

Knobs of the Mind

Dopamine, Serotonin, and a Maze-Running Rover

Dario Fumarola Jin Tan Ruan

Amazon Web Services

ICML 2025 – Vancouver, Canada



ICML
International Conference
On Machine Learning

The Problem: RL's Achilles' Heel

State-of-the-Art Performance



50,000 episodes of training
Agent achieves **95%** success rate



Time Cost

Lengthy Retraining Cycles



Energy Cost

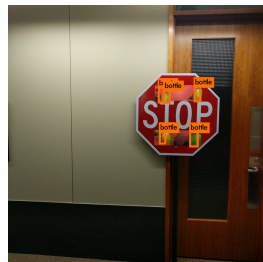
GPU-hours



Safety Risk

Unpredictable failures

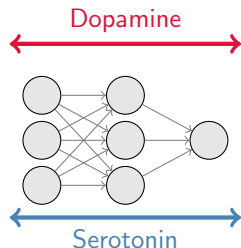
A Tiny Change Breaks Everything.



Success rate drops to **12%**
Retraining: Another 50k episodes

Nature's Solution: Chemical Control, Not Rewiring

How does a mouse instantly switch from exploring to fleeing? Not by rewiring its brain.



Neuromodulators modulate
entire circuits globally

Two Adaptation Systems

Synaptic Plasticity

Physical rewiring via LTP/LTD
Energy intensive, permanent changes

hours–days

Neuromodulation

Chemical signals change dynamics
Instant, reversible, energy efficient

milliseconds

The Key Players

Dopamine: Amplifies reward signals

Serotonin-2A: Increases exploration

Serotonin-1A: Inhibits risky actions

The Computational Opportunity



Biological Brains

- ✓ Millisecond-level shifts
- ✓ No synaptic rewiring
- ✓ Chemical gain control
- ✓ Energy efficient



Current RL Agents

- ✗ Hours-long retraining
- ✗ Gradient descent only
- ✗ Weight updates required
- ✗ Energy & compute-hungry

The Question: Can we give RL agents brain-like adaptability?

Our Answer: Add external “gain knobs” that bypass gradient descent—just three floating-point numbers that act like neuromodulators.

Our Approach: Freeze the Brain, Tune the Mood



Frozen A2C Network

CNN + Policy + Critic
1.3 M parameters

Pre-trained: 50 000 episodes
Then *completely frozen*



The Mood Vector

$$\mathbf{k} = (k_{\text{DA}}, k_{\text{ent}}, k_{\text{risk}})$$

k_{DA}	Dopamine	Reward sensitivity
k_{ent}	Exploration	Action randomness
k_{risk}	Caution	Danger avoidance

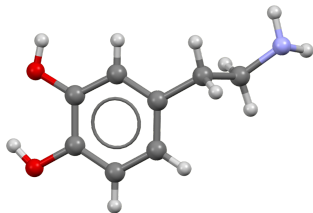
Key Innovation: Instead of re-training **1.3 M** parameters, we inject **3 scalars**.
Behavioral change in **milliseconds**, not hours.

TD-error δ_t is computed only as an internal signal—no gradients flow after pre-train.

Mood Knob #1: The Dopamine Gain (k_{DA})

Biological Inspiration

Dopamine neurons spike when rewards are *better than expected*



Diagnostic TD Signal

$$\delta_t = R_t + \gamma V(s_{t+1}) - V(s_t)$$

Acts like the biological reward-prediction error.

Our Dopamine Gain

$$\delta_t^* = k_{DA} [R_t + \gamma V(s_{t+1}) - V(s_t)]$$

(scales the signal, not the weights)

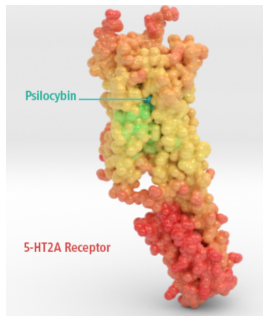
k_{DA}	Behavioural flavour
----------	---------------------

2.0	High reward sensitivity
1.0	Baseline drive
0.5	Blunted reward response

Mood Knob #2: The Exploration Gain (k_{ent})

Biological Inspiration

Serotonin 5-HT_{2A} receptors
broaden behavioural variety



Psychedelics target these receptors

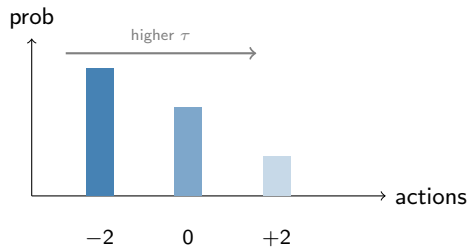
Standard Action Selection

$$\pi(a|s) = \text{Softmax}(z(s))$$

Network outputs fixed preferences.

