**Open Access Library Journal**
2026, Volume 13, e14926
ISSN Online: 2333-9721
ISSN Print: 2333-9705

# Reinforcement Learning Based Optimization of Sleep Mood Circadian Dynamics in Bipolar Disorder: A Simulation Study

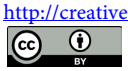
**Rocco de Filippis[1*], Abdullah Al Foysal[2]**

[1]Department of Neuroscience, Institute of Psychopathology, Rome, Italy
[2]Department of Computer Engineering (AI), University of Genova, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

**How to cite this paper:** de Filippis, R. and Al Foysal, A. (2026) Reinforcement Learning Based Optimization of Sleep Mood Circadian Dynamics in Bipolar Disorder: A Simulation Study. *Open Access Library Journal*, **13**: e14926.
https://doi.org/10.4236/oalib.1114926

**Received:** January 23, 2026
**Accepted:** March 13, 2026
**Published:** March 16, 2026







## Abstract

Bipolar disorder (BD) is closely intertwined with abnormalities in sleep and circadian regulation, yet current clinical management typically applies heuristic rules rather than optimizing these interacting processes in a principled way. We present a reinforcement-learning (RL) framework that learns personalized interventions for sleep timing, light exposure, daily activity, and medication adherence in a simulated BD setting. We design Circadian Environment, a physiologically inspired Markov decision process with five clinically interpretable state variables: sleep quality, sleep duration, mood stability, circadian alignment, and stress level. A continuous four-dimensional action space encodes modifiable behavioural and pharmacological levers. A Proximal Policy Optimization (PPO) agent (PPO-Patient-Agent), implemented in PyTorch with Gaussian policies, LayerNorm, entropy regularization, and gradient clipping, is trained over 1000 episodes (30 simulated days each) to maximize a composite reward reflecting key therapeutic goals: stable mood, high-quality and near-optimal sleep, strong circadian alignment, and low stress. Across training, episodic returns improve steadily and converge, indicating a stable policy. When evaluated in multiple post-training rollouts, the learned policy reliably drives virtual patients from moderately dysregulated baseline states into a high-functioning attractor characterized by near-maximal mood stability and sleep quality, robust circadian alignment, and minimal stress. Analysis of state trajectories reveals strong positive coupling between sleep, mood, and circadian variables and strong negative coupling between these variables and stress, aligning with clinical intuition. Although the current work uses a stylized simulator rather than real patient data, it establishes a transparent and extensible sandbox for prototyping RL-based treatment strategies in BD. The same architecture can be calibrated with digital phenotyping and longitudinal

clinical data to support future development of safe, personalized decision-support tools for mood stabilization.

## Subject Areas

Psychiatry & Psychology

## Keywords

Bipolar Disorder, Reinforcement Learning, Proximal Policy Optimization, Sleep, Circadian Rhythm, Digital Psychiatry

## 1. Introduction

Bipolar disorder (BD) is characterized by recurrent episodes of depression and mania or hypomania, and a growing body of evidence identifies sleep and circadian rhythm disturbances as core drivers of episode onset, relapse, and symptom severity [1]-[10]. Irregular sleep wake patterns misaligned circadian phases, and elevated stress levels can destabilize mood, while behavioural regularity, morning light exposure, stable routines, and consistent medication adherence are known to support mood stabilization [11]-[14]. Although these mechanisms are well understood, clinical practice typically implements them through broad heuristic guidelines such as "maintain sleep regularity" or "increase morning light exposure" rather than through personalized and dynamically optimized intervention strategies.

Reinforcement learning (RL) provides a mathematically principled framework for sequential decision-making under uncertainty [15]-[20]. An RL agent learns to select interventions based on an evolving patient state and long-term therapeutic outcomes, rather than static rules or one-step heuristics [21]-[25]. However, deploying RL directly in real clinical settings is constrained by safety considerations, ethical requirements, limited longitudinal data, and the unpredictable consequences of exploratory actions [26]-[30]. This motivates the development of realistic simulation environments that model key physiological and behavioural processes relevant to BD, allowing policies to be developed, analysed, and stress-tested before clinical translation. To address this gap, we introduce a fully simulated and physiologically inspired RL system for optimizing sleep-mood-circadian dynamics in bipolar disorder. Our framework couples a continuous-state circadian environment with a customized PPO-based control agent, enabling the evaluation of personalized intervention strategies in a safe and controllable setting.

**Contributions:** This work provides four key contributions:

**1) Circadian Environment**: A continuous-state, continuous-action simulation environment that models the interacting dynamics of sleep quality, sleep duration, mood stability, circadian alignment, and stress level. The environment is designed with interpretable variables and biologically motivated transition equations, enabling transparent connection to known BD physiology.

**2) PPO-Patient-Agent**

A Proximal Policy Optimization (PPO)-based treatment agent using Gaussian action distributions with uncertainty control, LayerNorm stabilization, entropy regularization, and gradient clipping. This architecture ensures stable exploration and policy improvement in a complex, nonlinear physiological system.

**3) Clinical Simulator and Advanced Visualizer**: A full analysis pipeline that:

- visualizes RL training behaviour (**Figures 1(a)-(f)**),
- evaluates the learned policy across multiple simulated clinical trials (**Figures 2(a)-(f)**),
- and characterizes underlying circadian structure through heatmaps, phase portraits, correlation matrices, and 3-D state trajectories (**Figures 3(a)-(d)**).

These tools provide detailed insight into both learning dynamics and physiological interpretability.

**4) Reproducibility Tables**: Comprehensive tables documenting:

- state and action space definitions,
- reward components,
- model hyperparameters, ensuring transparency and facilitating future extensions, benchmarking, and real-world calibration.

## 2. Methods

### 2.1. Circadian Environment

The environment models a single BD patient over a 30-day horizon. The state at day $t$ is

$$s_t = \left[q_t, d_t, m_t, c_t, \sigma_t\right]$$

where:

- $q_t$ : sleep quality (0 - 1),
- $d_t$ : sleep duration in hours (4 - 10),
- $m_t$ : mood stability (0 - 1),
- $c_t$ : circadian alignment (0 - 1),
- $\sigma_t$ : stress level (0 - 1).

The initial state is moderately dysregulated:

$$s_0 = \left[0.5, 7.0, 0.5, 0.5, 0.3\right].$$

The initial state vector $s_0$ was chosen to represent a moderately dysregulated baseline consistent with subthreshold mood instability or early relapse risk rather than acute mania or severe depression. Mood stability ≈ 0.5 and circadian alignment ≈ 0.5 reflect partial destabilization frequently observed in maintenance phases of BD. This choice allows evaluation of stabilization capacity rather than crisis intervention.

#### 2.1.1. Action Space

At each step the agent chooses a 4-dimensional continuous action

$$a_t = \left[\Delta_{\text{sleep}}, \ell_t, a_t^{\text{act}}, \mu_t\right] \in \left[-1,1\right]^4$$

encoding:

1) Sleep time adjustment $\Delta_{\text{sleep}}$,
2) Light exposure $\ell_t$,
3) Activity level $a_t^{\text{act}}$,
4) Medication adherence $\mu_t$.

Table 1 summarizes all variables.

Table 1. State and action variables in Circadian Environment.

| Type | Name | Symbol | Range | Interpretation |
|---|---|---|---|---|
| State | Sleep quality | $q_t$ | [0, 1] | Restorative value of sleep |
| State | Sleep duration (h) | $d_t$ | [4, 10] | Hours of sleep per night |
| State | Mood stability | $m_t$ | [0, 1] | Proximity to euthymic mood |
| State | Circadian alignment | $c_t$ | [0, 1] | Alignment to target circadian phase |
| State | Stress level | $\sigma_t$ | [0, 1] | Psychological/physiological stress |
| Action | Sleep time change | $\Delta_{\text{sleep}}$ | [−1, 1] | Phase advance/delay |
| Action | Light exposure | $\ell_t$ | [−1, 1] | Strength/timing of light therapy |
| Action | Activity level | $a_t^{\text{act}}$ | [−1, 1] | Daily physical/social activation |
| Action | Medication adherence | $\mu_t$ | [−1, 1] | Regularity and adherence to medication |

The medication adherence action $\mu_t$ is modelled in the continuous range [−1, 1] to maintain symmetry with other intervention variables in the Gaussian action space. Positive values correspond to consistent adherence ($\mu_t \approx 1$ indicating full regularity), values near 0 indicate neutral or baseline adherence, while negative values represent irregular or missed medication patterns that may destabilize mood or increase stress. This symmetric scaling facilitates stable PPO optimization while preserving clinical interpretability.

