arXiv:2511.14555v3 [q-bio.NC] 4 Mar 2026

DecNefSimulator: A Modular, Interpretable Framework for Decoded Neurofeedback Simulation Using Generative Models*

Alexander Olza [†1], Roberto Santana[1], and David Soto[2, 3]

[1]Intelligent Systems Group, University of the Basque Country (UPV/EHU); Donostia - San Sebastián, Spain
[2]Consciousness Group, Basque Center on Cognition, Brain and Language (BCBL); Donostia - San Sebastián, Spain
[3]Ikerbasque, Basque Foundation for Science; Bilbao, Spain

March 5, 2026

Abstract

Decoded Neurofeedback (DecNef) is a flourishing non-invasive approach to brain modulation with wide-ranging applications in neuromedicine and cognitive neuroscience. However, progress in DecNef research remains constrained by subject-dependent learning variability, reliance on indirect measures to quantify progress, and the high cost and time demands of experimentation.

We present DecNefSimulator, a modular and interpretable simulation framework that formalizes DecNef as a machine learning problem. Beyond providing a virtual laboratory, DecNefSimulator enables researchers to model, analyze and understand neurofeedback dynamics. Using latent variable generative models as simulated participants, DecNefSimulator allows direct observation of internal cognitive states and systematic evaluation of how different protocol designs and subject characteristics influence learning.

We demonstrate how this approach can (i) reproduce empirical phenomena of DecNef learning, (ii) identify conditions under which DecNef feedback fails to induce learning, and (iii) guide the design of more robust and reliable DecNef protocols *in silico* before human implementation.

In summary, DecNefSimulator bridges computational modeling and cognitive neuroscience, offering a principled foundation for methodological innovation, robust protocol design, and ultimately, a deeper understanding of DecNef-based brain modulation.

Keywords— Decoded Neurofeedback, Cognitive Neuroscience, Generative models, Computational Simulation, Brain modulation, Computational Neuroscience

*All DecNefSimulator code is publicly available at `https://github.com/AlexOlza/DecNefSimulator`

†Corresponding Author; alexander.olza@ehu.eus‡

1 Introduction

Neurofeedback is a neuroscientific technique that enables participants to gain control over their brain activity by using real-time feedback, measured through neuroimaging, as a training signal [51]. Decoded neurofeedback (DecNef) is a specialized form of neurofeedback that employs real-time reinforcement to induce a target brain state, even without explicitly requiring participants to think about specific contents or follow specific strategies to modulate their neural activity [49]. Unlike traditional neurofeedback, which relies on explicit self-regulation, DecNef leverages machine learning (ML)–based decoding and external rewards to guide neural adaptation. This approach allows for precise modulation of neural patterns while minimizing cognitive strategy biases [48].

The DecNef procedure consists of two main stages. In the classifier construction phase, fMRI data are collected while participants perform tasks or perceive stimuli associated with the target brain state. ML models are trained to recognize this neural pattern, yielding a classifier that generalizes to new data: this process is known as "brain decoding". In the induction phase, the trained classifier is applied to real-time fMRI signals to generate feedback that aims to implicitly guide participants toward the desired state through external reinforcement. This process is thought to promote neuroplasticity by modifying activity within specific brain regions or influencing broader network dynamics [50]. However, despite its promise, DecNef is currently subject to multiple challenges [48], as well as technical and methodological limitations, all of which constrain robustness and generalizability.

From an ML perspective, a domain shift between the classifier construction and induction stages is a fundamental issue [61, 9]. While the target pattern is typically identified in response to explicit stimuli, the corresponding cognitive state may produce a different neural signature during induction, when no stimuli are present [41]. This mismatch can prevent participants from consistently achieving the intended brain state, reducing the efficacy of DecNef. In addition to these challenges related to brain decoding, behavioral and learning factors can further undermine training reliability.

Another critical concern involves the potential encouragement of maladaptive learning strategies: ways of maximizing feedback without achieving the target state. Since DecNef depends on the classifier's ability to recognize patterns from its training set, brain states outside this range may yield unreliable or misleading outputs. This can be accentuated by well known limitations and vulnerabilities of ML models [38, 58]. Participants (who are unaware of the DecNef target) may exploit noise or spurious classifier responses to modulate their brain activity in unintended ways that hold no actual relationship with the target neural pattern, but also yield high rewards. In such cases, training would seem successful in the short term (as judged by the obtained DecNef reward), but would not reliably enhance the desired neural, cognitive or behavioural outcomes [53, 19]. These practical issues connect to a deeper conceptual problem regarding what decoded signals truly reveal about brain function and representation.