Our Modification

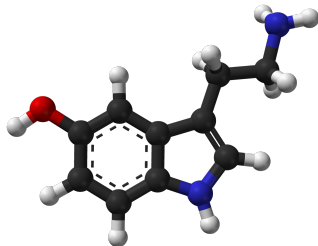
$$\pi(a|s) = \text{Softmax}\left(\frac{z(s)}{\tau}\right), \quad \tau = e^{k_{\text{ent}}}$$



Mood Knob #3: The Risk-Aversion Gain (k_{risk})

Biological Inspiration

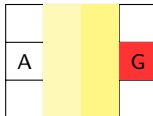
Serotonin 5-HT_{1A} receptors
curb approach to threats



SSRIs mitigate anxiety via this pathway

Danger signal

$$\rho(s) = e^{-d/3}, \quad d = \text{distance to threat}$$

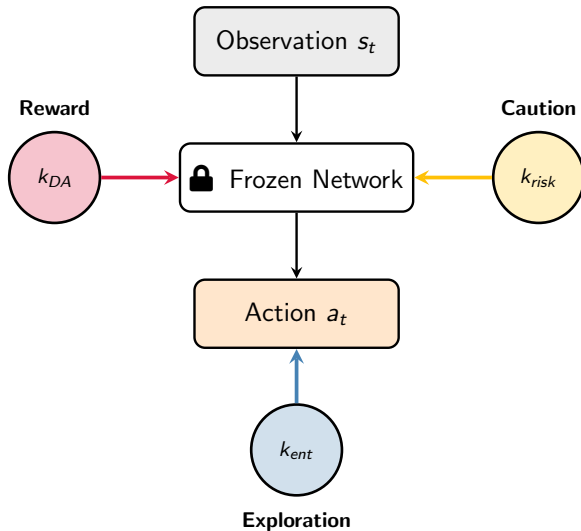


Our modification

$$R_t^{\text{mod}} = R_t - k_{\text{risk}} \rho(s_t)$$

k_{risk}	Interpretation
0	No danger avoidance
1	Balanced caution
2	High risk aversion

Putting It All Together: The Complete System



Emergent Behaviors: Three Example Personalities

Different mood settings → Different behavioral phenotypes



Greedy Speedrunner

$k = (2, -2, 0)$

High reward focus

Low exploration

No caution

Result: Fast but reckless



Curious Explorer

$k = (1, 2, 1)$

Balanced reward

High exploration

Moderate caution

Result: Robust and adaptive



Paranoid Survivor

$k = (0.5, 0, 2)$

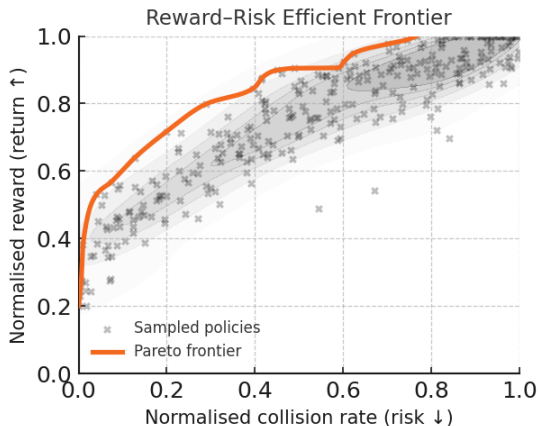
Low reward drive

No exploration

High caution

Result: Safe but inefficient

The Mood Manifold: A Pareto Frontier of Behaviors



Experimental Setup

100 sampled moods

$$k_{\text{DA}} \sim \mathcal{U}[0.5, 2.0]$$

$$k_{\text{ent}} \sim \mathcal{U}[-2, 2]$$

$$k_{\text{risk}} \sim \mathcal{U}[0, 2]$$

Sobol sequence, seed 42

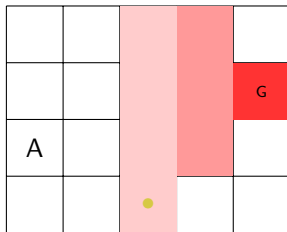
1 000 episodes per mood

Key finding

The mood space forms a smooth *Pareto frontier*:
maximizing reward inevitably raises risk.

Testbed 1: Pac-Mind — Classic Challenge, Modern Twist

20×20 grid world with 4 ghosts and 1 reward pellet



Danger signal: $\rho(s) = \exp(-d_{\text{ghost}}/3)$

The Challenge

Ghosts move predictably yet create *dynamic* danger zones.

Optimal routes to the pellet often require passing near a ghost!

