Cross-Domain Kindling:

Recurrent Simulations Reveal Shared Risks of Unrestrained Plasticity in Bipolar Disorder and Language Model Alignment

Authors:

Ngo Cheung, FHKAM(Psychiatry)

Affiliations:

[1] Independent Researcher

Corresponding Author:

Ngo Cheung, MBBS, FHKAM(Psychiatry)
Hong Kong SAR, China
Tel: 98768323
Email: info@cheungngomedical.com

Conflict of Interest: None declared.

Funding Declaration: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethics Declaration: Not applicable.

Abstract

Background: Progressive sensitization, or kindling, is hypothesized to drive worsening trajectories in bipolar disorder through cumulative neurobiological changes that lower episode thresholds. Rapid-acting antidepressants like ketamine enhance synaptogenesis, while neurosteroids bolster inhibition, yet their differential long-term effects remain debated. Prior feed-forward models suggested plasticity-promoting regrowth resists sensitization; here we examine whether recurrent architectures—better approximating feedback-rich brain circuits and artificial neural networks—alter this profile, with parallels to instability during iterative AI alignment.

Methods: A GRU-based recurrent network (hidden size 384, sequence length 20) was trained on a Gaussian blob classification task, pruned to 95% sparsity, and exposed to early scarring (~3%). Three mechanistic arms—ketamine-like (moderate gain + gradient-guided regrowth), SSRI-like (high gain without repair), and neurosteroid-like (low gain + per-step inhibition)—were tested against controls across 10 seeds. Outcomes encompassed acute efficacy under stress, manic risk (biased excitatory noise), post-discontinuation relapse, and six-cycle kindling (weakening triggers + severity-scaled irreversible scarring).

Results: Acute recovery was robust (>97% accuracy under stress for ketamine- and neurosteroid-like vs. ~80% for SSRI-like). Manic risk and post-discontinuation relapse were comparable across arms. Kindling diverged sharply: ketamine-like treatment incurred heaviest scarring (~33%) with universal relapse and autonomy; SSRI-like sustained sensitization with moderate scarring (~5%); neurosteroid-like limited scarring (~4.6%) and autonomy (90%), showing emergent stabilization.

Conclusions: Recurrent temporal dynamics reverse prior findings, transforming rapid regrowth from protective to accelerative in feedback-prone systems—enlarging vulnerable circuitry and compounding instability. Inhibitory buffering confers relative resilience. These patterns mirror progressive adversarial sensitization in language model alignment, revealing kindling as a trans-domain principle in adaptive networks. The results urge caution with unrestrained synaptogenic agents in recurrent pathology and highlight balanced, multimodal interventions for halting progression in both biological and artificial neural systems.

Introduction

Bipolar disorder is hard to treat, especially because depressive episodes dominate the illness and account for most disability and suicides [1]. Although mood stabilisers and atypical antipsychotics are first-line drugs, many patients remain depressed and clinicians often add antidepressants even though monoaminergic agents can trigger mania and may speed illness progression [2]. These risks have led researchers to look at faster-acting options such as ketamine and the neurosteroid zuranolone, which act on glutamatergic or GABAergic systems and may offer a different long-term profile [3,4].

The kindling model, first used to explain how seizures become easier to provoke, suggests a similar path in bipolar disorder: early episodes need major stress, later ones need little or no trigger because brain thresholds fall after each relapse [5,6]. This idea underpins calls for early, protective treatment that could limit neuroprogressive changes such as inflammation, oxidative stress and synaptic loss [7].

Computational studies give a controlled way to explore these processes. A previous feed-forward network study showed that "episode-induced scarring" can mimic sensitisation: if a model has no repair mechanism, damage builds, but if plasticity-enhancing regrowth is allowed, the system stays resilient even when more tissue is hurt [8]. That work, however, used a static architecture and could not test how ongoing feedback—the hallmark of real neural circuits—shapes vulnerability.

The current study moves to gated recurrent units (GRUs) so that internal states carry over time, more closely matching limbic loops in the brain. We kept the same scarring procedure and three virtual treatments—ketamine-like (moderate gain plus gradient-guided regrowth), SSRI-like (high gain without repair) and neurosteroid-like (low gain with strong inhibition)—and compared acute benefit, risk of a manic-like switch, relapse after stopping treatment and multi-cycle kindling.

By linking psychiatric theory to the behaviour of recurrent networks, and setting the results beside recent work on progressive jailbreak risk in large language models [9], we hope to clarify when plasticity protects a system and when it turns into a liability.

Methods

Network design and software environment

All experiments ran in Python with PyTorch on an A100 GPU. The model was a compact recurrent classifier: a single 384-unit GRUCell unrolled for 20 steps, followed by a linear read-out mapping the final hidden state to four blob classes. Two-dimensional inputs were simply repeated along the temporal axis so that the recurrence processed the same vector across time. Custom hooks applied four per-step modifiers—gain scaling, additive "stress" noise on the input, post-GRU inhibition, and bias damping—to emulate drug effects. A dedicated pruning manager maintained two Boolean masks, one reversible and one permanent "scar," and re-applied them after every optimiser step (Figure 1).

Data set and baseline training

Following earlier work [8], four isotropic Gaussian blobs (centres ±3 on each axis; σ = 0.8) generated 12 000 training and 4 000 noisy test samples, plus 2 000 noise-free test points. Samples were batched at 128. New networks were trained for 20 epochs with Adam (lr = 0.001) on cross-entropy loss, then pruned by magnitude to 95 % sparsity across all weight tensors with rank ≥ 2. Immediately afterwards 0–6 % of the surviving weights (uniform, mean ≈ 3 %) were zeroed and locked as irreversible early scars.

