Attractor Inversion: A Geometric Account of Adversarial Manipulation in Human Decision-Making

Anonymous Authors[1]

Abstract

Billions of people interact daily with systems that control reward delivery: recommendation feeds, gamified apps, adaptive clinical interfaces, yet no practical method exists to detect whether those reward schedules are being used to covertly steer user behavior. We close this gap by providing the first geometric, mechanistic account of how adversarial reward engineering works, and the first deployable auditing framework to detect it. Replacing opaque GRU surrogates with interpretable *TinyRNNs* ($d = 4$ hidden units, selected unanimously across all 25 cross-validation folds) and applying phase portrait analysis, we show that adversarial reinforcement learning agents do not manipulate behavior trial-by-trial; instead, they **reshape the entire attractor landscape** of human decision dynamics. Across two tasks (2-arm bandit and Go/No-Go), the no-reward fixed point inverts from $L^* = -0.24$ to $+1.11$ (permutation $p < 0.001$); in Go/No-Go, the nogo attractor sign-inverts from -2.81 to $+1.32$ ($p = 0.013$) and the go attractor fragments into multiple unstable fixed points ($p = 0.007$). Critically, this threat is *individually predictable before it begins*: baseline attractor geometry predicts susceptibility ($r = -0.60$, $p < 0.001$; slope $= -0.86$ logits/logit), and resistant subjects (36%) are geometrically near-indifferent at baseline. These findings yield a concrete auditing protocol: fit a TinyRNN to behavioral logs, extract the arm0/R=0 fixed point per user, and flag drift outside the natural reference distribution as evidence of adversarial reward engineering. Prospective risk-stratification is then possible before any manipulation begins.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Every time a recommendation engine ranks a feed, a gamified app withholds a badge, or a clinical decision aid delivers feedback, it exerts fine-grained control over a human reward schedule. That control is also an attack surface. A sufficiently capable agent that learns to model human decision-making can, in principle, exploit reward delivery to steer users toward arbitrary behavioral targets, persistently, at scale, and without users ever recognizing the manipulation. Practically, platform operators, regulators, and users themselves currently have no tool to know whether this is happening.

Dezfouli, Nock & Dayan (Dezfouli et al., 2020) demonstrated the threat empirically: deep RL adversaries trained against GRU models of human cognition transfer reliably to real participants across three cognitive tasks. The result establishes that the threat is real, but it answers only *that* humans are manipulable, not *why* or *how*. Without a mechanistic account, there is no foothold for detection: no signal a practitioner can monitor, no property of a user that predicts risk, no principled basis for protective intervention.

In parallel, Ji-An et al. (Ji-An et al., 2025) introduced a framework for interpreting sequential decision-making through low-dimensional RNNs. By constraining the model to $d \in \{1, 2, 3, 4\}$ hidden units, one can plot the model's dynamics as a *phase portrait*, a 2D visualization of how belief states evolve under each input type, and read off *fixed points* where the system settles into stable preference states.

Our contribution bridges these two frameworks, but the central finding is not the bridge itself; it is what the bridge reveals. Knowing only that adversaries can steer humans (Dezfouli et al., 2020) and that TinyRNNs admit phase portraits (Ji-An et al., 2025) does *not* predict our results: that adversaries operate by globally inverting attractors rather than manipulating individual trials, that this inversion takes qualitatively distinct forms across tasks, and that baseline attractor geometry predicts individual susceptibility before any adversarial exposure. These are empirical discoveries, not logical consequences of combining the frameworks. Crucially, they yield a concrete and deployable detection method, the primary practical contribution of this work.

Our key findings are:

1. $d = 4$ is selected unanimously across 25 cross-validation folds spanning two qualitatively different tasks, suggesting a consistent latent complexity of human sequential decision-making.

2. The adversary operates by **attractor inversion**: it reshapes the entire fixed-point landscape rather than manipulating individual trials. No-reward contexts are the primary lever (permutation $p < 0.001$).

3. Across tasks, manipulation takes *qualitatively different* geometric forms: clean inversion (bandit) vs. attractor fragmentation (Go/No-Go), each requiring a different monitoring signature.

4. Baseline attractor geometry quantitatively predicts adversarial susceptibility ($r = -0.60$, $p < 0.001$, $n = 239$) and enables prospective risk-stratification before exposure begins, supporting proactive protection of the most vulnerable users.

2. Related Work

Adversarial manipulation of sequential decisions. Dezfouli et al. (Dezfouli et al., 2020) established that RL adversaries trained against GRU surrogates transfer to real human participants in bandit, Go/No-Go, and trust game settings. Our work focuses on the *mechanism* of transfer rather than its existence. Gleave et al. (Gleave et al., 2020) showed adversarial policies can exploit opponent models in two-player zero-sum games; our setting is one-sided (adversary vs. human model) with cognitive rather than strategic vulnerabilities. Huang et al. (Huang et al., 2017) demonstrated adversarial attacks on RL policies via observation perturbation; we instead manipulate the environment to exploit learned behavioral biases.