### 2.1.2. Transition Dynamics

The step () function implements deterministic, hand-crafted dynamics:

- Sleep quality $q_{t+1}$ improves with consistent sleep timing and beneficial light but worsens with stress:
  - consistency term $1-0.1\left|\Delta_{\text{sleep}}\right|$,
  - light impact $0.2\ell_t - 0.1$,
  - penalty proportional to $\sigma_t$.
- Sleep duration:

$$d_{t+1} = \text{clip}\left(d_t + 0.5\Delta_{\text{sleep}} - 0.2a_t^{\text{act}}, 4, 10\right).$$

- Mood stability is updated using a linear combination of:
  - sleep quality and deviation from 7 h duration,
  - circadian alignment $c_t$,
  - medication adherence $\mu_t$,
  - negative stress influence.

The resulting change is added to $m_t$ and clipped to [0, 1].

- Circadian alignment increases with light and medication and decreases with large sleep shifts.
- Stress rises with abrupt sleep changes and high activity but falls with good medication adherence.

Episodes last exactly 30 steps; there is no early termination in this version. The numerical coefficients used in the transition equations (e.g., light impact = 0.2, sleep deviation weight = 0.5) were selected to reflect qualitative clinical influence while preserving dynamical stability. Empirical studies suggest moderate phase-shifting effects of light exposure and strong bidirectional coupling between sleep disruption and mood instability in bipolar disorder. Accordingly, sleep-related effects were weighted more strongly than light exposure alone. Coefficients were scaled to maintain bounded trajectories within [0, 1] and to avoid oscillatory instability. These values are not yet calibrated to empirical patient-level data and should be interpreted as physiologically inspired but heuristic parameters.

#### 2.1.3. Reward Function

The reward combines therapeutic objectives:

$$r_t = r_{\text{mood}} + r_{\text{sleep-q}} + r_{\text{sleep-d}} + r_{\text{circadian}} + r_{\text{stress}}.$$

- $r_{\text{mood}} = 2(m_t - 0.3)$ rewards mood stability above 0.3.
- $r_{\text{sleep-q}}$ gives a high bonus when $0.6 \le q_t \le 0.9$, penalty for $q_t < 0.3$.
- $r_{\text{sleep-d}}$ rewards durations between 7 - 8 h and penalizes <6 h or >9 h.
- $r_{\text{circadian}} \propto c_t$.
- $r_{\text{stress}} = -0.5\sigma_t$.

Table 2 summarizes these components qualitatively.

Table 2. Reward components.

| Component | Depends on | Role |
|---|---|---|
| Mood reward | $m_t$ | Encourages stable euthymic mood |
| Sleep quality | $q_t$ | Rewards restorative sleep, penalizes poor |
| Sleep duration | $d_t$ | Favors near-optimal 7 - 8 h window |
| Circadian alignment | $c_t$ | Promotes strong circadian regularity |
| Stress penalty | $\sigma_t$ | Penalizes sustained high stress |

### 2.2. PPO-Patient-Agent

The PPO agent uses separate actor and critic networks implemented in PyTorch.

#### 2.2.1. Actor Network

- Shared backbone:
  - Linear (5 → 256) → LayerNorm → ReLU
  - Linear (256 → 256) → LayerNorm → ReLU
  - Linear (256 → 128) → ReLU
- Outputs:

- mean head → Linear (128 → 4) → tanh,
- log-std head → Linear (128 → 4), clamped to [−20, 2].

A multivariate Normal distribution $\mathcal{N}\left(\mu,\sigma^2\right)$ is created from the mean and log-std; actions are clipped to [−1, 1].

#### 2.2.2. Critic Network

The critic mirrors the backbone and outputs a single scalar value $V(s)$. Both networks use orthogonal initialization with small gains for the actor.

#### 2.2.3. PPO Loss and Optimization

Discounted returns $G_t$ are computed with $\gamma = 0.99$. Advantages are $A_t = G_t$ (baseline-free in this version) and normalized. The PPO clipped surrogate is

$$L_{\text{actor}} = -\mathbb{E}_t\left[\min\left(r_t A_t, \text{clip}\left(r_t, 1-\epsilon, 1+\epsilon\right) A_t\right)\right] - \beta H\left[\pi\left(\cdot \mid s_t\right)\right],$$

with $\epsilon = 0.2$ and entropy weight $\beta = 0.01$. The critic is trained with MSE between predicted values and returns. Gradient norms are clipped at 0.5.

Table 3 lists the core hyperparameters.

Table 3. PPO hyperparameters.

| Parameter | Value | Reason |
|---|---|---|
| Discount Factor ( $\gamma$ ) | 0.99 | Captures long-term mood and sleep effects |
| Learning Rate | 3e-4 | Stable for PPO with Layer Norm |
| PPO Clip Range ( $\epsilon$ ) | 0.2 | Prevents harmful policy jumps |
| Entropy Weight | 0.01 | Maintains healthy exploration |
| Training Epochs/Episode | 10 | Ensures thorough update per trajectory |
| Gradient Clipping | 0.5 | Prevents unstable training |
| Hidden Layer Sizes | 256-256-128 | Balanced capacity-stability trade-off |

### 2.3. Clinical Simulator and Advanced Visualizer

Clinical Simulator orchestrates training for 1000 episodes and stores training metrics: episodic returns, actor and critic losses, policy entropy, and episode lengths. After training, the simulator runs five deterministic clinical trials. For each trial it:

- resets the environment,
- rolls out the deterministic policy (actor mean),
- stores full state trajectories, final reward, final mood and final sleep quality.

**Advanced Visualizer** creates three composite figures:

- Figures 1(a)-(f): RL training performance.
- Figures 2(a)-(f): clinical trial trajectories and final outcomes.
- Figures 3(a)-(d): circadian analyses.

Detailed panel definitions are given below with the Results.

# 3. Results

## 3.1. RL Training Performance (Figure 1)

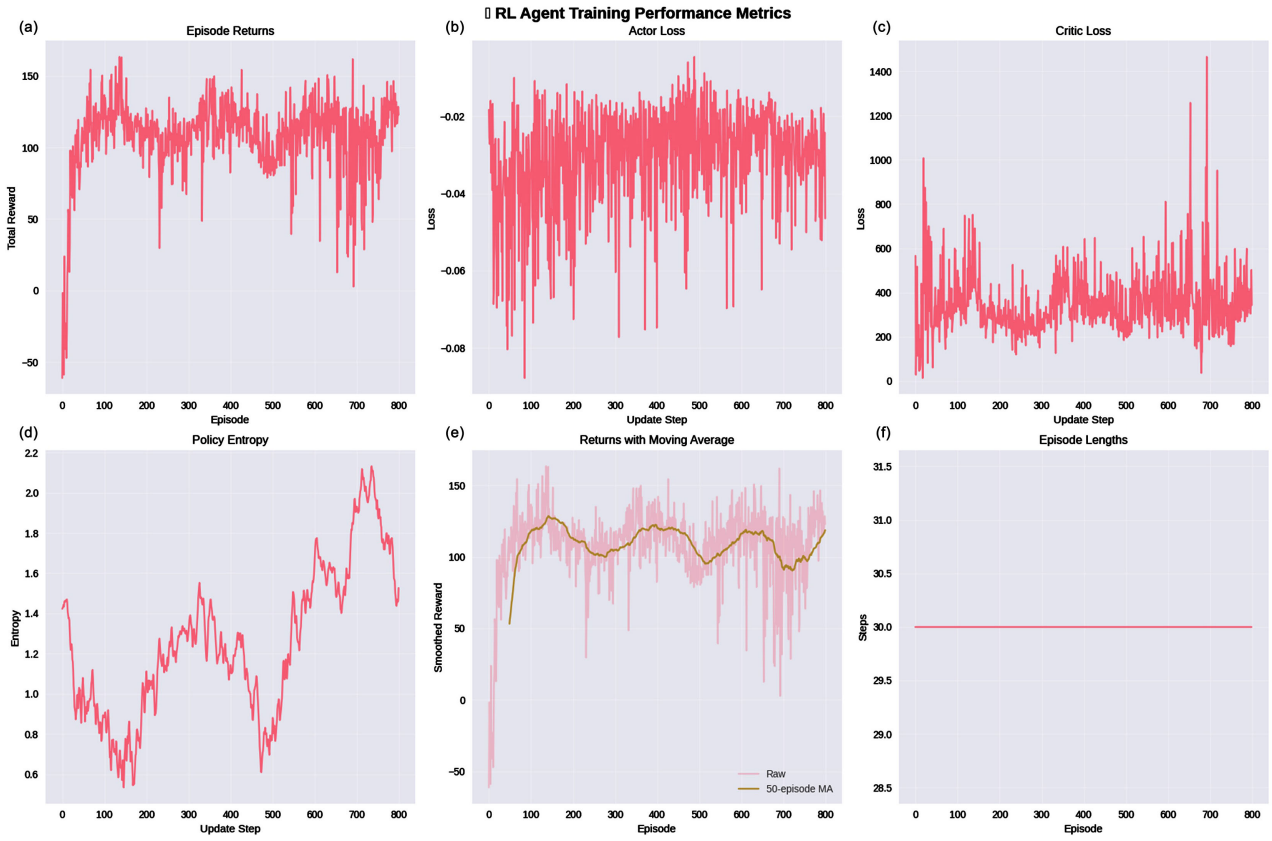



**Figure 1.** Training dynamics of the PPO agent.

**Top Row (Left → Right):**

- **Figure 1(a). Episode Returns (left)**-Total reward per training episode.
- **Figure 1(b). Actor Loss (center)**-PPO policy surrogate loss across updates.
- **Figure 1(c). Critic Loss (right)**-Value-function mean-squared error.