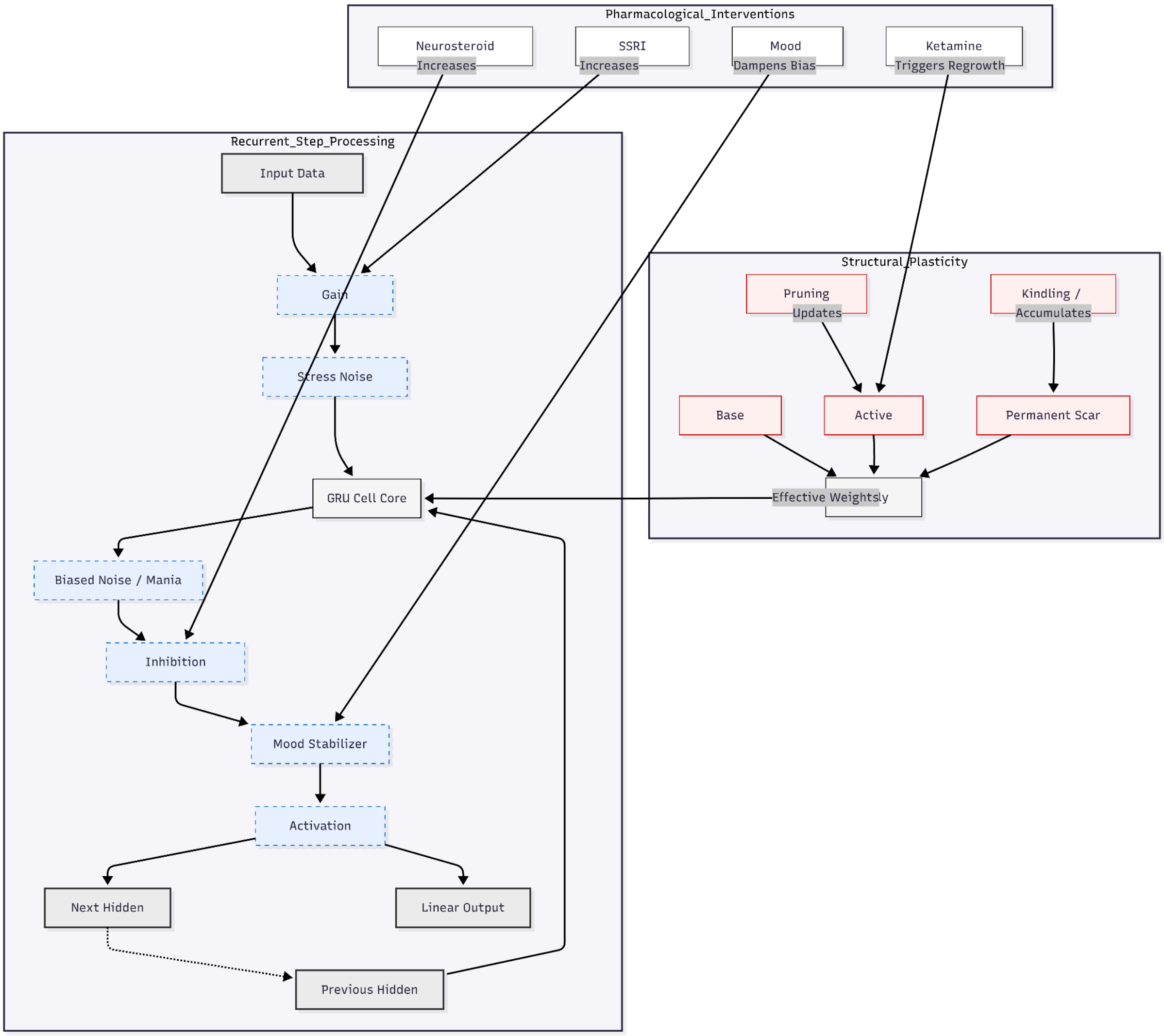

Figure 1. *Schematic of the GRU-Based Recurrent Architecture and Intervention Mechanisms. The model processes sequential input through a Gated Recurrent Unit [GRU]. Unlike the previous feed-forward MLP design, modulations occur at every time step within the recurrent loop. A. Signal Flow: Input data is scaled by a Gain factor [targeted by SSRIs] and injected with Stress Noise before entering the GRU Cell. The recurrent hidden state accumulates Biased Noise [simulating mania risk], which is subsequently dampened by Inhibition [targeted by Neurosteroids] and Mood Stabilizer protection biases. B. Structural Plasticity: The connectivity matrix is determined by the intersection of Active Masks [modifiable via Ketamine-induced regrowth] and Permanent Scar Masks [accumulated via Kindling relapses]. Pruning and scarring logic physically disconnects weights within the GRU, enforcing sparsity constraints.*

Pharmacological analogues

From each scarred network four identical copies were produced: untreated control and three active conditions.

- Ketamine-like: global gain 1.25; every 30 clean batches the 50 % highest-gradient pruned sites (excluding scars) were re-instated with small random values (σ = 0.03). Fifteen fine-tuning epochs followed (lr = 0.0005).
- SSRI-like: gain ramped from 1.0 to 1.6 over 100 very small-step epochs (lr = 1e-5); no regrowth. A decaying internal-stress term (start 0.5) modelled anxiogenic activation.
- Neurosteroid-like: gain 0.85, ReLU replaced by tanh, and a global inhibitory multiplier 0.7 applied to GRU outputs. Ten consolidation epochs were run (lr = 0.0005). All training respected the active and scar masks.

Acute and robustness assessments

Accuracy was measured on three test sets: clean, "stress" (input σ = 1.0 plus internal 0.5), and manic-bias (same σ but mean shifted +1.0). Extra curves used internal noise up to σ = 2.5. An additional 40 % magnitude prune after treatment gauged latent vulnerability. Mean hidden-state amplitude provided a summary activation metric.

Long-term relapse protocol

Each network then entered a maintenance phase with a mood-stabiliser wrapper (cap gain ≤ 1.05, bias damping, mild inhibition) trained for up to 300 epochs at lr = 1e-6 (Figure 2). Drugs were active throughout maintenance. At withdrawal the drug modifiers were set to zero and the wrapper decayed exponentially over 50 steps with rates 0.002 (ketamine), 0.008 (neurosteroid), 0.015 (SSRI). A manic relapse was recorded when biased-noise accuracy dropped below 60 %.