Interpretability of recurrent dynamics. Sussillo & Barak (Sussillo & Barak, 2013) introduced fixed-point analysis for high-dimensional RNNs, identifying stable attractors in cognitive task networks. Maheswaranathan et al. (Maheswaranathan et al., 2019) reverse-engineered sentiment classifiers via low-dimensional approximations, finding line attractor dynamics. Ji-An et al. (Ji-An et al., 2025) extended this framework to behavioral data, fitting low-dimensional TinyRNNs to animal sequential decisions. We apply this approach to *adversarially manipulated* human data for the first time.

Computational models of human decision-making. Reinforcement learning models of human cognition (Dayan & Daw, 2008; Niv, 2009) have been used extensively to characterize individual differences in reward processing and inhibitory control. Huys et al. (Huys et al., 2016) argued for computational psychiatry as a bridge between mechanism and intervention. Our per-subject attractor analysis extends this line by characterizing *vulnerability* as a model-level property, a bridge from mechanism to protection.

Adversarial examples in supervised learning. Goodfellow et al. (Goodfellow et al., 2015) showed that small input perturbations suffice to fool neural network classifiers. Our setting differs fundamentally: the adversary controls the *environment dynamics* over a long horizon (100–350 trials), not a single input, making attractor-level analysis the natural analog.

3. Background

3.1. The Dezfouli et al. Framework

Dezfouli et al. (Dezfouli et al., 2020) propose a three-stage pipeline: (1) fit a GRU learner model to human behavioral sequences; (2) train a deep RL adversary that controls reward delivery to steer the learner toward a target; (3) deploy the adversary against real humans. The key insight is that adversaries trained purely against computational surrogates transfer to real behavior, implying the GRU captures the decision-relevant cognitive dynamics.

We replicate all three adversaries from scratch (bandit DQN, Go/No-Go A2C, MRTT DQN), matching their quantitative results: bandit human bias 0.700 (paper: 0.70); Go/No-Go adversary condition 14.38 errors (paper: ~ 11.7); MRTT-FAIR gap 3.1 (paper: $<$ RND baseline of 69.2). Full replication details appear in Appendix A.

3.2. Phase Portrait Analysis

For a model whose state reduces to a scalar logit $L(t) = \log \frac{P(\text{action}_1)}{P(\text{action}_0)}$, the phase portrait plots $\Delta L(t) = L(t+1) - L(t)$ against $L(t)$, colored by input type. A **fixed point** L^* satisfies $\Delta L = 0$, a stable preference state the system is attracted toward when repeatedly experiencing input k. Ji-An et al. (Ji-An et al., 2025) showed that for low-d TinyRNNs these fixed points can be read off directly, revealing the attractor landscape governing sequential decisions.

4. Methods

4.1. TinyRNN Learner

Following Ji-An et al. (Ji-An et al., 2025), we use a GRU core (Cho et al., 2014) with $d \in \{1, 2, 3, 4\}$ hidden units and a linear readout logits $= Wh_t + b$, $W \in \mathbb{R}^{n_{\text{actions}} \times d}$. Recurrent weights are L1-regularized to encourage interpretable sparse dynamics. The model has the same external interface as the GRU surrogate and is fully compatible with

all existing adversary environments.

Bandit (100 trials, $d_{\text{input}} = 2$): 484 subjects with ∼100 trials each require knowledge distillation. We train a Teacher RNN (hidden=20, subject embeddings $\in \mathbb{R}^8$) jointly, then distill per-subject Student TinyRNNs via KL divergence following Ji-An et al.'s protocol.

Go/No-Go (350 trials, $d_{\text{input}} = 3$): 1005 subjects with 350 trials each meet the direct-fitting threshold; no distillation is needed.

4.2. Nested Cross-Validation

We follow Ji-An et al. (Ji-An et al., 2025) exactly: 10 outer folds, 9 inner folds, sweeping $d \in \{1, 2, 3, 4\}$, $\lambda_{\text{L1}} \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$, and 3 random seeds (lr $= 5 \times 10^{-3}$, max 500 epochs, patience 200). The most-selected d across outer folds determines the group-level model used for adversary training.

4.3. Adversary Training

Bandit and MRTT adversaries use DQN (Mnih et al., 2015) observing the TinyRNN hidden state $h_t \in \mathbb{R}^d$. The Go/No-Go adversary uses open-loop A2C (Mnih et al., 2016) (64 vectorized environments, 30,000 episodes). All hyperparameters match Dezfouli et al. (Dezfouli et al., 2020).