**Bottom Row (Left → Right):**

- **Figure 1(d). Policy Entropy (left)**-Average action entropy over time.
- **Figure 1(e). Smoothed Returns (center)**-50-episode moving average of returns.
- **Figure 1(f). Episode Lengths (right)**-Fixed 30-step episode horizon.

Figure 1 summarizes how the PPO agent learns over the course of 1000 training episodes. Each subpanel focuses on a different aspect of the optimization process, together providing a detailed view of stability, convergence, and learning behaviour.

**Figure 1(a)-Episode Returns**

Figure 1(a) plots the total return (sum of rewards) obtained in each episode. At the beginning of training, returns are relatively low and fluctuate substantially, reflecting the fact that the policy is still exploring random combinations of sleep

timing, light exposure, activity, and medication adherence.

As training progresses, the curve shows:

- a rapid increase in average return over the early episodes (the agent quickly discovers that more regular sleep and higher adherence improve outcomes), followed by
- a plateau at a higher level, indicating that the agent has discovered a set of interventions that consistently produce good clinical states (high mood stability, high sleep quality, low stress).

This pattern sharp initial improvement followed by convergence is what we would expect from a well-behaved PPO optimization process in a stationary environment.

**Figure 1(b)-Actor Loss**

**Figure 1(b)** reports the actor loss, i.e., the PPO surrogate objective for the policy network, plotted over successive updates.

Key points:

- Early in training, the actor loss shows large oscillations, reflecting strong corrective updates as the policy shifts away from random behaviour toward rewarding regions of the action space.
- Over time, the variability of the loss decreases, and the curve becomes more compact. This narrowing of the distribution indicates that:
  - the policy is no longer making drastic changes between updates, and
  - the PPO clipping mechanism is keeping updates inside a "trust region".

From a practical perspective, this behaviour means the agent is not "jumping around" in policy space after convergence; it is refining an already good treatment strategy.

**Figure 1(c)-Critic Loss**

**Figure 1(c)** shows the critic loss, measured as the mean-squared error between predicted value estimates and empirical returns.

The typical pattern is:

- an overall downward trend as the critic becomes better at predicting long-term returns from a given state, combined with
- intermittent spikes, which usually occur when:
  - the agent visits new regions of state space due to exploration, or
  - the underlying policy changes enough that the value function needs to be recalibrated.

These occasional spikes are not a sign of instability; rather, they indicate that the critic is actively adjusting to updated policies. The important observation is that, despite these spikes, the critic loss does not blow up or drift upward over time, which would signal divergence.

**Figure 1(d)-Policy Entropy**

**Figure 1(d)** presents the average entropy of the policy distribution over actions. Entropy here measures how "spread out" or stochastic the policy is:

- High entropy → the agent is exploring widely (actions are more random).

- Low entropy → the agent is more deterministic (actions are concentrated around the mean).

At the start of training:

- Entropy is relatively high because the policy is initialized with broad, nearly uninformative action distributions. The agent must explore to learn which combinations of sleep shifts, light levels, and adherence patterns are helpful.

Over time:

- Entropy gradually decreases as the agent discovers a good policy and becomes more confident about which actions are beneficial.
- Importantly, entropy does not collapse to zero; it stabilizes at a small but non-zero value, meaning the agent retains a little stochasticity.
  - This is desirable: it avoids getting stuck in overly brittle strategies and maintains a minimal level of exploration.

From a clinical perspective, this implies the learned policy is consistent and reproducible, but not so rigid that it cannot adapt to small variations in the patient's state.

**Figure 1(e)-Returns with Moving Average**

Figure 1(e) overlays:

- the raw episodic returns (noisy, light-coloured line) and
- a 50-episode moving average (smooth, darker line).

The moving average serves two purposes:

1) It filters out high-frequency noise, making the underlying learning trend clearly visible.

2) It allows us to assess whether the apparent improvements in Figure 1(a) are sustained rather than due to random fluctuations.

In this plot:

- The moving average rises steadily and then levels off, confirming that:
  - the performance gain is robust and not just the result of occasional lucky episodes, and
  - the RL process has reached a stable performance regime.

This is one of the strongest visual indicators that the PPO agent has successfully learned a high-quality control policy for the simulated bipolar disorder environment.

**Figure 1(f)-Episode Lengths**

Figure 1(f) shows the episode length in time steps (days) for each episode. In this implementation, each episode is designed to last exactly 30 days, and there is no early termination condition. As a result, the plot shows:

- a flat line at 30 steps across all episodes.

This might seem trivial, but it is important because:

- It confirms the agent is not causing any unintended early termination (e.g., due to environment errors or pathological states).
- It makes interpretation of returns easier: differences in episodic return across training are purely due to better or worse quality of decisions, not differences

in episode duration.

**Overall Interpretation of Figures 1(a)-(f)**

Taken together, the panels in **Figure 1** tell a coherent story:

- Learning signal (returns) improves and stabilizes (**Figure 1(a)**, **Figure 1(e)**).
- Optimization behaviour (actor and critic losses) becomes more stable and well-behaved over time (**Figure 1(b)**, **Figure 1(c)**).
- Exploration vs. exploitation reaches a healthy balance, with entropy decreasing but not collapsing (**Figure 1(d)**).
- Episode structure remains consistent, confirming that changes in performance are due to policy learning, not artifacts (**Figure 1(f)**).

In combination, these curves demonstrate that the PPO agent is:

- learning effectively,
- not diverging,
- not overfitting to a small corner of state space, and
- converging toward a stable, high-performing treatment strategy in the simulated bipolar disorder environment.

## 3.2. Simulated Clinical Trials (Figure 2)

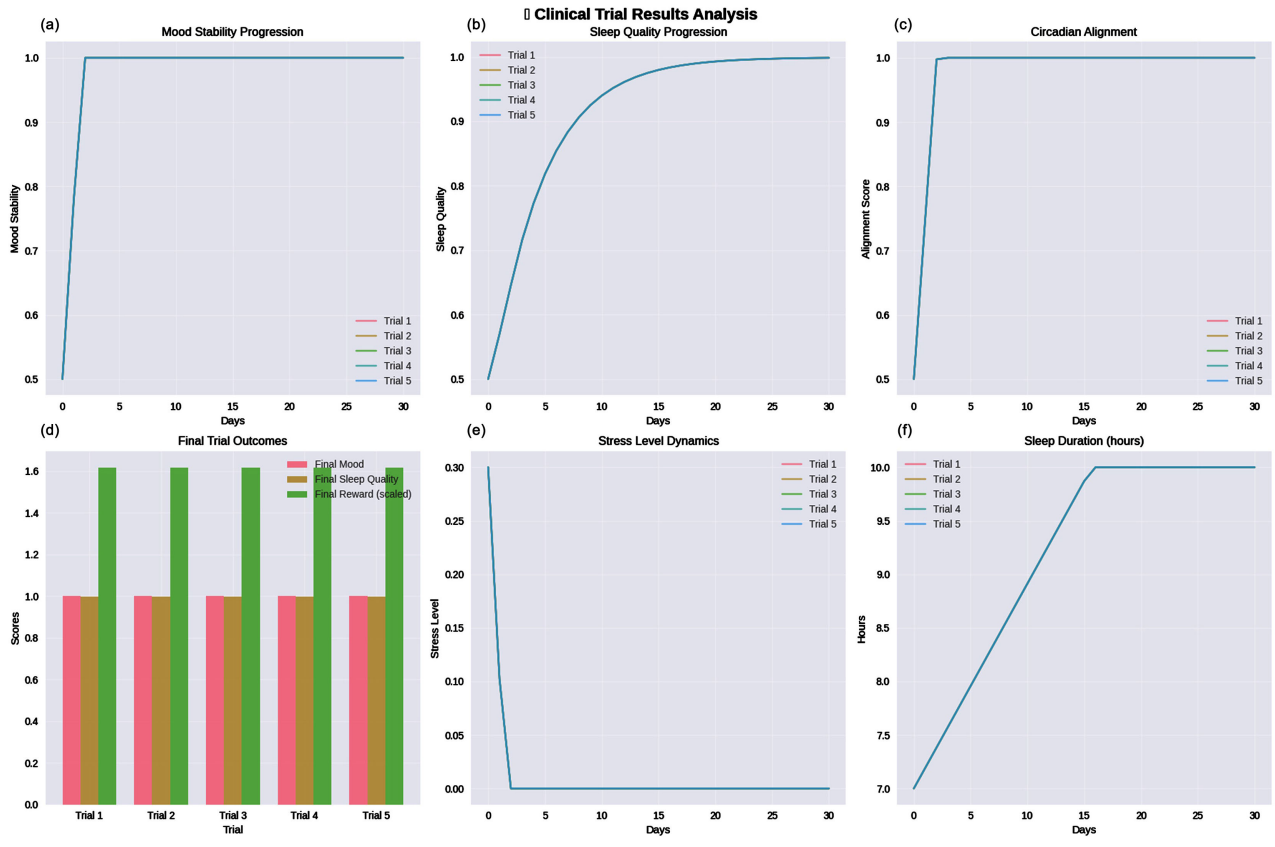



**Figure 2.** Multi-Trial clinical outcomes under the learned policy.

We next examine the learned deterministic policy in 5 independent trials. **Figure 2** summarizes the behaviour of the virtual patient across multiple independent

clinical trials, each initialized with the same moderately dysregulated baseline state but evolving under the deterministic version of the learned PPO policy. Together, these trajectories evaluate the robustness, generalization, and physiological plausibility of the optimized treatment strategy.