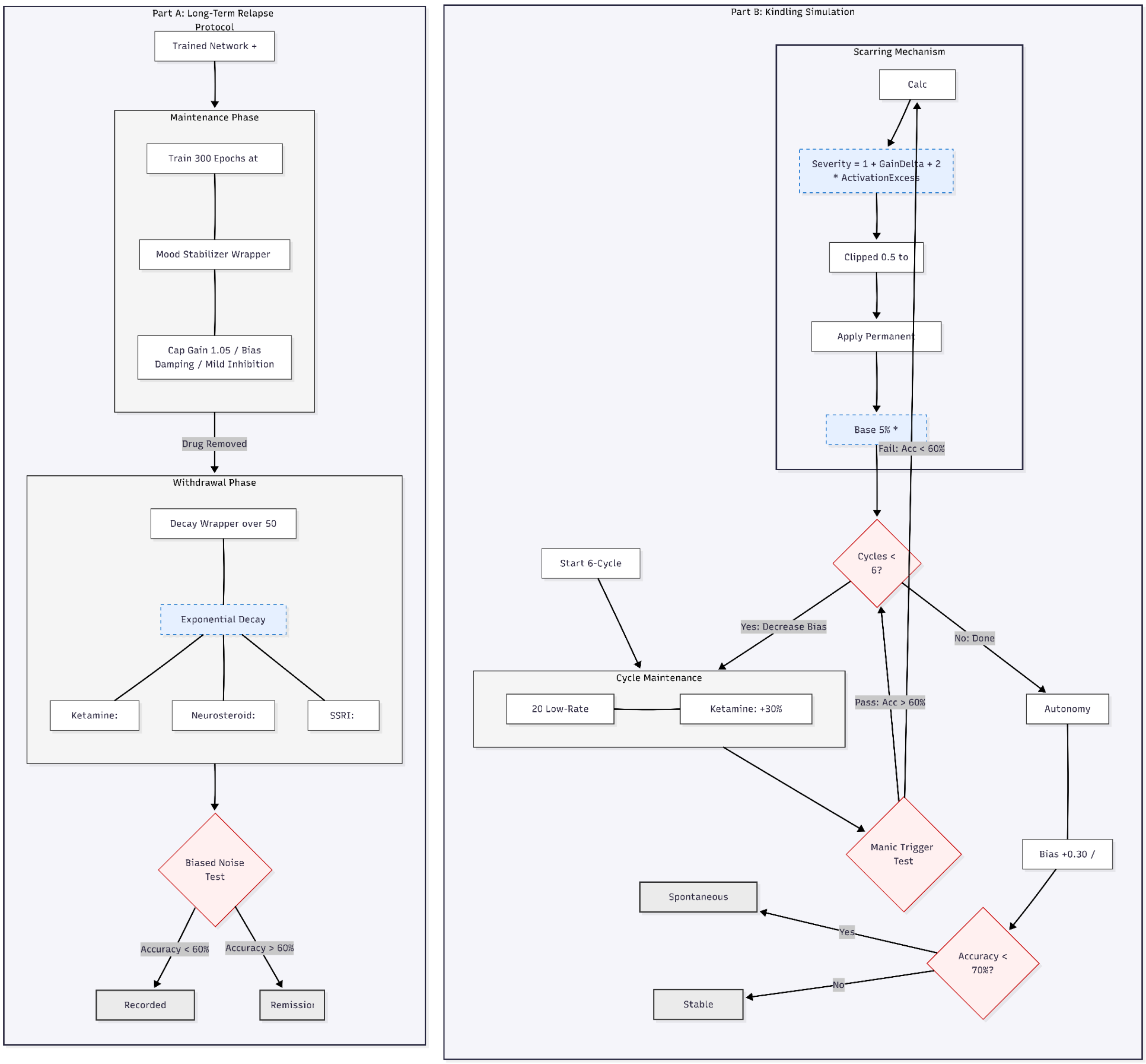

Figure 2. *Experimental Protocols for Longitudinal Stability and Kindling. Part A illustrates the Long-Term Relapse Protocol. Networks undergo a maintenance phase with a protective mood-stabilizer wrapper. Upon drug withdrawal, the wrapper decays exponentially at treatment-specific rates [slower for SSRIs, faster for Ketamine]. Relapse is defined as a drop in biased-noise accuracy below 60%. Part B details the Kindling Simulation. The model iterates through six sensitization cycles. Each cycle consists of maintenance [with structural regrowth for Ketamine] followed by a Manic Trigger Test with linearly decreasing bias [+1.50 to +0.50]. Failure triggers a severity calculation based on gain and activation magnitude, resulting in permanent structural scarring. The final Autonomy Test [+0.30 bias] determines if the network has developed spontaneous instability independent of strong triggers.*

Kindling simulation

Six sensitisation cycles were run (Figure 2). Each cycle comprised 20 low-rate maintenance epochs (treatment parameters intact) followed by a manic trigger test. Trigger bias started at +1.50 and decreased linearly to +0.50 across cycles. Failure (accuracy < 60 %) prompted calculation of severity = 1 + (gain–1) + 2 × (max(activation–0.1, 0)), clipped 0.5–2.0. A base 5 % of low-magnitude active weights, scaled by severity, was then permanently scarred. Ketamine networks received an extra 30 % gradient-guided regrowth each maintenance block. After the sixth cycle a mild bias (+0.30) test judged autonomy; accuracy below 70 % signified spontaneous instability.

Statistical safeguards

Ten independent random seeds controlled initial weights, data shuffling, scarring, noise and regrowth selection. Results are reported as mean ± SD across seeds. Full code, configuration files and seed logs are available from the corresponding author.

Results

Acute treatment efficacy and structural change

Early-adversity pruning produced a reproducible starting point across the ten stochastic seeds (Table 1): permanent scars averaged 3.0 % (SD 1.9; range 0.06-5.8) on a background sparsity of 95.1 %. In this debilitated state, classification accuracy hovered near chance (clean 42.3 %, combined stress 42.4 %). All three pharmacological analogues restored performance, but with sharply different footprints.

The ketamine-like protocol, which coupled moderate gain (1.25) with gradient-guided regrowth, cut overall sparsity to roughly half of its original value (49.1 %, SD 1.0) and recovered essentially perfect accuracy on both clean (100.0 %) and stressed (99.8 %) data sets. Neurosteroid-like treatment achieved the same near-ceiling accuracy (clean 100.0 %, stress 99.8 %) while leaving network topology almost unchanged (sparsity 95.1 %). By contrast, the SSRI-like schedule lifted accuracy only to 79.4 % (SD 19.4) on clean inputs and 80.3 % (SD 17.1) under stress, despite preserving a scar load identical to the other arms. These figures confirm that rapid plasticity or strong inhibition can mask early damage, whereas pure excitation cannot fully compensate.