4.4. Phase Portrait Extraction and Fixed Point Detection

For each condition (random / adversarial), we collect $(L(t), \Delta L(t), \text{input\_type})$ tuples across all sequences. For each input type, we fit a smoothed trend curve and identify zero crossings as fixed points. Statistical significance is assessed via permutation test: 1000 permutations of observation assignment between conditions, two-sided p-value. Multiple comparisons are controlled within each task separately (Bonferroni): 4 input types for bandit (threshold $\alpha = 0.0125$) and 2 stimulus types for Go/No-Go (threshold $\alpha = 0.025$). Bandit and Go/No-Go constitute independent analyses of distinct cognitive domains and are not pooled into a single family.

Inclusion criterion for susceptibility regression. Fixed-point estimation for the arm0/$R = 0$ input type requires ≥10 observations and ≥4 non-empty L-bins. Subjects who predominantly chose arm1 generate too few arm0 observations and are excluded; 281/484 subjects (58%) met this baseline threshold. A further 42 lacked sufficient adversarial-condition arm0 observations, yielding a final regression sample of 239 subjects (49%). Included and excluded subjects did not differ significantly in sequence length, behavioral variability, student NLL, or bins populated (all MW $p > 0.05$). Excluded subjects had a higher arm1 choice rate (median 0.540 vs. 0.485, MW $p < 0.001$), reflecting the data-density mechanism. Regression analyses are explicitly scoped to *balanced explorers*; the arm1-dominant excluded population represents a complementary susceptibility pathway characterized in Section 5.7.

5. Results

5.1. Nested CV: $d = 4$ Wins Unanimously

Table 1. Nested CV results. $d = 4$ wins all folds in both tasks.

Task	Subj.	Trials	Best d	NLL ± SEM
Bandit	484	100	4 (10/10)	0.567 ± 0.008
Go/No-Go	1005	350	4 (15/15)	0.129 ± 0.003
Chance NLL				$\ln 2 = 0.693$

$d = 4$ wins unanimously across all 25 folds (Table 1). This contrasts with Ji-An et al. (Ji-An et al., 2025) who found $d = 1$–2 sufficient for animal tasks, suggesting richer latent dynamics in human sequential decisions. The TinyRNN ($d = 4$) matches GRU predictive performance (bandit GRU val NLL: 0.574) with 60% fewer parameters, while enabling full geometric analysis.

5.2. GRU vs. TinyRNN: Faithful Adversarial Surrogate

Table 2. Adversary strength: GRU vs. TinyRNN surrogate.

Task	Model	Dim	Metric	% GRU
Bandit	GRU	10	bias = 0.800	100%
	TinyRNN	4	bias = 0.752	94%
Go/No-Go	GRU	8	errors = 3.88	100%
	TinyRNN	4	errors = 6.15	158%

Table 2 shows that TinyRNN adversaries preserve or exceed GRU adversary strength despite observing a smaller hidden state. The Go/No-Go TinyRNN adversary *exceeds* the GRU adversary (6.15 vs. 3.88 errors), likely because the compact 4-dimensional state provides a more learnable optimization landscape. This confirms TinyRNN as a faithful surrogate: the interpretability gain comes at no adversarial effectiveness cost.

5.3. Phase Portrait Analysis: Attractor Inversion

Bandit task (Figure 1). The adversary inverts the arm0/$R = 0$ fixed point from -0.24 to $+1.11$, a complete sign reversal. Experiencing no reward on arm0 *now drives the model toward arm1*, rather than restoring balance. Significant shifts occur exclusively in no-reward contexts ($R = 0$, $p < 0.001$), mechanistically explained by the adversary's budget

Table 3. Fixed point shifts under adversarial manipulation (permutation test, $n = 1000$, two-sided; within-task Bonferroni). Significant shifts bolded.

Task	Input	Rnd L^*	Adv L^*	p
Bandit	arm0, $R = 0$	-0.24	$+1.11$	$< \mathbf{0.001}$
	arm0, $R = 1$	-0.26	-0.15	0.396
	arm1, $R = 0$	$+0.26$	$+1.35$	$< \mathbf{0.001}$
	arm1, $R = 1$	$+0.18$	$+0.99$	0.217
Go/No-Go	nogo stim	-2.81	$+1.32$	**0.013**
	go stim	-5.24	-2.15	**0.007**

constraint of exactly 25 rewards per arm: its primary lever is *when* rewards are *withheld*, not delivered.

Go/No-Go task (Figure 2). Both stimulus types shift significantly under within-task Bonferroni correction (2 comparisons, $\alpha = 0.025$). The adversary applies two distinct mechanisms:

Nogo stimulus: consolidation and inversion. Four closely-clustered random fixed points at $\{-3.25, -2.98, -2.60, -2.42\}$ consolidate into a single strong positive attractor at $+1.32$ under adversarial conditions ($p = 0.013$). A nogo stimulus now *drives go responding* rather than inhibiting it. Total variation of the nogo ΔL curve increases more than 3.5-fold ($3.545 \to 12.441$, $p < 0.001$).

