# Diffusion Models and Reinforcement Learning : Novel Pathways to Modeling Decoded fMRI Neurofeedback

**Hojjat Azimi Asrari (hazimias@uci.edu)**

**Megan A. K. Peters (megan.peters@uci.edu)**
Department of Cognitive Sciences, University of California, Irvine, United States

## Abstract

**This study explores the application of diffusion models and reinforcement learning to model Decoded Neurofeedback (DecNef), as applied via functional magnetic resonance imaging (fMRI). Our methodology, Denoising Diffusion Policy Optimization (DDPO), integrates diffusion models trained via reinforcement learning to navigate the complex dynamics of brain activity changes. Using a pre-existing DecNef dataset, we implemented policy gradient methods to iteratively refine the diffusion models, aiming to produce target patterns of neural (voxel) activity. Our results demonstrate the potential of this approach for accurately modeling policies that allow the achievement of target brain states, offering a foundation for investigating the mechanisms of neurofeedback and its implications for basic science research and conducting more effective neurofeedback experiments.**

**Keywords:** Decoded Neurofeedback (DecNef), fMRI, Diffusion Models, Explainable Reinforcement Learning

## Introduction

Decoded neurofeedback (DecNef) is an innovative neurofeedback method that uses functional magnetic resonance imaging (fMRI) and pre-trained classifiers to non-invasively induce specific neural activity patterns (LaConte, 2011; Shibata, Watanabe, Sasaki, & Kawato, 2011). This technique enables unconscious control of neural activities, linking them to behavioral outcomes. Widely applied across various behaviors, DecNef demonstrates versatility and significant potential in both fundamental and clinical neuroscience research (Chiba et al., 2019; Li et al., 2021; Oblak, Lewis-Peacock, & Sulzer, 2021).

Despite its successes, however, DecNef can also exhibit inconsistent effectiveness. Several recent lines of research have begun to systematically investigate potential sources of this variability, focusing on neuro-cognitive and psychological mechanisms (Taschereau-Dumouchel, Cortese, Lau, & Kawato, 2021). For instance, (Shibata et al., 2011) introduced a "targeted neural plasticity model", suggesting DecNef induces specific behavioral changes through neuronal plasticity via reinforcement learning (RL), evidenced by increasingly similar brain activity patterns to targets during training, as shown in fMRI studies (Emmert et al., 2016; Shibata et al., 2011). This proposal is consistent with other neurofeedback research, which also links effective training to RL processes (Cortese et al., 2021). Additionally, (Lubianiker, Paret, Dayan, & Hendler, 2022) framed neurofeedback within a RL paradigm, offering a structured way to understand how training protocols could align with reinforcement learning components, helping participants modulate brain dynamics to reach desired outcomes.

Inspired by these, our ultimate goal is to establish a foundation for future work linking policies discovered through RL frameworks to specific neural activity patterns and mechanisms. As a first step, here we used RL to train diffusion models to examine the dynamics and policies that might be learned by the brain (or a human subject volitionally controlling their brain state) in order to achieve a target pattern of neural (voxel) activity in DecNef.

## Methods

We used existing data from one study published as part of the "DecNef collection" (Cortese et al., 2021) to examine whether a diffusion model trained through RL could successfully discover the policies required to transform a current brain state into a target brain state within a given region of interest (ROI), as required in DecNef studies. In this previously published study, 24 human subjects were tasked with producing a target pattern of voxel activity in the target ROI (cingulate cortex). This target pattern was defined in this study as the pattern associated with facial preferences, as identified through training an iterative sparse logistic regression. Data were preprocessed according to standard procedures (motion correction, slice timing correction, etc.) and voxel patterns of activity were extracted from the target ROI according to methods reported previously (Shibata, Watanabe, Kawato, & Sasaki, 2016). Each participant's data thus consisted of ROI voxel activity patterns across $\sim 720$ volumes, and a target voxel pattern of activity in this ROI that had resulted from training the logistic regression classifier.

We assume the changes the subjects (volitionally) cause to the ROI can be mapped to the denoising steps of a diffusion model. In this approach, the denoising sequence is formulated as a Markov Decision Process (MDP), where each state represents a step in the denoising process, actions correspond to the application of denoising steps, and the reward function is tailored to the specific objectives of the diffusion process; in our case, this objective is the same as the decoder used during the real DecNef experiments for each of the subjects. This method is called Denoising Diffusion Policy Optimization (DDPO) (Black, Janner, Du, Kostrikov, & Levine, 2023), a RL approach that employs policy gradient methods to directly optimize diffusion models with respect to the reward function (Figure1A).