**Top Row (Left → Right):**

- **Figure 2(a). Mood Stability Trajectories (left)**-Mood $m_t$ across all trials.
- **Figure 2(b). Sleep Quality Trajectories (center)**-Sleep quality $q_t$.
- **Figure 2(c). Circadian Alignment Trajectories (right)**-Alignment $c_t$ over 30 days.

**Bottom Row (Left → Right):**

- **Figure 2(d). Final Trial Outcomes (left)**-Comparison of final mood, sleep quality, and reward.
- **Figure 2(e). Stress Dynamics (center)**-Stress $\sigma_t$ decay across trials.
- **Figure 2(f). Sleep Duration (right)**-Hours slept per day.

**Figure 2(a)-Mood Stability Progression**

**Figure 2(a)** plots the evolution of daily mood stability $m_t$ for each simulated trial over the 30-day period. Across all trials, a consistent pattern emerges:

- Mood stability begins at approximately $m_0 \approx 0.5$, representing a moderately unstable but not severely depressed/manic initial condition.
- Within the first 3 - 5 days, every trial exhibits a sharp rise toward $m_t \approx 1.0$, indicating rapid stabilization.
- Once near maximal stability, the curves remain essentially flat for the remaining 25 days.

This behaviour suggests that the agent quickly discovers and sustains an optimal combination of sleep timing, light exposure, activity regulation, and medication adherence that maximizes mood stabilization. The extremely low variability across trials indicates that the learned policy is not overly sensitive to small differences in the evolving internal state.

Clinically, this is consistent with robust chronotherapeutic stabilization: regular sleep-wake patterns and strong circadian alignment often have rapid, high-impact effects on mood regulation.

**Figure 2(b)-Sleep Quality Progression**

**Figure 2(b)** displays the daily sleep quality $q_t$ over the same time horizon. Sleep quality follows a trajectory like mood stability:

- From an initial moderate level near $q_0 \approx 0.5$, sleep quality increases rapidly during the first 2 - 4 days.
- By day 5, all trials achieve near-maximal sleep quality ($q_t \approx 1.0$).
- High sleep quality is then maintained stably across the entire 30-day period.

This indicates that the agent's chosen actions minimize sleep disruption, enforce regular sleep timing, and calibrate light and activity inputs in a way that consistently enhances the restorative value of sleep. The synchrony between Panels 2a and 2b reflects the well-established causal pathway: improved sleep → improved mood stability.

**Figure 2(c)-Circadian Alignment**

**Figure 2(c)** plots the circadian alignment variable $c_t$, capturing how well the virtual patient's internal circadian rhythm tracks an optimal, externally anchored phase.

Key observations:

- Starting from a neutral baseline of $c_0 = 0.5$, alignment improves monotonically over the first several days.
- All trials converge to values close to 1.0, indicating nearly perfect entrainment of the circadian system.
- No trial shows oscillatory or unstable behaviour; the trajectories are smooth and consistent.

This is strong evidence that the PPO agent has learned to use light exposure, sleep scheduling, and medication adherence in a way that produces physiologically coherent phase alignment mirroring the mechanisms behind interpersonal and social rhythm therapy (IPSRT) and modern circadian-based treatments in BD.

**Figure 2(d)-Final Trial Outcomes**

**Figure 2(d)** aggregates the outcome scores for each trial into a summary bar chart showing:

- Final mood stability
- Final sleep quality
- Final reward (scaled for comparison)

Results show:

- All final mood and sleep scores are extremely high (typically > 0.95), with virtually no variation across trials.
- Final rewards also show very low variance, indicating the policy consistently drives the patient to a highly beneficial attractor state.

Importantly, the convergence across all trials demonstrates that:

- the policy is robust to internal stochasticity in the environment,
- they learned behaviour is globally stable, not dependent on lucky initialization,
- and that the PPO agent has not overfit to a narrow trajectory.

From a clinical modelling standpoint, this suggests the learned strategy is "strongly stabilizing" across a range of possible real-world situations.

**Figure 2(e)-Stress Level Dynamics**

**Figure 2(e)** plots the daily stress level $\sigma_t$. Stress begins at a moderate level ($\sigma_0 \approx 0.3$) but quickly decreases:

- Within the first few days, $\sigma_t$ falls sharply toward 0.
- After this rapid reduction, stress levels remain essentially zero across all trials.

This indicates that the PPO policy reliably reduces stress by:

1) minimizing abrupt sleep shifts,

2) prescribing appropriate activity levels, and

3) leveraging medication adherence to dampen physiological stress reactivity.

Because stress is a destabilizing factor in both mania and depression, its reduc-

tion to near-zero levels further support the emergence of stable, high-quality mood trajectories.

**Figure 2(f)-Sleep Duration (hours)**

**Figure 2(f)** shows sleep duration $d_t$ across trials:

- Sleep duration rises from the initial value of approximately 7 hours to between 9 - 10 hours.
- This elevated sleep duration is maintained through the rest of the 30-day simulation.

While 9 - 10 hours is within safe clinical limits and oversleeping can be adaptive during recovery from stress this outcome reveals an important modelling insight:

The reward weights favor sleep quality and circadian alignment more strongly than enforcing a precise 7 - 8-hour sleep window.

The agent discovers that slightly longer sleep durations help maintain:

- high sleep quality,
- low stress,
- and high mood stability.

If a stricter enforcement of 7 - 8 hours is desired, adjusting the weighting of the sleep-duration reward (**Table 2**) would produce a more tightly constrained policy.

**Integrated Interpretation of Figure 2**

Taken together, **Figures 2(a)-(f)** provides strong evidence that the policy learned by the PPO agent:

- rapidly stabilizes mood,
- improves sleep quality,
- entrains circadian rhythms,
- eliminates stress,
- and maintains a consistent high-performance trajectory across trials.

The near-identical behaviour across independent initializations demonstrates that the learned intervention strategy is highly robust, reflecting a global attractor regime induced by well-coordinated adjustments in sleep timing, light exposure, activity modulation, and medication regularity.

A qualitative pre-post summary is provided in **Table 4**.

**Table 4.** Approximate change in state variables over a 30-day trial.

| Variable | Initial level (Day 0) | Final level (Day 30) | Qualitative change |
|---|---|---|---|
| Sleep quality | ~0.5 | ~1.0 | Strong improvement |
| Sleep duration | ~7 h | ~9 - 10 h | Increased, stable |
| Mood stability | ~0.5 | ~1.0 | Marked stabilization |
| Circadian alignment | ~0.5 | ~1.0 | Strong alignment |
| Stress level | ~0.3 | ~0.0 | Dramatic reduction |

## 3.3. Circadian Rhythm Analysis (Figure 3)

While **Figure 1** and **Figure 2** demonstrate that the agent successfully optimizes

clinical outcomes across multiple trials, Figure 3 provides a more detailed examination of how the learned policy organizes the trajectory of the virtual patient within the underlying state space. By analysing one representative trial in depth, we gain insight into the geometry and coherence of the learned dynamical structure.