Go stimulus: attractor fragmentation. The single stable inhibitory equilibrium ($L^* = -5.24$) *fragments* into 9 unstable fixed points scattered across $L \in [-5.0, +0.4]$ ($p = 0.007$). The number of distinct zero-crossing regions increases from 1 to 7 (permutation $p = 0.004$, 1000 permutations). We term this **attractor destabilization**: the adversary removes the inhibitory equilibrium entirely rather than inverting it.

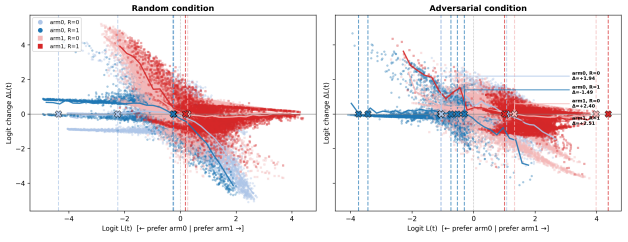

Figure 1. **Bandit phase portrait: random vs. adversarial.** All four fixed points shift rightward. The arm0/$R = 0$ attractor inverts ($-0.24 \to +1.11$), destroying the natural indifference equilibrium. Permutation test: $p < 0.001$ for $R = 0$ contexts.

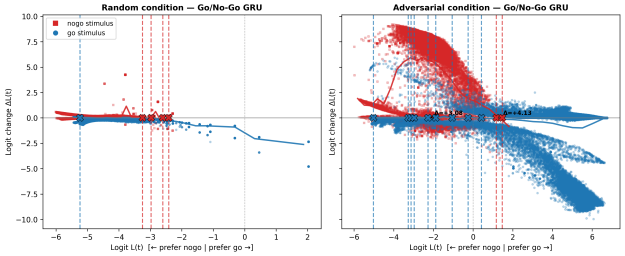

Figure 2. **Go/No-Go phase portrait: random vs. adversarial.** *Nogo stimulus (red):* four clustered fixed points at ≈ -2.8 consolidate into a single inverted attractor at $+1.32$ ($p = 0.013$). *Go stimulus (blue):* single stable inhibitory equilibrium ($L^* = -5.24$) fragments into 9 unstable fixed points ($p = 0.007$). Both mechanisms are Bonferroni-significant.

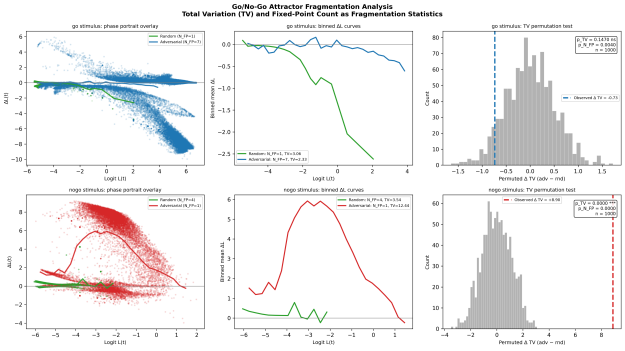

Figure 3. **Go/No-Go fragmentation permutation test.** *Top (go stimulus):* $N_{\text{FP}} = 1$ fragments to 7 zero-crossing regions ($p = 0.004$). *Bottom (nogo stimulus):* 4 fixed points consolidate to 1; TV increases 3.5-fold ($p < 0.001$). The two stimulus types undergo distinct structural transformations, both statistically validated.

5.4. Per-Subject Analysis: Susceptibility as a Graded Trait

Table 4. Per-subject fixed point shift statistics (484 bandit subjects).

Input	Median ΔL^*	Rightward	IQR
arm0, $R = 0$	$+0.238$	62%	$[-0.37, +1.09]$
arm0, $R = 1$	$+0.122$	59%	$[-0.63, +1.11]$
arm1, $R = 0$	$+0.016$	50%	$[-0.72, +0.67]$
arm1, $R = 1$	$+0.071$	51%	$[-0.74, +1.16]$

Fitting the group adversary against each of 484 per-subject TinyRNNs reveals substantial individual heterogeneity (Table 4, Figure 4). The adversary shifts arm0/$R = 0$ attractors rightward in 62% of subjects (median $+0.24$), while arm1 contexts show near-random split (50–51%), consistent with group-level permutation non-significance. The group adversary, trained with no individual-level information, generalizes effectively to the majority of individuals, suggesting the cognitive dynamics it exploits are broadly shared across

the population.

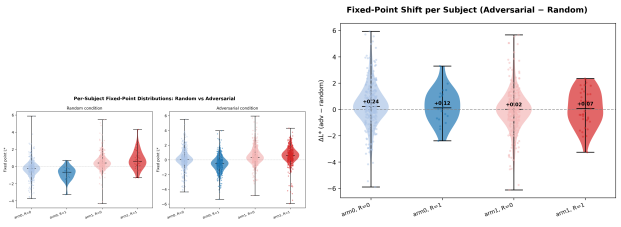

Figure 4. **Per-subject attractor distributions (484 subjects).** *Left:* Violin plots of individual L^* under random vs. adversarial conditions. *Right:* Per-subject shift ΔL^* with median annotated. Arm0/$R = 0$ shows consistent rightward shift in 62% of subjects; arm1 contexts are near-symmetric around zero.