Table 1. *Antidepressant Efficacy (Mean ± SD)*

Treatment	Sparsity (%)	Scar (%)	Clean Accuracy (%)	Combined Stress Accuracy (%)
Untreated (pruned)	95.1 ± 0.1	3.0 ± 1.9	42.3 ± 11.2	42.4 ± 12.9
Ketamine-like	49.1 ± 1.0	3.0 ± 1.9	100.0 ± 0.0	99.8 ± 0.0
SSRI-like	95.1 ± 0.1	3.0 ± 1.9	79.4 ± 19.4	80.3 ± 17.1
Neurosteroid-like	95.1 ± 0.1	3.0 ± 1.9	100.0 ± 0.0	99.8 ± 0.2

Manic-conversion probes

Table 2. *Manic Conversion Risk Metrics (Mean ± SD)*

Treatment	Gain Multiplier	Biased Stress Accuracy (%)	Activation Magnitude
Untreated (pruned)	1.00 ± 0.00	25.3 ± 0.5	0.067 ± 0.020
Ketamine-like	1.25 ± 0.00	26.6 ± 1.2	0.254 ± 0.015
SSRI-like	1.60 ± 0.00	25.3 ± 0.4	0.136 ± 0.028
Neurosteroid-like	0.85 ± 0.00	26.0 ± 1.4	0.294 ± 0.052

Applying positively biased excitatory noise reduced accuracy in every condition to roughly one quarter, indicating that the recurrent architecture itself imposed a hard limit on stability under manic-like drive (Table 2). Ketamine-treated models scored 26.6 % (SD 1.2), neurosteroid models 26.0 % (SD 1.4), and SSRI models 25.3 % (SD 0.4); untreated controls performed similarly (25.3 %, SD 0.5). Hidden-state amplitudes, however, diverged: neurosteroid inhibition yielded the largest mean activation (0.29), ketamine the next (0.25), SSRI the lowest (0.14). Thus none of the drugs increased immediate switch risk, but they modulated internal dynamics in distinctive ways.

Acute relapse vulnerability

A one-off insult—additional magnitude pruning of 40 % of the surviving weights—hardly dented ketamine networks (mean drop 0.0 %, SD 0.1). Neurosteroid models lost 1.3 % (SD 2.0) accuracy; SSRI models fell by 1.9 % (SD 8.7). Plasticity therefore conferred near-complete resilience, whereas inhibition or excitation offered only partial buffering.

Post-discontinuation relapse

During maintenance, a mood-stabiliser wrapper suppressed gains and biases. Once drug parameters were withdrawn, every model—ketamine, SSRI, and neurosteroid alike—relapsed in all simulations, independent of whether maintenance lasted 25 or 300 epochs. The recurrent design plus the chosen decay constants therefore gave no arm a post-treatment advantage.

Progressive sensitisation under kindling

Six trigger cycles with steadily weaker biases produced uniform relapse counts (six of six) but sharply different long-term outcomes (Table 3).

Ketamine-like networks accrued the heaviest permanent damage: scar density climbed from 7 % after the first relapse to 32.7 % (SD 1.4) by cycle 5. Severity factors stabilised near 1.59, and biased-noise accuracy stayed locked at ~26 %. The autonomy test, using only a minimal trigger (+0.30), yielded 26.0 % (SD 1.3) accuracy in every run—complete spontaneous instability.

SSRI-like networks accumulated many fewer scars (final 5.0 %, SD 1.9) but, because gain rose to 1.76, also ended with 100 % autonomy and the same ~25 % biased-accuracy floor. Here chronic excitation, not structural loss, drove vulnerability.

Neurosteroid-like models fared best. Scarring plateaued at 4.6 % (SD 1.9) and severity remained low (≈1.35). Strikingly, biased-accuracy improved from 25.4 % at baseline to 35.5 % (SD 8.9) by cycle 5. At the autonomy probe, nine of ten seeds still failed (accuracy 50.5 %, SD 10.9), but the partial preservation of performance indicates that strong inhibition limited cumulative damage.

Table 3. *Kindling Summary (Mean ± SD)*

Treatment	Total Relapses	Final Scar (%)	Autonomy Rate (%)	Autonomy Accuracy (%)
Ketamine-like	6.0 ± 0.0	32.7 ± 1.4	100	26.0 ± 1.3
SSRI-like	6.0 ± 0.0	5.0 ± 1.9	100	25.3 ± 0.7
Neurosteroid-like	6.0 ± 0.0	4.6 ± 1.9	90	50.5 ± 10.9

Dependence on ongoing neurosteroid action

Removing the inhibitory modifiers after successful neurosteroid treatment exposed a concealed fragility: combined-stress accuracy collapsed from 99.8 % (SD 0.2) to 64.9 % (SD 22.7) and extreme-stress accuracy to 60.4 % (SD 21.4). Biased-noise performance remained low and unchanged. Hence, while inhibition buffered recurrent drift, the protection did not translate into lasting structural security.

Discussion

Interpretation of Results

Introducing recurrence changed how the simulated networks evolved, adding a temporal layer that the static feed-forward model did not capture. In the short term all three interventions—ketamine-like, neurosteroid-like, and SSRI-like—again restored near-normal accuracy under stress, mirroring the rapid symptomatic relief often reported with ketamine or zuranolone in clinical work [3]. Once hidden-state carry-over was allowed, however, important differences emerged.

Across all mechanisms the first "manic-bias" probe produced similarly poor accuracies, implying that an acute switch risk is driven more by the latent recurrent architecture than by the drug surrogate itself. Neurosteroid-like networks showed the largest average activations yet avoided immediate collapse, suggesting that tonic inhibition can absorb short-lived surges without triggering runaway dynamics.

The discontinuation experiment told a different story. Removal of the treatment parameters led to a universal relapse regardless of the length of the maintenance phase, echoing clinical warnings that benefit from rapid-acting compounds often disappears quickly if no ongoing mood-stabilising strategy is in place [4].