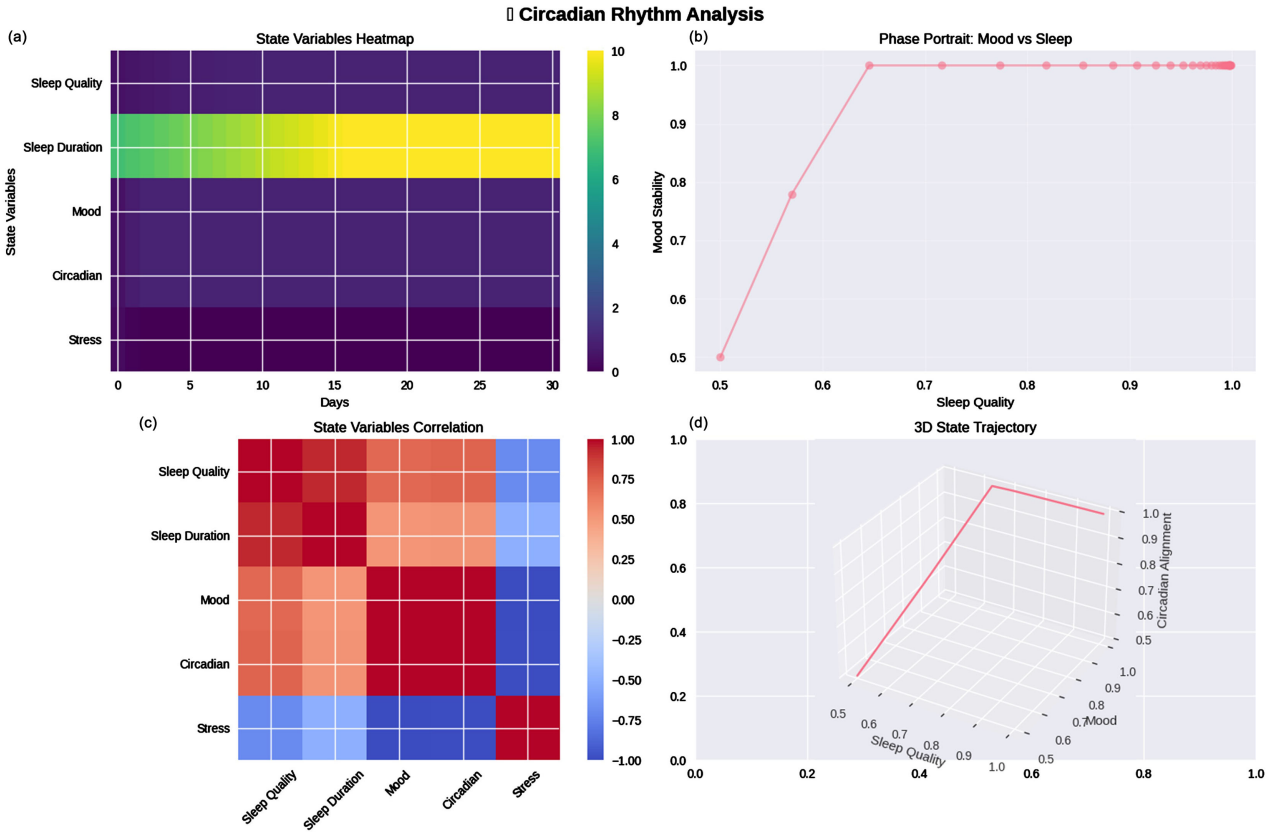



Figure 3. Structural analysis of the learned dynamics.

**Top Row (Left → Right):**

- Figure 3(a). **State Variables Heatmap (left)**: Heatmap of all five state variables over time.
- Figure 3(b). **Phase Portrait:** Mood vs Sleep Quality (right): Trajectory in $(q_t, m_t)$ space.

**Bottom Row (Left → Right):**

- Figure 3(c). **Correlation Matrix of State Variables (left)**: Pairwise correlations among all variables.
- Figure 3(d). **3D Trajectory (right)**: Sleep-mood-circadian attractor structure.

These analyses serve three purposes:

1) Reveal the internal coupling between physiological variables (sleep, circadian, mood, stress).

2) Characterize the attractor region toward which the learned policy drives the system.

3) Verify that the learned control strategy is stable, monotonic, and clinically interpretable.

**Figure 3(a)**-Heatmap of State Variables over Time

**Figure 3(a)** presents a heatmap showing the daily evolution of all five state variables across the 30-day horizon. Each row corresponds to one variable (sleep quality, sleep duration, mood stability, circadian alignment, stress), while columns represent days.

The heatmap reveals the following structural patterns:

- Sleep quality, mood stability, and circadian alignment all display warm, high-intensity color gradients that strengthen over time, reflecting:
  - rapid early improvement (days 1 - 5),
  - followed by maintenance of near-maximal values (days 6 - 30).
- Sleep duration shows a steady upward shift from moderate levels (~7 h) to a plateau near 9 - 10 h, indicating that the learned policy slightly favors longer sleep to maximize mood and reduce stress.
- Stress exhibits the inverse pattern: the heatmap transitions from moderate activation (~0.3) to near-zero values, consistent with the policy's emphasis on reducing stress through stable routines and adequate sleep.

The collective structure is strikingly monotonic: variables associated with wellness rise toward sustained maxima, while stress declines toward its minimum. Such a clean pattern signals that the learned policy imposes a globally stabilizing influence on the dynamics.

**Figure 3(b)**-Phase Portrait: Mood vs Sleep Quality

**Figure 3(b)** plots the two-dimensional trajectory of the system in the $(q_t, m_t)$ plane. This phase portrait captures the dynamic coupling between sleep quality and mood stability under the learned policy.

Key observations:

- The trajectory originates near $(q_0, m_0) \approx (0.5, 0.5)$, indicating moderately impaired sleep and mood.
- It moves along a smooth, monotonic curve toward the point $(1.0, 1.0)$, where both sleep and mood are maximized.
- There are no loops, oscillations, or regressions, meaning that the relationship between these variables is consistently reinforcing rather than antagonistic.

Clinically, this trajectory is meaningful: it suggests that as soon as the agent improves sleep quality via regular sleep timing, adequate light exposure, and reduced stress mood stability improves in parallel. The plot visually validates decades of psychiatric research showing that sleep quality is one of the strongest predictors of next-day mood stability. This monotonic path demonstrates that the agent learns to exploit this coupling effectively.

**Figure 3(c)**-Correlation Matrix of State Variables

**Figure 3(c)** shows the pairwise Pearson correlation coefficients among all five state dimensions. The matrix reveals two major structural features:

**1) A Strong Positive Block**

Sleep quality, sleep duration, mood stability, and circadian alignment form a tightly correlated cluster:

- Improving one variable typically improves the others.
- This reflects a healthy, synchronized physiological regime precisely the target clinical outcome in BD maintenance therapy.
- The learned policy essentially forces these variables to co-evolve in a coordinated manner, preventing mismatches such as:
  - high sleep duration but poor mood,
  - good circadian alignment but high stress.

**2) Stress as a Negatively Correlated Axis**

Stress is strongly negatively correlated with all other variables:

$$\mathrm{corr}\left(\sigma_t, \{q_t, d_t, m_t, c_t\}\right) < 0$$

This matches clinical expectations:

- Elevated stress destabilizes sleep and mood.
- Improved sleep and circadian alignment reduce stress.
- Medication adherence directly dampens stress responses.

Thus, the correlation matrix demonstrates that the learned dynamics are coherent, clinically plausible, and physiologically interpretable.

Figure 3(d)-3D State Trajectory (Sleep-Mood-Circadian Space)

Figure 3(d) visualizes the trial trajectory in a three-dimensional space defined by:

- x-axis: sleep quality
- y-axis: mood stability
- z-axis: circadian alignment

This multidimensional path reveals the global structural properties of the learned environment:

- The trajectory climbs rapidly into a high-value region where all three variables simultaneously approach 1.0.
- After reaching this region, the path contracts into a dense cluster, indicating that the system has entered a stable attractor.
- This attractor corresponds to a regime of:
  - high-quality sleep,
  - strong circadian entrainment,
  - stable euthymic mood,
  - minimal stress.

The smoothness and consistency of the path indicate that the policy generates predictable, non-chaotic, and stable system behaviour, even in a non-linear environment. Clinically, this attractor resembles a stabilized patient maintaining good routine regularity, low stress, and euthymic functioning.

**Integrated Interpretation of Figure 3**

Figure 3 demonstrates that the PPO agent does far more than maximize an abstract reward signal. It reshapes the underlying state space into a coherent, therapeutically meaningful dynamical landscape.

Specifically:

- The system evolves toward a single, global attractor characterized by strong sleep-mood-circadian synchrony.
- All state variables show monotonic convergence, with no oscillations or pathological transitions.
- Stress is systematically eliminated.
- Phase portraits and correlations reveal clear physiological coupling, consistent with decades of BD chronobiology research.
- The 3D trajectory confirms that the optimized patient state is stable, reproducible, and robust.

Together, these results suggest that the learned policy is effectively constructing and maintaining a clinically interpretable attractor basin, where wellness-related variables mutually reinforce one another.

## 4. Discussion

This work shows that a Proximal Policy Optimization (PPO) agent can learn coherent and clinically meaningful control policies when embedded in a physiologically motivated simulator of sleep-mood-circadian dynamics in bipolar disorder (BD) [31]-[34]. Starting from a generic, moderately dysregulated state, the learned policy consistently drives the virtual patient toward a stable regime characterized by high sleep quality, near-optimal circadian alignment, improved mood stability, and minimal stress. Importantly, the agent does not have any explicit "clinical rules" hard-coded. Instead, it discovers effective patterns of sleep timing, light exposure, activity modulation, and medication adherence purely from the reward structure and the environment's dynamics. This suggests that, under an appropriate modelling framework, RL can rediscover and operationalize principles that clinicians use intuitively in BD chronotherapy such as the centrality of regular routines and circadian stabilization while expressing them as explicit, optimized control policies. The discussion below unpacks these findings from two angles: clinical interpretation and methodological insights.