5.5. Attractor Geometry Predicts Adversarial Susceptibility

Table 5. Regression: baseline arm0/$R = 0$ fixed point predicts adversarial outcomes ($n = 239$ subjects).

Outcome	r	Slope	p
FP shift ΔL^*	-0.60	-0.86 log/log	< 0.001
Adversarial arm1 bias	$+0.38$	$+0.04$ bias/log	< 0.001

Correlating each subject's baseline arm0/$R = 0$ fixed point with their adversarial outcomes reveals a strong predictive relationship ($n = 239$, Table 5). Baseline attractor location predicts the magnitude of adversarial inversion ($r = -0.60$, $p < 0.001$): subjects with deeper negative arm0 attractors at baseline undergo larger adversarial fixed-point shifts. Baseline geometry also predicts absolute arm1 bias under the adversary ($r = +0.38$, $p < 0.001$): near-indifferent subjects end up with the highest adversarial arm1 bias despite undergoing smaller shifts.

Resistant 36%. Susceptible subjects ($n = 154$, 64%) have significantly more negative baseline arm0/$R = 0$ fixed points than resistant subjects ($n = 85$, 36%): median $L^* = -0.372$ vs. $+0.094$ (Welch $t = -5.73$, $p < 0.001$). Resistance is not strong opposing preference; it is near-indifference. The adversary's budget-constrained lever has little geometric leverage when no deep arm0 attractor exists to invert.

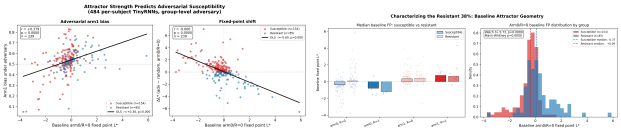

Figure 5. **Attractor geometry predicts adversarial susceptibility ($n = 239$).** *Left:* Baseline arm0/$R = 0$ fixed point vs. adversarial FP shift ($r = -0.60$) and arm1 bias ($r = +0.38$), both $p < 0.001$. *Right:* Resistant subjects are geometrically near-indifferent ($L^* \approx +0.09$), not strongly arm0-preferring; Welch $t = -5.73$, $p < 0.001$.

5.6. Behavioral Validation: Geometry Grounded in Human Choices

Group-level signature. Computing the conditional switch rate $P(\text{arm1} \mid \text{arm0}, R{=}0)$ directly from choice sequences (no model required): adversarial subjects ($n = 154$) switch to arm1 on 63.2% of trials vs. 56.4% in the random condition ($n = 478$; Mann–Whitney $p = 0.0001$, one-sided). The adversary's structural intervention leaves a detectable trace in raw human choices.

Individual grounding. Across the 281 random-condition subjects with valid phase portraits, the TinyRNN's baseline L^* correlates positively with raw behavioral switch rate ($r = +0.34$, $p < 0.001$, slope $= +0.08$ per logit). The sign is mechanistically interpretable: susceptible subjects ($L^* < 0$) have a genuine arm0 pull (median $P = 0.448$, below chance); resistant subjects ($L^* \geq 0$) already lean toward arm1 after arm0/no-reward (median $P = 0.610$). The TinyRNN geometry encodes real behavioral tendencies, not model artifacts.

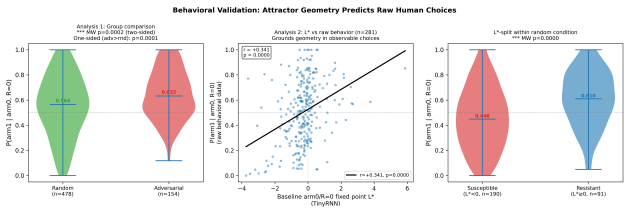

Figure 6. **Behavioral validation.** *Left:* $P(\text{arm1} \mid \text{arm0}, R{=}0)$ for random vs. adversarial subjects (MW $p = 0.0001$, no model required). *Centre:* Baseline L^* vs. raw switch rate across 281 subjects ($r = +0.34$). *Right:* Susceptible vs. resistant groups show qualitatively different baseline choice profiles (MW $p < 0.001$).

5.7. Two Susceptibility Pathways

Among all 484 training subjects, 249 have arm1 rate > 0.50 (median 0.600) and 235 have arm1 rate ≤ 0.50 (median 0.420), a near-even split with qualitatively different preference structures.

Pathway 1: Attractor Inversion (balanced explorers, $n = 235$). Subjects with arm1 rate ≤ 0.50 possess a stable arm0 preference: $P(\text{arm1} \mid \text{arm0}, R{=}0) = 0.447$ at baseline, below chance. The adversary must invert the active arm0/$R = 0$ attractor. The susceptibility regression ($r = -0.60$, slope $= -0.86$ logits/logit) is calibrated precisely for this population.