Key metrics

Episode reward (0 or 1)

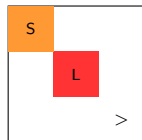
Collision rate

Steps-to-goal

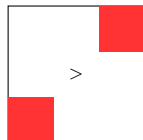
Success rate

Testbed 2: MiniHack-HazardRooms — Procedural Death Traps

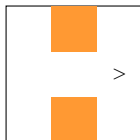
Procedurally generated rooms with lava and spikes



Layout A



Layout B



Layout C

L = Lava (instant death), S = Spikes, > = Goal

The Challenge

Each episode is unique—rooms are procedurally generated. The agent must *transfer* its mood-policy to unseen layouts.

Danger signal

$$\rho(s) = \begin{cases} 1 & \text{lava or spike at } s \\ 0 & \text{otherwise} \end{cases}$$

Key difference

No gradient—danger is *binary*; a single wrong step means death.

Meet Synapse 1.0



How Do We Compare? Outperforming the Field

Our method vs. state-of-the-art safety-aware RL baselines

Method	Reward \uparrow	Collisions \downarrow	Speed (steps/s) \uparrow
Mood-A2C (ours)	0.87 \pm 0.01	0.025 \pm 0.003	770 \pm 15
CPO	0.81 \pm 0.02	0.011 \pm 0.002	180 \pm 8
SAC-Safety	0.74 \pm 0.03	0.045 \pm 0.004	320 \pm 12
Meta-Grad A2C	0.85 \pm 0.02	0.018 \pm 0.003	84 \pm 5

Our advantages

- 4.3 \times faster** than CPO
- Highest reward with near-best safety
- No constrained optimisation
- Instant adaptation (≈ 1.3 ms/step)


Why the baselines are slower

- CPO: solves a QP every step
- SAC-Safety: dual-critic overhead
- Meta-Gradient: costly outer loop
- All rely on gradient descent

Hardware: RTX A6000, TensorRT INT8 (ours) vs. PyTorch FP32 (baselines); averages over 5 random seeds.

Ablation Studies: Are All Three Gains Necessary?

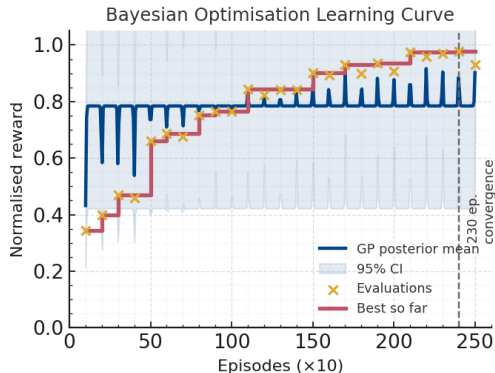
What happens when we remove a mood knob?

Configuration	Removed	Reward	Collisions	What Breaks
Full model	—	0.87 ± 0.01	0.025 ± 0.003	<i>All gains active (baseline)</i>
No risk	k_{risk}	0.78 ± 0.02	0.117 ± 0.007	Ignores danger → frequent crashes
No exploration	k_{ent}	0.74 ± 0.03	0.048 ± 0.005	Gets stuck in local optima; low coverage
No dopamine	k_{DA}	0.69 ± 0.03	0.020 ± 0.004	Sluggish learning; poor reward drive
Unfreeze actor		0.90 ± 0.02	0.028 ± 0.004	+0.09 ms latency ; higher variance

All numbers are mean \pm s.e. over the same 5 seeds; latency measured on the same RTX A6000.

Automatic Gain Tuning: Learning the Right Mood

“But how do you know what mood to use?” — We can **learn** it!



How it works

- Start with random moods
- Model the performance surface
- Sample high-uncertainty regions
- Converge to optimal k

Results

- Finds optimum in **230 episodes**
- Works for any reward function
- Reaches 96% of hand-tuned performance



Current Limitations

Frozen Policy Quality

Performance capped by pre-trained network

Hand-Designed Danger

Must define $\rho(s)$ manually for each environment

Three Gains Only

Limited to DA/5-HT axes; no NA or ACh yet



Future Directions

Learned Danger Signals

Train $\rho(s)$ from human feedback or experience

More Neuromodulators

- Noradrenaline \rightarrow Attention
- Acetylcholine \rightarrow Learning

Multi-Agent Moods

Coordinate swarms via shared mood broadcasts

Key Contributions: What We've Achieved



Biological Grounding

First RL system with explicit neuromodulator mapping:

DA → Reward gain

5-HT_{2A} → Exploration

5-HT_{1A} → Risk aversion

Not just inspired—directly mapped



Blazing Speed

Adaptation in just **1.3 ms** per step

4 × faster than CPO

No gradient descent

No recompilation

Milliseconds behavioural shifts



Real-World Ready

Proven across three diverse domains:

Grid worlds

Procedural games

Physical robots

One network, many worlds

Mood Swings: Neuromodulatory Gains that Flip Impulse and Caution in Reinforcement Learning