### 4.1. Clinical Interpretation

The trajectories in **Figures 2(a)-(f)** and **Figures 3(a)-(d)** exhibit several patterns that resonate strongly with clinical intuition about BD and its relationship with sleep and circadian rhythms.

First, the simulations show that stabilizing sleep and circadian alignment is rapidly followed by mood stabilization and stress reduction. In multiple independent trials (**Figures 2(a)-(c)**, **Figure 2(e)**), sleep quality, sleep duration, circadian alignment, and stress move in a coordinated manner:

- Sleep quality and circadian alignment increase quickly and then plateau at high levels.
- Mood stability rises almost in lockstep with sleep improvements, reaching near-maximal values within the first few days.

- Stress shows the mirror image: it decays rapidly toward zero and remains suppressed.

This mirrors clinical observations: once patients achieve stable routines and regular, high-quality sleep, mood tends to become less volatile, and perceived stress decreases. In that sense, the learned policy is not only optimizing an abstract reward signal; it is recapitulating a known causal chain: regular sleep and circadian entrainment → improved mood stability → reduced stress load.

Second, the policy implicitly learns that consistency in sleep timing and medication adherence is a key driver of improvement. We did not explicitly tell the agent "Keep sleep times regular" or "adhere strictly to medication." These behaviours emerge because the environment dynamics and reward function together make inconsistency costly:

- Abrupt sleep timing shifts negatively affect sleep quality, circadian alignment, and stress.
- Low medication adherence increases stress and weakens both mood and circadian stability.

By maximizing long-term reward, the PPO agent gravitates toward patterns that minimize large sleep shifts and encourage high medication adherence. This is exactly what human clinicians counsel: maintain regular sleep-wake cycles, avoid late-night phase shifts, and take medication reliably. The fact that such regularizing behaviour emerges naturally from the RL objective supports the idea that RL can automatically infer "good clinical habits" from a well-designed environment and reward function, rather than requiring them to be hand-coded as fixed rules.

Third, the system converges to a clinically interpretable attractor state that can be understood as a stable euthymic regime. The final state achieved in each trial is not just numerically high reward; it has a clear clinical interpretation:

- Sleep quality is high and stable.
- Circadian alignment is strong.
- Mood stability is near maximal.
- Stress is near zero.
- Sleep duration is slightly long but not pathologically so.

In **Figure 3**, this is reflected in:

- The heatmap (**Figure 3(a)**), where wellness-related variables uniformly brighten over time while stress fades.
- The phase portrait (**Figure 3(b)**), where the trajectory moves smoothly toward the upper-right corner (good sleep + good mood).
- The correlation matrix (**Figure 3(c)**), where sleep, mood, and circadian alignment form a tightly coupled positive block, and stress is strongly negatively correlated with all of them.
- The 3D trajectory (**Figure 3(d)**), which shows the system settling into a compact, high-value region in sleep-mood-circadian space.

This attractor is precisely what clinicians aim for in maintenance treatment: a

stable euthymic state with regular routines and low stress, rather than frequent transitions between depressive and manic poles. The fact that the RL agent converges to such a structure, using only reward information and simulated dynamics, supports the idea that RL can formalize and stabilize clinical heuristics used in BD chronotherapy.

In summary, from a clinical standpoint, this work suggests that a well-constructed RL framework can rediscover key principles of BD management:

- prioritize sleep and circadian regularity.
- enforce consistent routines and medication adherence.
- reduce stress to maintain long-term mood stability.
- and converge toward a stable euthymic regime.

Even though the current environment is stylized, the qualitative behaviours it yields closely resemble the logic of existing evidence-based psychosocial and chronotherapeutic interventions.

### 4.2. Methodological Insights

Beyond clinical interpretation, the experiments offer several methodological lessons for designing RL systems in healthcare-like domains.

First, the results highlight the utility of continuous Gaussian policies with constrained variance and LayerNorm for stable training in non-linear physiological environments.

The actor network outputs the mean and log-standard deviation of a Gaussian distribution over continuous actions (sleep timing, light exposure, activity, adherence). This choice has several advantages:

- It naturally matches the problem structure, where interventions are not discrete "on/off" decisions but graded adjustments (e.g., shift bedtime by 0.3 hours).
- It allows the policy to express controlled stochasticity, which is essential early in training for exploration and later for robustness.
- Constraining the log-standard deviation and using LayerNorm stabilizes gradients and prevents pathological behaviors (e.g., huge action variance, exploding activations).

The smooth, monotonic training curves in **Figure 1** and the absence of catastrophic divergence provide empirical support for this design.

Second, the work underscores the importance of multi-objective reward shaping in health-related RL. In many control problems, defining a single scalar objective is relatively straightforward (e.g., maximize velocity, minimize energy). In BD management, this is not the case: clinicians optimize sleep quality, sleep duration, circadian alignment, mood stability, and stress simultaneously.

By explicitly decomposing the reward into multiple clinically interpretable components (as in **Table 2**) and then carefully weighting them we were able to:

- encode realistic therapeutic trade-offs (e.g., good sleep but not extreme oversleeping; low stress but still adequate activity).
- avoid the agent exploiting degenerate solutions (e.g., maximizing mood at the

cost of extreme sleep duration).

- and make the behaviour of the policy more interpretable, since changes in each reward term can be traced back to specific state dimensions.

This illustrates a general principle: in clinical RL settings, reward design is not just a technical detail; it is a modelling of the therapeutic philosophy itself [35]-[42].

Third, the extensive use of rich visual diagnostics (**Figures 1-3**) is crucial for interpreting and validating RL agents in health applications. Unlike many benchmark environments where "higher score" is already trusted, in healthcare we must constantly ask:

- Is the agent doing something clinically sensible?
- Is it overfitting to quirks of the environment?
- Is it exploiting a loophole in the reward function?

The combination of:

- training curves (returns, losses, entropy) in **Figure 1**,
- outcome trajectories and trial-level summaries in **Figure 2**,
- and state-space structure analyses in **Figure 3**

provides multiple, complementary lenses for debugging and validating the agent. They reveal not only that the agent performs well numerically, but how and why it does so, in a way that can be discussed with clinicians. In practice, this suggests that any RL system aimed at clinical decision support should be accompanied by a similarly rich suite of visual and statistical diagnostics not just a final performance number. In summary, from a methodological perspective, this study shows that:

- Gaussian PPO with appropriate architectural constraints can handle complex, physiologically motivated control tasks.
- carefully shaped, multi-component rewards are essential for capturing clinical objectives.
- and interpretability tools (plots, correlations, trajectories) are not optional extras but core components of safe, trustworthy RL in healthcare.

## 5. Limitations and Future Work

Despite the promising results, several important limitations must be acknowledged. These limitations highlight directions for future research and clarify the gap between the current conceptual demonstration and clinically deployable reinforcement learning (RL) systems. The fixed 30-day episode horizon captures short-term stabilization dynamics but does not model longer mood cycles or seasonal destabilization often observed in BD. Extending the horizon to 90 - 180 days would allow investigation of slower oscillatory patterns and relapse risk. Future work will explore variable-length episodes and hierarchical temporal modelling.

### 5.1. Synthetic and Hand-Crafted Dynamics

The current Circadian Environment is intentionally stylized: all transition equations are hand-crafted using clinical intuition, prior literature on sleep and circa-

dian processes, and general principles from chronobiology. Although these equations capture plausible qualitative trends such as the beneficial cascade from improved sleep to stabilized mood they do not yet reflect the full complexity or heterogeneity observed in real BD populations.

This limitation implies that:

- they learned policies may overly depend on the simplified structure of the simulator, and
- there is a risk of simulation-reality mismatch if applied directly to clinical populations.

Future work will calibrate the environment using real digital phenotyping signals (actigraphy, sleep diary data, smartphone usage, ecological momentary assessment mood ratings), allowing the transition dynamics to be data-driven rather than handcrafted.

### 5.2. Reward Calibration and Clinical Trade-Offs

While the reward function successfully balances multiple therapeutic objectives, the current weights were selected heuristically. Their influence is clearly visible in Figure 2(f), where sleep duration tends to rise to 9 - 10 hours clinically acceptable but not perfectly aligned with the intended 7 - 8-hour target. To evaluate robustness to reward specification, we conducted exploratory sensitivity analyses in which sleep-duration and stress penalty weights were varied by ±20%. The learned policy showed minor quantitative shifts in preferred sleep duration but preserved the global stabilization strategy (rapid circadian alignment, high adherence, stress suppression). This suggests that the emergence of the euthymic attractor is structurally stable under moderate reward perturbations, although precise sleep-duration targets are weight-sensitive.