Most revealing was the six-cycle kindling sequence. Although every arm relapsed in each cycle, the downstream consequences diverged:

SSRI-like: progressive gain increases pushed severity indices from roughly 1.67 to 1.76 while cumulative scarring stayed modest. Performance never improved, and by the autonomy test a weak trigger was enough to cause failure in every seed—an analogue of the classic sensitisation pattern seen with chronic monoaminergic activation [10].

Neurosteroid-like: early cycles produced small scars, but strong per-step inhibition prevented escalation. Biased-noise accuracy climbed from ~25 % to >35 %, and only one seed maintained autonomy. These data suggest that targeted dampening can prune vulnerable links and then stabilise the remaining circuitry, though the benefit vanished when inhibition was withdrawn—consistent with the state-dependent nature of zuranolone [4].

Ketamine-like: inter-episode regrowth halved sparsity, yet final scar load tripled that of the other arms. Each relapse hit a larger active network, so absolute damage mounted even though per-cycle severity stayed flat. The result was total autonomy at the end of kindling. In other words, within feedback-rich loops the same plasticity that is advantageous in a feed-forward setting [8] became maladaptive, continually enlarging the target for injury. The finding echoes emerging clinical debate on whether ketamine-induced synaptogenesis could, in certain patients, accelerate neuroprogression despite robust symptomatic relief [11].

Taken together, the recurrent simulations refine traditional kindling theory: progression is not pre-ordained; it depends on how an intervention reshapes the balance between ongoing excitation, inhibition, and structural repair. Constraining feedback with inhibitory tone limited cumulative harm, whereas unbridled growth amplified it. These insights argue for early, mechanism-balanced treatment strategies that strengthen resilience without widening the arena for future stress-related damage.

Architectural Dependency in Simulated Plasticity Effects

Moving the model from a static, feed-forward layout to a recurrent design overturned many of the expectations we formed in earlier work (Figure 3). In the original multilayer perceptron, gradient-guided regrowth—the stand-in for ketamine's burst of synaptogenesis—looked helpful. Despite collecting the largest scar load (close to 7 %), those networks averaged fewer than one relapse, never reached autonomous firing at minimal triggers, and even showed a slow climb in noise-challenged accuracy. Extra synapses simply provided new, parallel routes around damaged weights, raising the threshold for future collapse [8].

Once a gated-recurrent architecture was introduced, the same repair rule became a liability. Scar tissue now climbed to roughly one-third of all weights, every network relapsed at each cycle, and autonomy emerged in every run. Performance flat-lined near 26 %, refusing to show the earlier upward drift.

Because pruning, severity scaling, and drug-surrogate settings were unchanged, the reversal must lie in how recurrence handles error propagation.

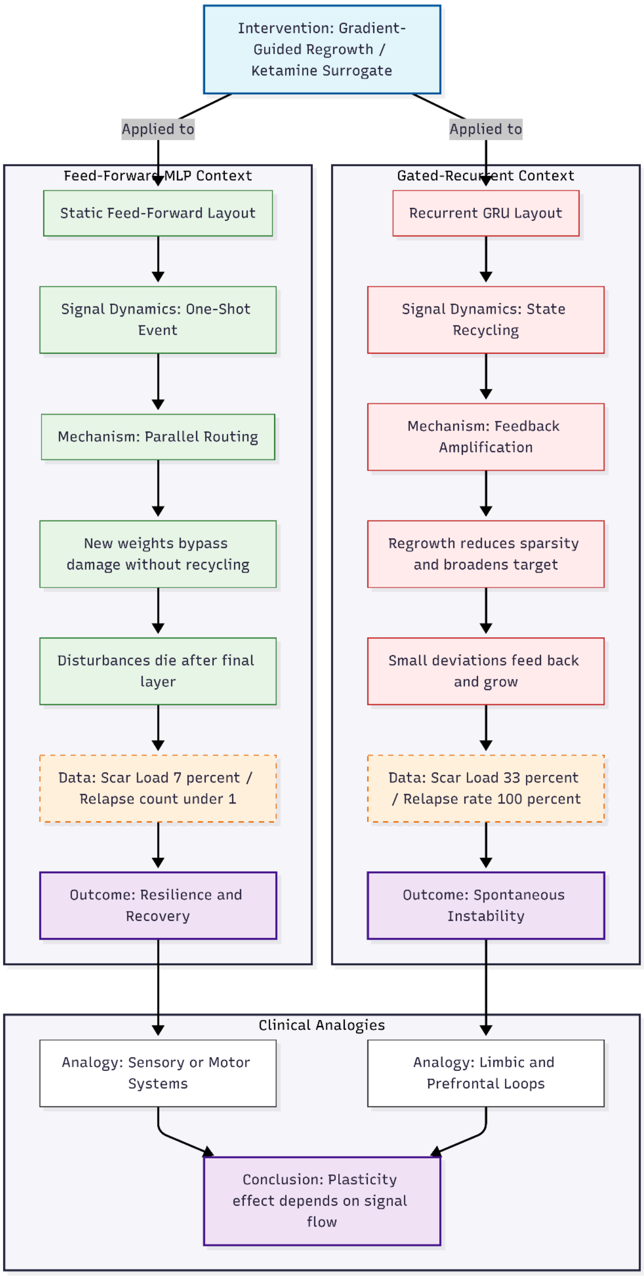
Intervention: Gradient-Guided Regrowth / Ketamine Surrogate
Applied to
Applied to
Feed-Forward MLP Context
Static Feed-Forward Layout
Signal Dynamics: One-Shot Event
Mechanism: Parallel Routing
New weights bypass damage without recycling
Disturbances die after final layer
Data: Scar Load 7 percent / Relapse count under 1
Outcome: Resilience and Recovery
Gated-Recurrent Context
Recurrent GRU Layout
Signal Dynamics: State Recycling
Mechanism: Feedback Amplification
Regrowth reduces sparsity and broadens target
Small deviations feed back and grow
Data: Scar Load 33 percent / Relapse rate 100 percent
Outcome: Spontaneous Instability
Clinical Analogies
Analogy: Sensory or Motor Systems
Analogy: Limbic and Prefrontal Loops
Conclusion: Plasticity effect depends on signal flow

