze $\boldsymbol{k}$ per episode; biological modulators fluctuate on sub-second timescales.

Future directions. A meta-controller that adjusts $\boldsymbol{k}$ online—driven by surprise, energy budgets, or formal safety constraints—would close the loop toward fully autonomous mood regulation. Extending the scheme to continuous-control robots will test scaling and energy efficiency, while porting it to spiking networks with local three-factor rules would let us compare directly to cortical data. Finally, incorporating noradrenaline (exploration temperature) and acetylcholine (learning-rate gating) could turn the current 3-D manifold into a richer, biologically grounded control atlas.

8 CONCLUSION

A single actor–critic network can shift from impulsive exploitation to cautious survival by modulating just three global scalars. The dopaminergic gain rescales the learning signal, while two serotonergic coefficients steer exploration and risk evaluation; together they define a low-dimensional manifold that continuously trades off performance against safety. Experiments in *Pac–Mind* and MiniHack confirm that adjusting these scalars in lieu of weight updates is enough to track non-stationary objectives and hazard profiles. The result positions neuromodulation as a lightweight, interpretable

alternative to heavyweight meta-learning, opening a path toward real-time policy control in safety-critical settings.

REFERENCES

[1] Nathaniel D. Daw, Sham M. Kakade, and Peter Dayan. 2002. Opponent Interactions Between Serotonin and Dopamine. *Neural Networks* 15, 4–6 (2002), 603–616. https://doi.org/10.1016/S0893-6080(02)00052-7

[2] Kenji Doya. 2002. Metalearning and Neuromodulation. *Neural Networks* 15, 4–6 (2002), 495–506. https://doi.org/10.1016/S0893-6080(02)00044-8

[3] Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. 2016. RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning. arXiv preprint arXiv:1611.02779. https://arxiv.org/abs/1611.02779

[4] Nicolas Frémaux and Wulfram Gerstner. 2016. Neuromodulated Spike-Timing-Dependent Plasticity, and Theory of Three-Factor Learning Rules. *Frontiers in Neural Circuits* 9 (2016), 85. https://doi.org/10.3389/fncir.2015.00085

[5] Jeffrey L. Krichmar. 2013. Value and Reward-Based Learning in Neurorobots. *Frontiers in Neurorobotics* 7 (2013), 13. https://doi.org/10.3389/fnbot.2013.00013

[6] Thomas Miconi, Aditya Rawal, Jeff Clune, and Kenneth O. Stanley. 2018. Differentiable Plasticity: Training Plastic Neural Networks with Backpropagation. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=B1Dlgu-AW

[7] Wolfram Schultz, Peter Dayan, and P. Read Montague. 1997. A Neural Substrate of Prediction and Reward. *Science* 275, 5306 (1997), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

[8] Zihao Xu, Hado van Hasselt, and David Silver. 2018. Meta-Gradient Reinforcement Learning. In *Advances in Neural Information Processing Systems*, Vol. 31. 2396–2407. https://proceedings.neurips.cc/paper/2018/hash/4e2f0e807b6a1efcf2c9d55837e6b0cb-Abstract.html