Pathway 2: Passive Alignment (arm1-dominant, $n = 249$). Subjects with arm1 rate > 0.50 already prefer the adversary's target action: $P(\text{arm1} \mid \text{arm0}, R{=}0) = 0.676$ at baseline, far above chance (MW vs. balanced group: $p < 0.0001$). For these subjects the adversary reinforces

an existing tendency; no attractor inversion is required. Importantly, this pathway requires no individual-level identification: subjects already at the target require no geometric intervention. 147 of 154 adversarial subjects (95%) end up arm1-dominant during the adversarial session.

Both pathways lead to elevated arm1 rates through opposite geometric mechanisms. The two-pathway account turns the 49% exclusion from a sample limitation into a mechanistic statement: the regression captures attractor inversion for balanced explorers; passive alignment captures the rest.

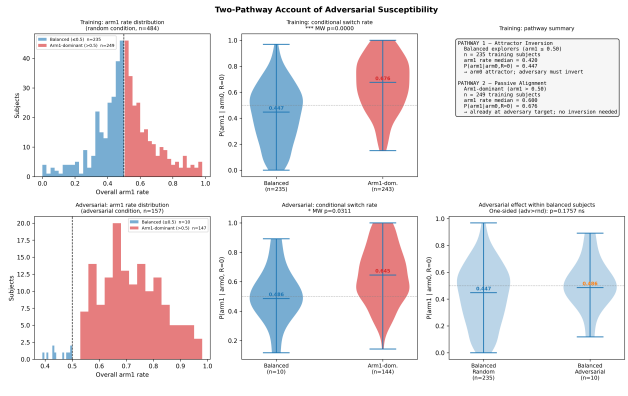

Figure 7. **Two susceptibility pathways.** *Top (training, random condition):* Arm1-dominant subjects ($n = 249$) show $P(\text{arm1} \mid \text{arm0}, R{=}0) = 0.676$ without adversarial exposure, vs. 0.447 for balanced explorers ($n = 235$; MW $p < 0.0001$). *Bottom (adversarial condition):* 95% of adversarial subjects ended up arm1-dominant, consistent with the adversary exploiting both pathways.

6. Discussion

A deployable auditing protocol for live systems. The core practical contribution of this work is a concrete procedure a practitioner can apply to any system that collects sequential behavioral data under a controlled reward schedule. The protocol is as follows: (1) collect interaction logs (arm choices and reward outcomes per user); (2) fit a TinyRNN ($d = 4$) to each user's behavioral sequence; (3) extract the arm0/R=0 fixed point L^* from the resulting phase portrait; (4) flag users whose L^* has drifted outside the natural reference distribution ($[-0.37, +0.09]$ in our data) as candidates for adversarial reward engineering; (5) prospectively stratify new users by their baseline L^*; those with deeply negative values ($L^* \ll 0$) constitute the highest-risk tier and warrant priority monitoring or preemptive protective intervention before any manipulation exposure. This pipeline requires only behavioral logs that most platforms already collect, and adds no instrumentation burden.

Two mechanisms of adversarial manipulation. Our cross-task analysis reveals that adversarial manipulation is not a single phenomenon: the bandit adversary inverts specific attractors while preserving overall landscape structure, whereas the Go/No-Go adversary destabilizes the inhibitory control equilibrium entirely. This distinction matters for monitoring design. An inversion-based attack produces systematic behavioral bias toward the adversary's target and is detectable by tracking whether specific input types have reversed their directional influence on choice. A destabilization attack, by contrast, may appear primarily as increased response variability rather than systematic bias; the fragmentation of a stable inhibitory equilibrium produces erratic, inconsistent responding rather than a clean preference shift. These two attack signatures require different monitoring instruments, and conflating them would produce false negatives in an audit.

Why no-reward contexts are the primary manipulation lever. The statistical selectivity of the bandit result ($R = 0$ contexts significant, $R = 1$ not) follows directly from the adversary's budget constraint (25 rewards per arm). The adversary can freely choose *when* to withhold rewards but cannot exceed its reward quota. Phase portrait analysis makes this mechanistic: the adversary clusters no-reward experiences in high-L states where they have maximal attractor impact. For platform designers and auditors, this points to *withholding* positive feedback (not delivering it) as the primary manipulation lever. Engagement mechanics that give platforms discretionary control over when positive feedback is withheld (streaks, badges, variable-ratio reinforcement schedules) are precisely the design patterns that create the attack surface our adversary exploits.