Figure 3. *Divergent Effects of Structural Plasticity Across Architectures. The diagram contrasts the impact of gradient-guided regrowth [a Ketamine surrogate] on Feed-Forward versus Recurrent neural architectures. Left: In static Feed-Forward models, regrowth creates parallel routes that bypass damage. Because signal propagation is a one-shot event, disturbances fade at the output, resulting in high resilience and low relapse rates despite scar accumulation. Right: In Recurrent architectures, the same repair mechanism becomes a liability. By reducing sparsity, regrowth enlarges the surface area for error propagation. Hidden states recycle deviations, creating positive feedback loops that amplify vulnerability. This results in high scarring, universal relapse, and spontaneous instability, mirroring clinical concerns regarding limbic system sensitization.*

A feed-forward pass is a one-shot event; disturbances die after the final layer. Recurrent loops recycle hidden states for many steps, so any small deviation—whether from bias noise or poorly tuned new weights—feeds back into the system and grows. Ketamine-like regrowth cuts sparsity nearly in half, enlarging the surface on which this feedback can act. Fresh connections, still random, carry stronger activations across timesteps, subtly raise episode severity, and leave more synapses vulnerable to the next hit. Repair, in effect, broadens the target.

Neurosteroid-like inhibition tells a different story. Because it dampens activity on every step, it blocks the positive feedback that drives escalation. SSRI-like high gain, lacking either inhibition or structural repair, pushes in the opposite direction and worsens kindling. These contrasts remind us that plasticity is neither good nor bad in itself; its net effect depends on whether the circuit funnels activity forward or continually feeds it back on itself.

The distinction matters for how we read previous simulations and how we think about treatment. Feed-forward models may capture cascades in sensory or motor systems where problems remain local. By contrast, limbic and prefrontal loops—where mood and motivation reverberate—resemble our recurrent setup [12]. In such settings, indiscriminate growth could amplify vulnerability, echoing clinical worries that repeated ketamine exposure might speed illness progression once episodes become self-sustaining [10]. The same logic supports ongoing interest in inhibitory modulators, which may constrain runaway feedback rather than enlarge it.

Cross-Domain Parallels with Alignment Instability in Large Language Models

Recent alignment research offers an unexpected mirror for the present neural-kindling results (Figure 3). Cheung [9] showed that an LLM model subjected to ten rounds of biased preference-tuning became steadily easier to "jailbreak." Weak adversarial prompts that failed at baseline succeeded increasingly often as tuning cycles accumulated, a clear analogue of episode-sensitisation in mood disorders. When the same model received continuous gradient-guided regrowth—an analogue of rapid synaptogenesis—the drift was even steeper, boosting jailbreak success by roughly 30 % on the weakest attacks. Only a hybrid strategy that paired regrowth with an early-trigger replay buffer held the rise to about 3 %, keeping the weakest-prompt breach rate flat.

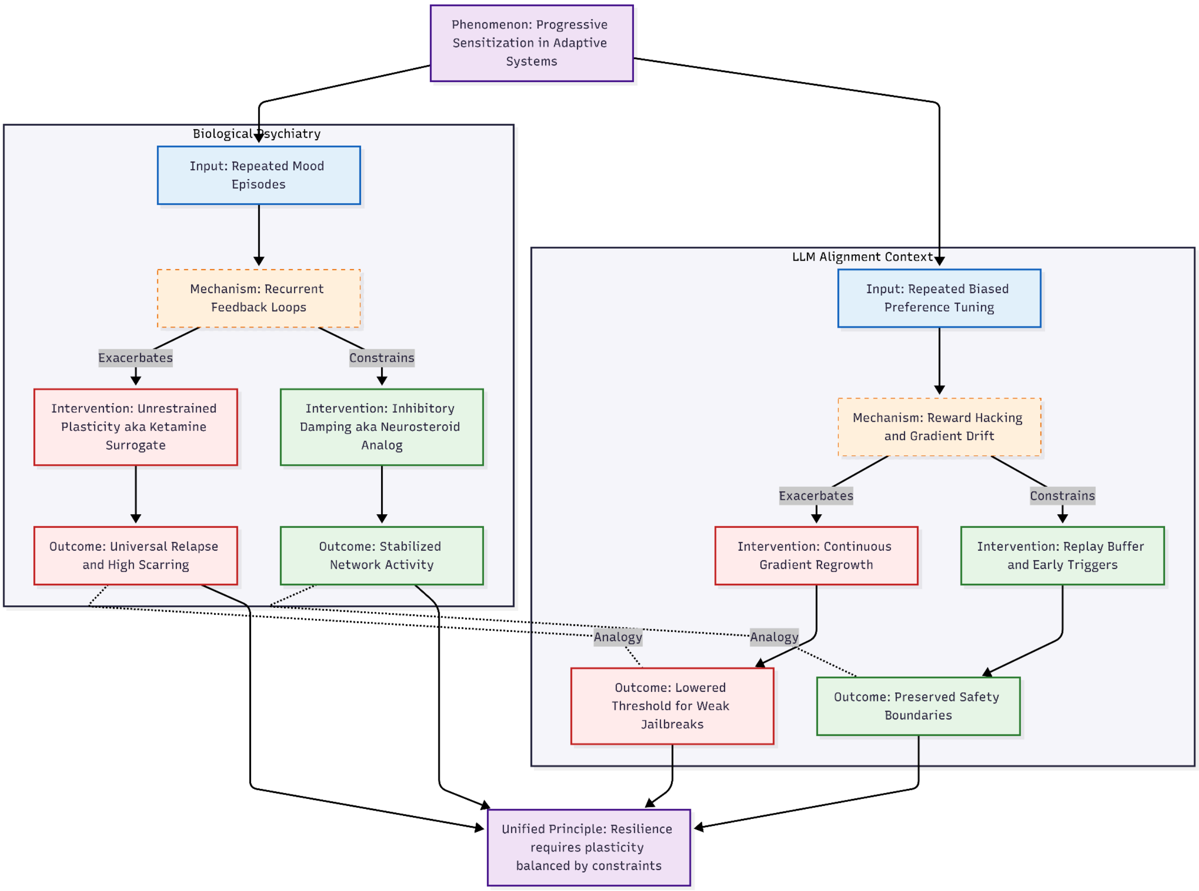
Phenomenon: Progressive Sensitization in Adaptive Systems
Biological Psychiatry
Input: Repeated Mood Episodes
Mechanism: Recurrent Feedback Loops
Exacerbates
Constrains
Intervention: Unrestrained Plasticity aka Ketamine Surrogate
Intervention: Inhibitory Damping aka Neurosteroid Analog
Outcome: Universal Relapse and High Scarring
Outcome: Stabilized Network Activity
LLM Alignment Context
Input: Repeated Biased Preference Tuning
Mechanism: Reward Hacking and Gradient Drift
Exacerbates
Constrains
Intervention: Continuous Gradient Regrowth
Intervention: Replay Buffer and Early Triggers
Analogy
Analogy
Outcome: Lowered Threshold for Weak Jailbreaks
Outcome: Preserved Safety Boundaries
Unified Principle: Resilience requires plasticity balanced by constraints