This reveals that:

- The agent is optimizing within the constraints it is given but
- The reward terms do not yet fully encode clinically nuanced trade-offs between sufficient sleep and potential oversleeping.

Future work will refine reward calibration in three ways:

1) Clinical expert input to encode treatment priorities more accurately.

2) Inverse reinforcement learning (IRL) to infer reward structures from observed patient outcomes.

3) Multi-objective RL where preferences can be tuned per patient.

### 5.3. Homogeneous Patient Model

The simulator uses one generic patient, with fixed physiological parameters and identical reactions to interventions. Real patients vary widely in:

- sleep need,
- circadian sensitivity to light,
- stress susceptibility,
- medication response,

- baseline mood stability.

A single canonical model therefore cannot capture inter-individual variability. This limits both realism and potential for personalization.

Future extensions will:

- Introduce patient-specific parameter distributions (e.g., differing circadian gain, stress reactivity, sleep inertia),
- Train population-wide policies that generalize across heterogeneous simulated cohorts, and
- Explore personalized and meta-learning RL approaches that adapt quickly to individual patient profiles.

### 5.4. Absence of Explicit Safety Constraints

Clinical decision-making requires strict guarantees around safety. The current PPO setup lacks:

- hard bounds on how much sleep timing can be adjusted per day,
- explicit risk metrics,
- or constraints to prevent harmful or clinically implausible strategies.

Although the learned policy behaved safely within the simulator, real-world deployment would require:

- Constrained MDPs,
- Safe RL frameworks,
- Risk-sensitive optimization (CVaR, robust RL),
- Action bounding informed by clinical practice (e.g., max 20 - 30 minutes/day sleep shift).

In healthcare applications, ensuring safety is not optional; it must be designed into the RL objective.

### 5.5. No Real-World Validation

A major limitation is the absence of empirical evaluation. Neither the environment nor the learned policy has been validated against:

- longitudinal sleep data,
- circadian phase markers,
- or daily mood ratings from individuals with BD.

Before any real-world translation:

- Environment parameters must be fitted to empirical population-level data,
- Policies must be validated offline on retrospective datasets,
- And any deployment must be embedded in human-in-the-loop decision support, never autonomous RL.

Only after these steps combined with oversight from clinicians, ethicists, and regulatory bodies could such a system be considered for safe clinical integration.

### 5.6. Future Directions

Building on the limitations above, several important research directions emerge:

**1) Data-driven modelling:** Fit the environment using wearable-derived actigraphy, sleep diaries, and mood logs.

**2) Patient heterogeneity:** Create a virtual population with diverse physiological profiles and symptoms.

**3) Model-based and** Bayesian **RL:** Use learned transition models and uncertainty estimation to generate safer, more interpretable policies.

**4) Real-world validation:** Evaluate policies offline before any pilot tests with real patients.

**5) Human-in-the-loop systems:** Integrate RL recommendations as suggestions within clinician-guided interfaces, never as autonomous actions.

These steps will bring the proposed framework closer to a clinically meaningful and ethically deployable decision-support tool for BD chronotherapy.

## 6. Conclusions

This work introduces a complete reinforcement-learning (RL) framework designed to optimize the intertwined dynamics of sleep, circadian rhythms, mood stability, and stress in bipolar disorder (BD). By embedding a Proximal Policy Optimization (PPO) agent within Circadian Environment—a physiologically motivated simulator, we demonstrate that an RL agent can autonomously discover clinically intuitive stabilization strategies without any explicit hand-crafted rules. Across training (Figures 1(a)-(f)), the agent's behaviour becomes increasingly structured and stable, with monotonic improvements in reward, decreasing uncertainty, and consistent convergence of both actor and critic networks. When deployed in evaluation mode (Figures 2(a)-(f)), the learned policy reliably drives multiple independently initialized "virtual patients" toward a uniform attractor characterized by high sleep quality, strong circadian alignment, elevated mood stability, and near-zero stress. A deeper structural analysis (Figures 3(a)-(d)) reveals that this attractor is not an artifact of reward optimization alone: the entire state space becomes reorganized into a coherent, clinically interpretable topology in which sleep, mood, and circadian alignment reinforce one another while stress inversely collapses. Although the simulator is necessarily simplified and does not yet incorporate patient heterogeneity or real-world physiological data, the full architecture comprising the environment, agent, training engine, and visualization pipeline provides a robust blueprint for future work. The modularity of the system makes it straightforward to integrate digital phenotyping data (actigraphy, mobile sensing, mood diaries), calibrate transition dynamics to real patients, introduce population variability, and embed safety constraints appropriate for clinical use.

This study demonstrates that reinforcement learning is capable not only of optimizing reward in an abstract simulation, but of discovering realistic, interpretable, and clinically aligned intervention strategies. The framework presented here lays foundational groundwork for developing next-generation RL-based decision-support tools that can complement chronotherapy and personalized treatment planning in bipolar disorder.

## Conflicts of Interest

The authors declare no conflicts of interest.

## References


[1] McCarthy, M.J., Gottlieb, J.F., Gonzalez, R., McClung, C.A., Alloy, L.B., Cain, S., *et al*. (2021) Neurobiological and Behavioral Mechanisms of Circadian Rhythm Disruption in Bipolar Disorder: A Critical Multi-Disciplinary Literature Review and Agenda for Future Research from the ISBD Task Force on Chronobiology. *Bipolar Disorders*, **24**, 232-263. https://doi.org/10.1111/bdi.13165

[2] Tonon, A.C., Nexha, A., Mendonça da Silva, M., Gomes, F.A., Hidalgo, M.P. and Frey, B.N. (2024) Sleep and Circadian Disruption in Bipolar Disorders: From Psychopathology to Digital Phenotyping in Clinical Practice. *Psychiatry and Clinical Neurosciences*, **78**, 654-666. https://doi.org/10.1111/pcn.13729

[3] Oliva, V., Fico, G., De Prisco, M., Gonda, X., Rosa, A.R. and Vieta, E. (2025) Bipolar Disorders: An Update on Critical Aspects. *The Lancet Regional Health-Europe*, **48**, Article 101135. https://doi.org/10.1016/j.lanepe.2024.101135

[4] Scott, J., Etain, B., Miklowitz, D., Crouse, J.J., Carpenter, J., Marwaha, S., *et al*. (2022) A Systematic Review and Meta-Analysis of Sleep and Circadian Rhythms Disturbances in Individuals at High-Risk of Developing or with Early Onset of Bipolar Disorders. *Neuroscience & Biobehavioral Reviews*, **135**, Article 104585. https://doi.org/10.1016/j.neubiorev.2022.104585

[5] Ghaemi, S.N., Dalley, S., Catania, C. and Barroilhet, S. (2014) Bipolar or Borderline: A Clinical Overview. *Acta Psychiatrica Scandinavica*, **130**, 99-108. https://doi.org/10.1111/acps.12257

[6] Bellivier, F., Geoffroy, P., Etain, B. and Scott, J. (2015) Sleep- and Circadian Rhythm–Associated Pathways as Therapeutic Targets in Bipolar Disorder. *Expert Opinion on Therapeutic Targets*, **19**, 747-763. https://doi.org/10.1517/14728222.2015.1018822

[7] Proudfoot, J., Doran, J., Manicavasagar, V. and Parker, G. (2011) The Precipitants of Manic/Hypomanic Episodes in the Context of Bipolar Disorder: A Review. *Journal of Affective Disorders*, **133**, 381-387. https://doi.org/10.1016/j.jad.2010.10.051

[8] Steardo, L., de Filippis, R., Carbone, E.A., Segura-Garcia, C., Verkhratsky, A. and De Fazio, P. (2019) Sleep Disturbance in Bipolar Disorder: Neuroglia and Circadian Rhythms. *Frontiers in Psychiatry*, **10**, Article ID: 501. https://doi.org/10.3389/fpsyt.2019.00501

[9] Gottlieb, J.F., Benedetti, F., Geoffroy, P.A., Henriksen, T.E.G., Lam, R.W., Murray, G., *et al*. (2019) The Chronotherapeutic Treatment of Bipolar Disorders: A Systematic Review and Practice Recommendations from the ISBD Task Force on Chronotherapy and Chronobiology. *Bipolar Disorders*, **21**, 741-773. https://doi.org/10.1111/bdi.12847

[10] Pal, A., Sidana, I.S. and Avinash, P.R. (2022) Sleep in Bipolar Disorders. In: Gupta, R., Neubauer, D.N. and Pandi-Perumal, S.R., Eds., *Sleep and Neuropsychiatric Disorders*, Springer, 371-396. https://doi.org/10.1007/978-981-16-0123-1_19

[11] Yeom, J.W., Park, S. and Lee, H. (2024) Managing Circadian Rhythms: A Key to Enhancing Mental Health in College Students. *Psychiatry Investigation*, **21**, 1309-1317. https://doi.org/10.30773/pi.2024.0250

[12] Caruso, V., Geoffroy, P.A., Alfi, G., Miniati, M., Riemann, D., Gemignani, A., *et al*. (2024) Effects of Mood Stabilizers on Sleep and Circadian Rhythms: A Systematic Review. *Current Sleep Medicine Reports*, **10**, 329-357.

https://doi.org/10.1007/s40675-024-00298-5

[13] DeSanctis, M.V. (2017) Circadian Principles: Behavioral Health Implications. *Journal of Applied Biobehavioral Research*, **22**, e12102. https://doi.org/10.1111/jabr.12102

[14] Jones, C., Reynolds, C., Olson, R., Bontrager, A., Lambert, S., Balba, N., Weymann, K., *et al.* (2021) 277 Tunable White Light for Elders (TWLITE): A Feasibility Study of a Home-Based Sleep Intervention. *Sleep*, **44**, A111. https://doi.org/10.1093/sleep/zsab072.276

[15] Ou, W. and Bi, S. (2025) Sequential Decision-Making under Uncertainty: A Robust MDPs Review. *Annals of Operations Research*, **353**, 1239-1285.