To operationalize DDPO, we utilize a policy gradient algorithm that iteratively updates the diffusion model parameters. This algorithm optimizes the expected reward, which is calculated over multiple trajectories of the denoising process. The use of a policy gradient approach is particularly advantageous for this setting as it enables the integration of both immediate and cumulative rewards, thus allowing for the refinement of the model to produce outputs that align closely with the predefined classifer.

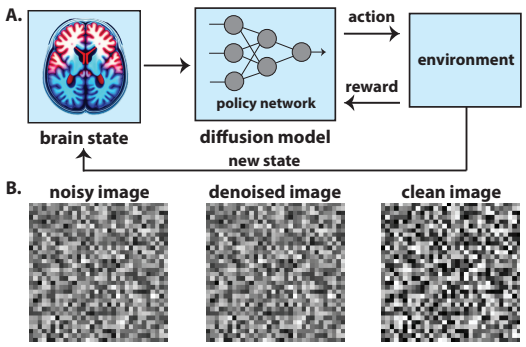


Figure 1: **Model:** A. Denoising Diffusion Policy Optimization (DDPO) uses reinforcement learning to align ROI pattern adjustments with denoising steps in a diffusion model. It frames denoising as states in a Markov Decision Process, optimizing rewards based on experiment objectives through a policy gradient algorithm to enhance model alignment with target classifiers. B. During the initial phase, our model was trained and evaluated on a simulated 32×32 pattern dataset, utilizing Sum of Squared Error (SSE) as the objective function. The model demonstrated progressive improvement in cumulative reward across successive training epochs.

In the DDPO framework, the MDP is formalized to integrate the multi-step denoising process of diffusion models. The MDP formulation is crucial for applying RL techniques to optimize the denoising process in alignment with predefined objectives. Formally, the MDP is described by first defining $s_t \equiv (c, t, x_t)$ as the state at time $t$, with $c$ representing the context and $x_t$ the data state at $t$. The action $a_t \equiv x_{t-1}$ signifies the denoising step from $x_t$ to $x_{t-1}$. The transition probability $P(s_{t+1} \mid s_t, a_t) \equiv \delta(c, t-1, x_{t-1})$ describes a deterministic shift to the next state based on the denoising process. The reward function $R(s_t, a_t) \equiv$ provides $r(x_0, c)$ if $t = 0$ and zero otherwise, focusing on the quality of the initial state $x_0$ in the context $c$.

We apply these equations and learning rules in simple networks with two fully connected layers, with input defined as one subject's voxel pattern of activity on every temporal volume of an acquired fMRI image in the target ROI in the DecNef study used, and the goal state defined as the target voxel pattern of activity in that same ROI according to the ROI-based classifiers trained in that study. We tested whether this diffusion model would be able to learn to satisfy the pre-trained classifier on the ROI across 24 subjects in the DecNef dataset.

## Results and Discussion

In the initial phase, our model underwent training and evaluation using a simulated dataset (random 32×32 patterns), employing the Sum of Squared Error (SSE) as the objective function. The model progressively enhanced its cumulative reward across successive training epochs (Figure 1B).

Subsequently, the model was trained using the dataset provided by (Shibata et al., 2016). Target ROIs across subjects varied in size/dimensions between 219 and 221. A separate model was developed for each of the 24 participants. The learning curve for the model across all subjects is depicted in Figure 2. That each individualized model can successfully learn to produce the individualized target pattern of voxel activity for each subject demonstrates that policies have been successfully discovered for transforming any voxel input image into the target images that satisfy the pre-trained classifier.

The primary objective of this study is to discover a policy network employed in the diffusion model that may in future be mapped to neural activity patterns, e.g. via encoding models or similar. Consequently, we maintained the model's simplicity, opting for a two-layer fully connected network to avoid complexity that might obscure the underlying mechanisms. While this configuration yielded satisfactory performance on test datasets with dimensions under 1000, more complex architectures will likely be required to handle higher dimensional data. For such scenarios, we plan to implement more complex models similar to those described in (Orouji et al., 2022; Orouji & Peters, 2022), which have demonstrated promising capabilities in learning latent space representations of real fMRI data. Ultimately, our approach has the potential to map learned policies for successful—or unsuccessful—DecNef to brain states and networks, significantly enhancing our understanding of when and how DecNef can be used to successfully modulate brain states in awake, behaving humans.