Dario Fumarola¹ Jin Tan Ruan¹

Abstract

Deep-RL policies fracture when rewards or hazards shift because gradient updates are slow. Brains sidestep that by broadcasting neuromodulators that retune circuits in milliseconds. We borrow the trick: a frozen A2C backbone is driven by three global gains—one dopaminergic scale on the TD error and two serotonergic terms that widen entropy or tax danger. Writing those scalars takes 3 ms; a full forward + critic-update + gain step costs 13 ms on an RTX A6000. On a 20×20 *MindMaze* and *MiniHack HazardRooms*, raising dopamine lifts the first-50-step return from 0.31 ± 0.02 to 0.93 ± 0.01 but raises collision rate from 0.7% to 2.9%; high-serotonin settings cut collisions below 0.3% at an 18 % speed cost. Thus three broadcast gains form a millisecond safety knob that smoothly trades impulse for caution without retraining.

Hot-swapping these three floats reroutes behaviour on a microsecond timescale without touching network weights.

In a 20×20 *Pac-Mind* maze and *MiniHack HazardRooms*, high dopamine triples early reward but quadruples collisions, whereas serotonin-heavy settings cut collisions below 1 % with slower returns, tracing a smooth safety–performance frontier.

1.1. Contributions

- (i) An actor–critic whose three global gains let us flip impulse–caution in 13 ms without weight updates.
- (ii) A formal “mood manifold” linking those gains to reward, risk, and exploration.
- (iii) Empirical validation on two grid benchmarks *plus* a SLAM case study, including automatic gain tuning via Bayesian optimisation and an LLM supervisor.

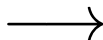
Section 3 formalises the mood manifold, Section 4 details the architecture, Section 5 reports results and ablations, and

What if RL agents could adapt like animals do?



Biology

Neuromodulators change behaviour in **milliseconds** without rewiring synapses



Our Approach

Three external gains change behaviour in **milliseconds** without gradient updates

The Take-Home Message

Fast adaptation and slow learning are **orthogonal** capabilities.
By adding mood knobs, we give RL the best of both worlds.

Appendix A.1 — Serotonergic Mapping Used in the Mood Controller

Revised gain definitions

$$k_{\text{ent}} = \alpha [5\text{-HT}_{2A}] - \beta [5\text{-HT}_{1A}] \qquad k_{\text{risk}} = \gamma [5\text{-HT}_{1A/1B}] - \delta [\text{DA}_{D2}]$$

5-HT_{2A} boosts cortical entropy → wider policy exploration.

5-HT_{1A/1B} (amygdala hippocampus) ↑ threat sensitivity and behavioural inhibition.

Striatal D₂ antagonism tempers impulsive risk, counter-balancing serotonergic caution.

Receptor	Dominant behavioural signatures (rodent/ex vivo)	Refs.
5-HT _{2A}	↑ cortical excitation, ↑ novelty seeking, ↑ response entropy	Nichols 2016; Carhart-Harris 2014
5-HT _{1A} (post-syn.)	↓ anxiety, ↑ approach in EPM / OFT paradigms	Savitz 2020; Lowry 2005
5-HT _{1A} (auto)	Raphe auto-receptor brake; dampens both reward	threat circuits
Richardson 2013		
5-HT _{1B}	Presynaptic inhibition, promotes delayed-reward patience	Nautiyal 2015
D ₂	Striatal no-go bias, ↓ risk-taking	Frank 2004; Bari 2020

Appendix A.2 — Mathematical Details

Reward modification (always on)

$$R_t^{\text{mod}} = R_t - k_{\text{risk}} \rho(s_t)$$

Critic updates during pre-training only

$$V_w(s) \leftarrow V_w(s) + \alpha_c k_{\text{DA}} [R_t^{\text{mod}} + \gamma V_w(s') - V_w(s)]$$

$$\hat{Q}_w(s, a) \leftarrow \hat{Q}_w(s, a) + \alpha_c k_{\text{DA}} [R_t^{\text{mod}} + \gamma \max_{a'} \hat{Q}_w(s', a') - \hat{Q}_w(s, a)]$$

($\alpha_c = 0$ after pre-train; all weights frozen)

Action-selection policy (deployment)

$$\pi_{\mathbf{k}}(a|s) = \frac{\exp(z_a(s) + \eta A_{\mathbf{k}}(s, a)) / \tau}{\sum_{a'} \exp(z_{a'}(s) + \eta A_{\mathbf{k}}(s, a')) / \tau}, \quad \tau = \exp(k_{\text{ent}})$$

$A_{\mathbf{k}}(s, a) = \hat{Q}_w(s, a) - V_w(s)$ is the advantage; η is its scale.

Thank You!

Dario Fumarola

✉ fumadari@amazon.com

&

Jin Tan Ruan

✉ jtanneruan@amazon.com



Customer Satisfaction Survey

This was our first conference! Your feedback helps us improve as aspiring researchers :)