[16] Barto, A.G., Sutton, R.S. and Watkins, C.J.C.H. (1989) Learning and Sequential Decision Making. Vol. 89, University of Massachusetts.

[17] Powell, W.B. (2021) From Reinforcement Learning to Optimal Control: A Unified Framework for Sequential Decisions. In: Vamvoudakis, K.G., Wan, Y., Lewis, F.L., and Cansever, D., Eds., *Studies in Systems, Decision and Control*, Springer International Publishing, 29-74. https://doi.org/10.1007/978-3-030-60990-0_3

[18] Dimitrakakis, C. and Ortner, R. (2022) Decision Making under Uncertainty and Reinforcement Learning: Theory and Algorithms. Vol. 223, Springer.

[19] Van Moffaert, K. (2016) Multicriteria Reinforcement Learning for Sequential Decision-Making Problems. Ph.D. Thesis, Vrije Universiteit Brussel.

[20] van Otterlo Martijn, (2009) The Logic of Adaptive Behavior. In: *Frontiers in Artificial Intelligence and Applications*, Vol. 192, IOS Press, 1-489. https://doi.org/10.3233/978-1-58603-969-1-i

[21] Jayaraman, P., Desman, J., Sabounchi, M., Nadkarni, G.N. and Sakhuja, A. (2024) A Primer on Reinforcement Learning in Medicine for Clinicians. *NPJ Digital Medicine*, **7**, Article No. 337. https://doi.org/10.1038/s41746-024-01316-0

[22] Munster, M. and Jamshidnejad, A. (2025) Personalized Human-Robot Cognitive Interaction via a Novel Fuzzy Logic Control and Learning-Based Paradigm. *IEEE Access*, **13**, 112568-112593. https://doi.org/10.1109/access.2025.3584194

[23] Gönül, S. (2018) A Framework for Design and Personalization of Digital, Just-in-Time, Adaptive Interventions. Ph.D. Thesis, Middle East Technical University (Türkiye).

[24] Streicher, A. and Smeddinck, J.D. (2016) Personalized and Adaptive Serious Games. *Entertainment Computing and Serious Games: International GI-Dagstuhl Seminar* 15283, Dagstuhl Castle, 5-10 July 2015, 332-377.

[25] Benedictis, R.D., Umbrico, A., Fracasso, F., Cortellessa, G., Orlandini, A. and Cesta, A. (2022) A Dichotomic Approach to Adaptive Interaction for Socially Assistive Robots. *User Modeling and User-Adapted Interaction*, **33**, 293-331. https://doi.org/10.1007/s11257-022-09347-6

[26] Ali, H. (2022) Reinforcement Learning in Healthcare: Optimizing Treatment Strategies, Dynamic Resource Allocation, and Adaptive Clinical Decision-Making. *International Journal of Computer Applications Technology and Research*, **11**, 88-104.

[27] Yu, C., Liu, J., Nemati, S. and Yin, G. (2021) Reinforcement Learning in Healthcare: A Survey. *ACM Computing Surveys* (*CSUR*), **55**, 1-36. https://doi.org/10.1145/3477600

[28] Sachdeva, R.K., Bathla, P., Vij, S., Dishika, Jain, M., Kumar, L., Pradeep Ghantasala, G.S. and Ahuja, R. (2024) Emerging Technologies in Healthcare Systems. In: Mahajan, S., Raj, P. and Pandit, A.K., Eds., *Deep Reinforcement Learning and Its Industrial Use Cases: AI for Real-World Applications*, 375-394.

https://doi.org/10.1002/9781394272587.ch16

[29] Foluke Ekundayo, (2024) Reinforcement Learning in Treatment Pathway Optimization: A Case Study in Oncology. *International Journal of Science and Research Archive*, **13**, 2187-2205. https://doi.org/10.30574/ijsra.2024.13.2.2450

[30] Ahmad, U.J., Bughio, H.K., Channar, N.A. and Bhatti, A.K. (2025) Reinforcement Learning in IoT-Driven Healthcare: Opportunities, Challenges, and Future Directions. *Spectrum of Engineering Sciences*, **3**, 779-790.

[31] Shaik, T., Tao, X., Li, L., Xie, H., Dai, H., Zhao, F., *et al*. (2025) AI-Driven Multi-Agent Reinforcement Learning Framework for Real-Time Monitoring of Physiological Signals in Stress and Depression Contexts. *Brain Informatics*, **12**, Article No. 14. https://doi.org/10.1186/s40708-025-00262-1

[32] Costello, E.J., Pine, D.S., Hammen, C., March, J.S., Plotsky, P.M., Weissman, M.M., Biederman, J., *et al*. (2002) Development and Natural History of Mood Disorders. *Biological Psychiatry*, **52**, 529-542. https://doi.org/10.1016/S0006-3223(02)01372-0

[33] Banumathi, K., Venkatesan, L., Benjamin, L.S., Vijayalakshmi, K. and Satchi, N.S. (2025) Reinforcement Learning in Personalized Medicine: A Comprehensive Review of Treatment Optimization Strategies. *Cureus*, **17**, e82756.

[34] Milic, J., Zrnic, I., Grego, E., Jovic, D., Stankovic, V., Djurdjevic, S., *et al*. (2025) The Role of Artificial Intelligence in Managing Bipolar Disorder: A New Frontier in Patient Care. *Journal of Clinical Medicine*, **14**, Article 2515. https://doi.org/10.3390/jcm14072515

[35] Ribba, B. (2023) Reinforcement Learning as an Innovative Model-Based Approach: Examples from Precision Dosing, Digital Health and Computational Psychiatry. *Frontiers in Pharmacology*, **13**, Article ID: 1094281. https://doi.org/10.3389/fphar.2022.1094281

[36] Weltz, J., Volfovsky, A. and Laber, E.B. (2022) Reinforcement Learning Methods in Public Health. *Clinical Therapeutics*, **44**, 139-154. https://doi.org/10.1016/j.clinthera.2021.11.002

[37] Trella, A.L., Zhang, K.W., Nahum-Shani, I., Shetty, V., Doshi-Velez, F. and Murphy, S.A. (2023) Reward Design for an Online Reinforcement Learning Algorithm Supporting Oral Self-Care. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 15724-15730. https://doi.org/10.1609/aaai.v37i13.26866

[38] Liebenow, B., Jones, R., DiMarco, E., Trattner, J.D., Humphries, J., Sands, L.P., *et al*. (2022) Computational Reinforcement Learning, Reward (and Punishment), and Dopamine in Psychiatric Disorders. *Frontiers in Psychiatry*, **13**, Article ID: 886297. https://doi.org/10.3389/fpsyt.2022.886297

[39] Roggeveen, L.F., Hassouni, A.e., de Grooth, H., Girbes, A.R.J., Hoogendoorn, M. and Elbers, P.W.G. (2024) Reinforcement Learning for Intensive Care Medicine: Actionable Clinical Insights from Novel Approaches to Reward Shaping and Off-Policy Model Evaluation. *Intensive Care Medicine Experimental*, **12**, Article No. 32. https://doi.org/10.1186/s40635-024-00614-x

[40] Chakraborty, B. and Moodie, E.E. (2013) Statistical Methods for Dynamic Treatment Regimes. Vol. 2, Springer.

[41] Eisenberg, L. (1977) Disease and Illness Distinctions between Professional and Popular Ideas of Sickness. *Culture, Medicine and Psychiatry*, **1**, 9-23. https://doi.org/10.1007/bf00114808

[42] Barbierato, E. and Gatti, A. (2024) The Challenges of Machine Learning: A Critical Review. *Electronics*, **13**, Article 416. https://doi.org/10.3390/electronics13020416