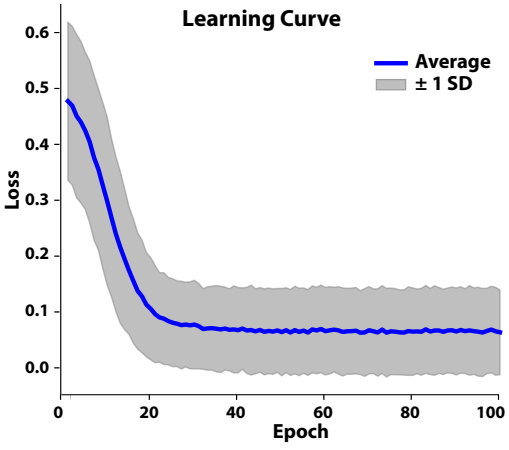


Figure 2: **Model Learning Curve on Real Data:** Following initial training, the model was further developed using (Shibata et al., 2016)'s dataset, with separate models tailored for each of the 24 participants based on their specific data.

## References


Black, K., Janner, M., Du, Y., Kostrikov, I., & Levine, S. (2023). Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*.

Chiba, T., Kanazawa, T., Koizumi, A., Ide, K., Taschereau-Dumouchel, V., Boku, S., ... others (2019). Current status of neurofeedback for post-traumatic stress disorder: a systematic review and the possibility of decoded neurofeedback. *Frontiers in human neuroscience*, *13*, 448460.

Cortese, A., Tanaka, S. C., Amano, K., Koizumi, A., Lau, H., Sasaki, Y., ... Kawato, M. (2021). The decnef collection, fmri data from closed-loop decoded neurofeedback experiments. *Scientific data*, *8*(1), 65.

Emmert, K., Kopel, R., Sulzer, J., Brühl, A. B., Berman, B. D., Linden, D. E., ... others (2016). Meta-analysis of real-time fmri neurofeedback studies using individual participant data: How is brain regulation mediated? *Neuroimage*, *124*, 806–812.

LaConte, S. M. (2011). Decoding fmri brain states in real-time. *Neuroimage*, *56*(2), 440–454.

Li, L., Wang, Y., Zeng, Y., Hou, S., Huang, G., Zhang, L., ... Zhang, Z. (2021). Multimodal neuroimaging predictors of learning performance of sensorimotor rhythm up-regulation neurofeedback. *Frontiers in Neuroscience*, *15*, 699999.

Lubianiker, N., Paret, C., Dayan, P., & Hendler, T. (2022). Neurofeedback through the lens of reinforcement learning. *Trends in Neurosciences*, *45*(8), 579–593.

Oblak, E., Lewis-Peacock, J., & Sulzer, J. (2021). Differential neural plasticity of individual fingers revealed by fmri neurofeedback. *Journal of Neurophysiology*, *125*(5), 1720–1734.

Orouji, S., & Peters, M. (2022). Extracting task-relevant low dimensional representations under data sparsity.

Orouji, S., Taschereau-Dumouchel, V., Cortese, A., Odegaard, B., Cushing, C., Cherkaoui, M., ... Peters, M. A. (2022). " task-relevant autoencoding" enhances machine learning for human neuroscience. *arXiv preprint arXiv:2208.08478*.

Shibata, K., Watanabe, T., Kawato, M., & Sasaki, Y. (2016). Differential activation patterns in the same brain region led to opposite emotional states. *PLoS biology*, *14*(9), e1002546.

Shibata, K., Watanabe, T., Sasaki, Y., & Kawato, M. (2011). Perceptual learning incepted by decoded fmri neurofeedback without stimulus presentation. *science*, *334*(6061), 1413–1415.

Taschereau-Dumouchel, V., Cortese, A., Lau, H., & Kawato, M. (2021). Conducting decoded neurofeedback studies. *Social Cognitive and Affective Neuroscience*, *16*(8), 838–848.