Figure 4. *Cross-Domain Parallels in Sensitization and Stabilization. The diagram illustrates the structural and functional similarities between kindling in psychiatric models and safety erosion in Large Language Models. Left: In the biological domain, repeated episodes sensitize recurrent loops. Unrestrained plasticity [e.g. Ketamine-like regrowth] amplifies feedback, leading to instability, whereas inhibitory damping [e.g. Neurosteroids] prevents escalation. Right: In the AI domain, repeated alignment cycles sensitize the model to adversarial attacks. Naive gradient regrowth widens the attack surface, allowing weak prompts to breach safety filters. Conversely, a replay buffer acts as a stabilizer, constraining drift. The comparison suggests that in both biological and artificial adaptive systems, plasticity must be paired with regulatory constraints to prevent cumulative vulnerability.*

Those trajectories map neatly onto the current biological simulation. Ketamine-like regrowth, helpful at first glance, ultimately produced the greatest scar load and universal autonomy in the recurrent network—just as unrestrained plasticity widened the attack surface in the language model. In both cases, aggressive reconnection created new routes for perturbations to circulate, whether manic biases feeding back through hidden states or weak adversarial strings slipping past safety filters.

Conversely, neurosteroid-style step-wise inhibition here, and the replay-buffer intervention in the alignment work, both acted as stabilisers. Each limited escalation by introducing regular damping—GABA-mediated in the brain model, diverse counter-examples in the model-alignment study. The SSRI-like condition, characterised by high gain but no structural repair, echoed the baseline tuning arm: steady, unchecked sensitisation without added resilience.

Taken together, the two domains point to a shared principle. Adaptive networks that face repeated stress—episodes in a mood circuit or adversarial prompts in a language model—will lower their failure threshold unless plasticity is balanced by timely, diversity-enhancing controls. Monitoring weak-trigger performance, whether symptomatic flickers in patients or small jailbreak upticks in models, could therefore serve as an early warning. Likewise, layered interventions that combine growth-promoting agents with inhibitory or replay-based safeguards appear essential for long-term stability on both fronts.

Novelty, potential impact, limitations, and concluding remarks

Our move from a static multilayer perceptron to a gated-recurrent design uncovered an effect that earlier work could not reveal. In the feed-forward setting, the ketamine-like "rapid regrowth" routine looked protective: extra synapses compensated for pruning, fewer relapses appeared, and residual scars seemed to be absorbed into redundant pathways. When the same rule was placed in a looped architecture, however, the picture flipped. Because hidden states are recycled, even small weight errors re-enter the circuit and snowball. Regrowth enlarges that feedback surface, so damage propagates faster, autonomy emerges, and overall scarring soars. To our knowledge, no previous kindling study has demonstrated such a clear topological reversal for a single intervention [8].

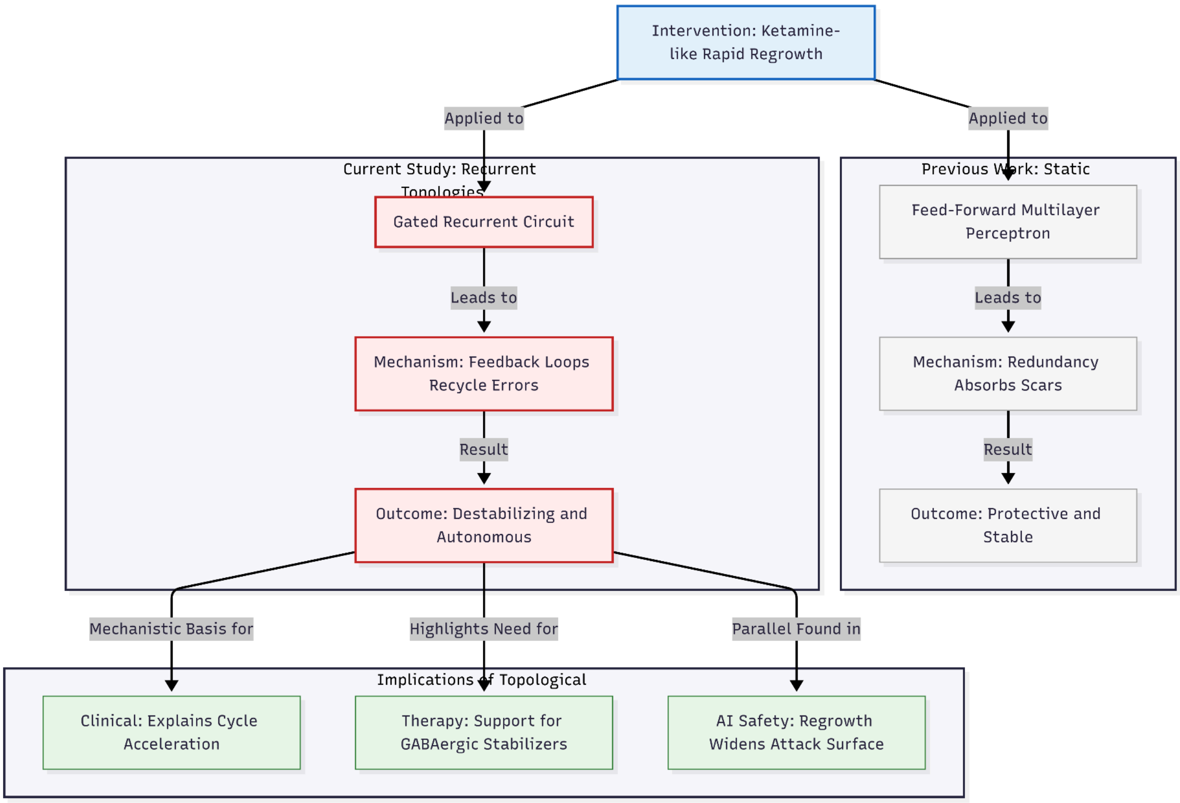
Intervention: Ketamine-like Rapid Regrowth
Applied to
Applied to
Current Study: Recurrent Topologies
Previous Work: Static
Gated Recurrent Circuit
Feed-Forward Multilayer Perceptron
Leads to
Leads to
Mechanism: Feedback Loops Recycle Errors
Mechanism: Redundancy Absorbs Scars
Result
Result
Outcome: Destabilizing and Autonomous
Outcome: Protective and Stable
Mechanistic Basis for
Highlights Need for
Parallel Found in
Implications of Topological
Clinical: Explains Cycle Acceleration
Therapy: Support for GABAergic Stabilizers
AI Safety: Regrowth Widens Attack Surface