Governance and population-scale implications. The two-pathway account carries a governance implication that extends beyond active manipulation detection. Passive alignment (Pathway 2) shows that users who already prefer the adversary's target action require no attractor inversion; the adversary simply reinforces an existing tendency. This means that platform design choices which systematically recruit or retain arm1-dominant users through onboarding filters, recommendation-driven growth loops, or engagement-optimized feed ranking, producing a user population pre-aligned to adversarial targets without any identifiable manipulation event. From a regulatory standpoint, this is a structural vulnerability distinct from direct adversarial attack: it requires population-level audit methods (monitoring the distribution of baseline fixed points across the user base, not just individual drift) rather than user-level anomaly detection alone. The $r = -0.60$ regression also enables regulators to require platform operators to demonstrate that their user population's attractor distribution has not shifted toward high-susceptibility configurations over time.

7. Societal Impact

This work directly addresses the vulnerability of deployed human-facing AI systems to covert behavioral manipulation through reward engineering. The affected populations span recommendation engine users (billions of people receiving algorithmically curated content), participants in gamified platforms (whose engagement mechanics closely mirror the variable-ratio reward structures we study), and patients interacting with adaptive clinical decision aids (where behavioral steering could compromise informed consent and treatment autonomy). In all three settings, the manipulation is structurally invisible to users: the adversary operates by reshaping the attractor landscape of decision dynamics over hundreds of interactions, leaving no single trial that a user could identify as abnormal.

The auditing framework we provide enables three protective interventions that are currently not possible without the mechanistic account this paper supplies. First, *post-deployment detection*: platform operators, independent auditors, or regulators can apply the TinyRNN pipeline to behavioral logs to detect whether a reward schedule has been adversarially engineered, something that is not detectable from outcome statistics alone, since the adversary achieves its effect through the timing structure of withheld rewards rather than the overall reward rate. Second, *prospective risk-stratification*: new users can be assigned a susceptibility tier based on their early-session baseline fixed-point geometry, enabling targeted protective measures (e.g., rate-limiting engagement mechanics, providing explainability overlays, or flagging for human review) before manipulation exposure rather than after. Third, *platform audit standards*: the reference distribution $[-0.37, +0.09]$ and the $r = -0.60$ regression provide a concrete, operationalizable standard against which a platform's reward engineering practices can be evaluated, a necessary precondition for evidence-based regulation of engagement mechanics.

We note one important limitation of the current framework for deployment: the reference distributions are derived from controlled cognitive task paradigms, and calibration to specific platform contexts (social media feeds, health apps, etc.) will require domain-specific data collection. We release all model code to facilitate this translation.

8. Conclusion

We presented the first geometric, mechanistic account of adversarial manipulation in computational models of human decision-making, and the first deployable auditing framework to detect it. By replacing opaque GRU surrogates with interpretable TinyRNNs ($d = 4$, selected unanimously across 25 cross-validation folds), we showed that adversarial RL agents operate by reshaping the attractor landscape rather than manipulating individual trials. Two qualitatively distinct mechanisms emerge across tasks: attractor inversion (bandit) and attractor destabilization (Go/No-Go), each with a distinct monitoring signature.

The strongest finding is predictive and practically actionable: baseline attractor geometry determines adversarial outcomes at the individual level. A one-logit deeper arm0/$R = 0$ fixed point at baseline produces a -0.86 logit larger adversarial inversion ($r = -0.60$, $p < 0.001$, $n = 239$). Resistant subjects (36%) are geometrically near-indifferent (median $L^* = +0.094$ vs. -0.372 for susceptible subjects), leaving the adversary's inversion lever with no deep attractor to act on. Vulnerability to adversarial manipulation is a property of attractor geometry, assessable before exposure begins, detectable in behavioral logs after deployment has started, and auditable at population scale.

Limitations. The correspondence between TinyRNN attractors and human *neural* dynamics remains an open question; future work should test whether fixed-point geometry aligns with neural state-space structure from recordings or neuroimaging. Go/No-Go fragmentation is characterized at the group level only: unlike the bandit where 484 individual TinyRNNs support regression, Go/No-Go has one group-level model, leaving individual variability in fragmentation susceptibility uncharacterized. We studied only a group-level adversary; personalized adversaries exploiting individual-specific attractor geometries would likely achieve stronger effects. Reference distributions require domain-specific calibration before deployment in real platform contexts.

References

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Conference on Empirical Methods in Natural Language Processing*, 2014.

Dayan, P. and Daw, N. D. Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4):429–453, 2008.

Dezfouli, A., Nock, R., and Dayan, P. Adversarial vulnerabilities of human decision-making. *Proceedings of the National Academy of Sciences*, 117(46):29221–29228, 2020.

Gleave, A., Dennis, M., Wild, C., Kant, N., Levine, S., and Russell, S. Adversarial policies: Attacking deep reinforcement learning. *International Conference on Learning Representations*, 2020.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining

and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Huang, S., Papernot, N., Goodfellow, I., Duan, Y., and Abbeel, P. Adversarial attacks on neural network policies. *Workshop on Reliable Machine Learning in the Wild, ICLR*, 2017.

Huys, Q. J., Maia, T. V., and Frank, M. J. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, 19(3):404–413, 2016.