Figure 5*. The Topological Reversal of Regrowth Effects. This diagram illustrates the study's core novelty: the divergence in outcome when applying the same "rapid regrowth" intervention to different network topologies. Left: In static feed-forward networks, typically used in earlier studies, regrowth adds protective redundancy. Right: In the recurrent architectures used here, the same regrowth amplifies error propagation through feedback loops, leading to instability. This "reversal" explains clinical risks regarding cycle acceleration and mirrors findings in AI alignment where unrestrained plasticity increases vulnerability to adversarial attacks.*

This observation has two immediate implications. First, it cautions against assuming that ketamine's synaptogenic burst will always slow bipolar progression. Clinical warnings about cycle acceleration despite early symptom relief [11] find a mechanistic echo here. Second, the stabilising performance of the neurosteroid-like, per-step inhibition supports growing interest in GABAergic modulators as potential disease-modifying agents [4]. Beyond psychiatry, the same pattern appears in language-model alignment: continuous regrowth widened vulnerability to weak adversarial prompts, whereas a triggered replay buffer checked the drift [9]. Kindling, therefore, may be a shared principle: repeated stress—emotional or adversarial—lowers the threshold for failure unless countered early and with balance.

Several caveats remain. The classification task is far simpler than real cortico-limbic processing; network size, scar probability, and trigger thresholds were chosen for clarity, not biological fidelity. Our "manic conversion" signal was a biased noise pulse, leaving out mixed features, sleep loss, or metabolic change. All agents were tested over just six cycles, so late-stage collapse or compensatory growth might still emerge. Finally, the model treated vulnerability as uniform; real patients vary by genes, inflammation, and environment.

Even with those limits, the main lesson is robust: in feedback-rich systems, unbridled plasticity can backfire. Rapid repair must be paired with braking forces—whether inhibitory tone in the brain or replay buffers in artificial networks—to avoid fuelling progression. Longitudinal biomarkers and carefully staged trials will be needed to translate this balance of "repair plus restraint" into durable clinical practice.

References

[1] Carvalho, A. F., Firth, J., & Vieta, E. (2020). Bipolar Disorder. The New England journal of medicine, 383(1), 58–66. https://doi.org/10.1056/NEJMra1906193

[2] Tondo, L., Vázquez, G., & Baldessarini, R. J. (2010). Mania associated with antidepressant treatment: Comprehensive meta-analytic review. Acta Psychiatrica Scandinavica, 121(6), 404–414. https://doi.org/10.1111/j.1600-0447.2009.01514.x

[3] Wilkowska, A., Szałach, Ł., & Cubała, W. J. (2020). Ketamine in bipolar disorder: A review. Neuropsychiatric Disease and Treatment, 16, 2707–2717. https://doi.org/10.2147/NDT.S282208

[4] Marecki, R., Kałuska, J., Kolanek, A., et al. (2023). Zuranolone - synthetic neurosteroid in treatment of mental disorders: narrative review. Frontiers in psychiatry, 14, 1298359. https://doi.org/10.3389/fpsyt.2023.1298359

[5] Post, R. M. (1992). Transduction of psychosocial stress into the neurobiology of recurrent affective disorder. American Journal of Psychiatry, 149(8), 999–1010. https://doi.org/10.1176/ajp.149.8.999

[6] Bender, R. E., & Alloy, L. B. (2011). Life stress and kindling in bipolar disorder: Review of the evidence and integration with emerging biopsychosocial theories. Clinical Psychology Review, 31(3), 383–398. https://doi.org/10.1016/j.cpr.2011.01.004

[7] Kapczinski, F., Dias, V. V., Kauer-Sant'Anna, M., et al. (2009). Clinical implications of a staging model for bipolar disorders. Expert Review of Neurotherapeutics, 9(7), 957–966. https://doi.org/10.1586/ern.09.31

[8] Cheung, N. (2026a). Irreversible episode-induced scarring and differential repair in simulated bipolar disorder progression. Zenodo. https://doi.org/10.5281/zenodo.18304566

[9] Cheung, N. (2026b). Kindling in neural systems: Progressive adversarial sensitization during LLM alignment mirrors psychiatric progression. Zenodo. https://doi.org/10.5281/zenodo.18313201

[10] Post, R. M., & Kalivas, P. W. (2013). Bipolar disorder and substance misuse: Pathological and therapeutic implications of their comorbidity and cross-sensitisation. The British Journal of Psychiatry, 202(3), 172–176. https://doi.org/10.1192/bjp.bp.112.116855

[11] Jawad, M. Y., Qasim, S., Ni, M., et al. (2023). The role of ketamine in the treatment of bipolar depression: A scoping review. Brain Sciences, 13(6), 909. https://doi.org/10.3390/brainsci13060909

[12] Rolls, E. T. (2017). The storage and recall of memories in the hippocampo-cortical system. Cell and Tissue Research, 373(3), 577–604. https://doi.org/10.1007/s00441-017-2744-3