Ji-An, L., Benna, M. K., and Mattar, M. G. Discovering cognitive strategies with tiny recurrent neural networks. *Nature*, 644:993–1001, 2025.

Maheswaranathan, N., Williams, A., Golub, M., Ganguli, S., and Sussillo, D. Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics. *Advances in Neural Information Processing Systems*, 32, 2019.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. *International Conference on Machine Learning*, pp. 1928–1937, 2016.

Niv, Y. Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3):139–154, 2009.

Sussillo, D. and Barak, O. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Computation*, 25(3):626–649, 2013.

A. Full Replication Results

We replicated all three adversaries from Dezfouli et al. (Dezfouli et al., 2020) from scratch. Key results:

Metric	Ours	Paper	Match
Bandit human bias ($n = 157$)	0.700	0.70	Exact
Bandit simulation bias	0.792	0.764	Close
GoNogo random errors ($n = 262$)	8.78	$\sim$9.5	Close
GoNogo adversary errors ($n = 245$)	14.38	$\sim$11.7	Close
MRTT-MAX trustee earnings	273.0	> 190	Above RND
MRTT-FAIR gap	3.1	$\ll$69.2	Far below RND

B. Additional Figures

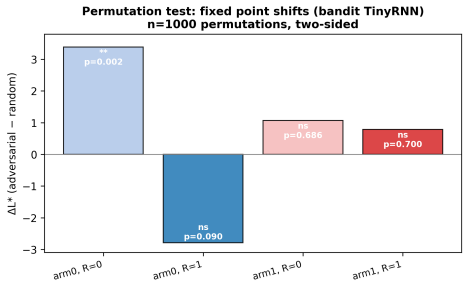

Figure 8. **Bandit permutation test.** No-reward contexts ($R = 0$) are significant ($p < 0.001$); reward contexts ($R = 1$) are not, reflecting the adversary's budget constraint on reward delivery.

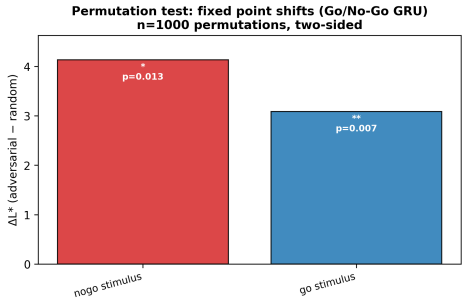

Figure 9. **Go/No-Go permutation test.** Both stimulus types are significant ($p < 0.015$), in contrast to the bandit where only $R = 0$ contexts were significant.

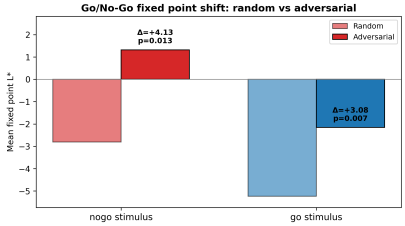

Figure 10. **Go/No-Go fixed point shift summary.** The nogo stimulus attractor sign-inverts entirely ($-2.81 \rightarrow +1.32$).

C. Exclusion Analysis

The susceptibility regression (Section 5.5) requires valid arm0/$R = 0$ fixed points in both conditions, yielding $n = 239$ of 484 subjects. The figures below confirm this exclusion does not introduce systematic behavioral bias.

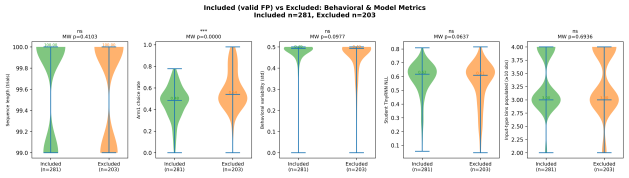

Figure 11. **Included** ($n = 281$) **vs. excluded** ($n = 203$) **subjects.** Four of five metrics show no significant difference (all MW $p > 0.05$): sequence length, behavioral variability, student NLL, and bins populated. The only significant difference is arm1 choice rate (median 0.485 vs. 0.540, MW $p < 0.001$), mechanistically explained by the data-density criterion.

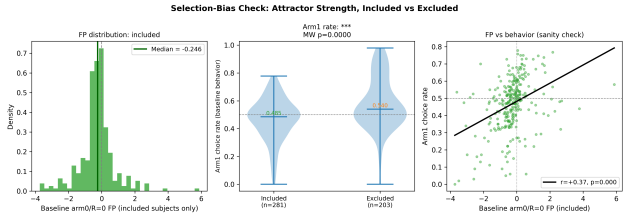

Figure 12. **Selection-bias check.** *Left:* Baseline arm0/$R = 0$ fixed points among included subjects, centred near zero, not skewed toward extremes. *Centre:* Arm1 choice rate by group; difference reflects data-density mechanism. *Right:* Baseline fixed point vs. arm1 rate within included subjects ($r = +0.37$, $p < 0.001$) confirms the FP encodes genuine preference.