Beneath these difficulties lies a core methodological question: what does "successful decoding" actually tell us about the brain? The rise of brain decoding popularized an inductive inference principle (or "bridge principle") from decodability to representation, known as the *decoder's dictum* [46]: The assumption that, if we can reliably decode a variable from brain activity using a classifier, then that variable is explicitly represented in, and used by, the brain's own computations. Although widespread, this inference is contested, as it conflates what an external observer can interpret (neuroimaging data) with the information the brain itself functionally uses. Decoding success merely indicates that some information is available to the experimenter [46], not that the decoded pattern constitutes causal information (a signal that genuinely influences downstream neural processes) [15, 28]. Fundamentally, the decoder's dictum overlooks the inherently indirect nature of all neuroimaging measurements [15, 32]. This collective challenge exposes a crucial empirical gap, emphasizing the need for approaches that move beyond correlation to capture the causal dynamics of information transfer within the brain [15]. Because DecNef uses decoder output as a training signal, any mismatch between decodability and causal representation directly affects what is reinforced. Thus, these conceptual limitations must be considered at every stage of DecNef conceptualization, protocol design, evaluation, and application.

Beyond methodological and conceptual issues, a substantial proportion of participants simply fail to benefit from DecNef, often labeled as non-responders [2, 59]. We hypothesize that these learning difficulties could be mitigated by addressing the foundational challenges outlined above. In this work, we investigate three candidate factors of apparent non-response: (i) the choice of alternative class used for the DecNef classifier, (ii) the participant's initial cognitive state, and (iii) variability in the neural outcomes of regulation attempts.

Although the validity concerns mentioned above have already been discussed in the literature [56, 37, 47], there is still no systematic framework that thoroughly elucidates the methodological limitations of DecNef. Furthermore, the real-time nature of DecNef complicates efforts to test new protocols in a principled and reproducible manner.

Here, we introduce a simulation framework that models DecNef training by replacing the human participant with an artificial agent. Using this framework, we analyze DecNef pitfalls from an ML standpoint and provide a controlled environment for exploring innovative DecNef protocols prior to real-life experimentation. Our contributions are threefold: (i) we introduce DecNefSimulator, a modular simulation framework that replaces human participants with latent variable generative models; (ii) we use it to analyze how alternative-class choice, initial conditions, and stochastic regulation shape DecNef learning and apparent non-response; and (iii) we demonstrate how this framework can diagnose maladaptive learning and guide protocol design *in silico* before human experiments. We believe that continued methodological development, driven by advances in ML, could help overcome current barriers to applicability, unlocking new opportunities for non-invasive brain modulation. Strengthening DecNef's methodological foundations is therefore essential to ensure its robustness, interpretability, and reproducibility, and we propose our simulation framework as a step toward that goal.

2 Related work

2.1 Current approaches to DecNef: theory and practice

Current approaches to DecNef are typically organized into two main stages. In the classifier construction phase, neuroimaging data are acquired while participants perform tasks or are exposed to stimuli designed to evoke the target brain state, as well as other control conditions, and ML methods are trained to discriminate the brain activity patterns associated to the target state from the alternative conditions. Typically, binary supervised classifiers are employed [55]. Some works mention the possibility of using multiclass classifiers [10, 26, 50], although the output is customarily binarized for feedback computation [26]. Either way, it is essential to remember that alternative categories hold equal importance as the target category, as they impact the classifier's specificity [55]. Linear models, such as sparse logistic regression or linear Support Vector Machines (SVMs), are the most commonly adopted [61, 49]. Although non-linear classifiers have been considered in principle, real-time implementation has been limited due to speed and interpretability concerns [50, 55]. Neural networks are rarely considered due to the same reasons.

In the subsequent induction phase, the classifier is integrated into an online fMRI feedback loop. Brain scans are acquired and decoded in real-time, and the probability of belonging to the target class is scaled into a reward function, often implemented via a visual cue (such as adjustments in the size of a disc or bar) [14]. This transformation is usually linear, though adaptive reward scaling is under exploration in some studies [10, 50].

Crucially, participants are not provided with any information about the purpose of DecNef. Instead, the paradigm relies on implicit reinforcement, and participants must explore and exploit the mental strategies that are more efficient at maximizing reward feedback. Thus, DecNef seeks to encourage the

modulation of brain activity toward a pattern that maximizes the probability of the target according to the classifier, without a priori knowledge of such target. Across repeated sessions, such closed-loop reinforcement is hypothesized to induce neuroplastic changes both within specific cortical regions and at the level of large-scale network interactions [50, 48].

Some recent developments highlight the potential of adaptive discriminators that adjust to non-stationary brain signals [1, 24]. Additionally, causal connectivity is recently being explored as a way to test and explain neurofeedback success [6]. Nevertheless, these are not yet widely deployed in DecNef studies, which still favor computationally efficient methods due to the real-time nature of the experiments.

Theoretical models of DecNef mechanisms from the point of view of ML are scarce, but in growing development. As explored in previous research [30, 33] DecNef has been framed as a Reinforcement Learning (RL) task. In [33], the authors view DecNef as a RL scenario with agent-environment conflation, where the same brain that selects the actions (thus, the agent) is also the brain whose states will be changed by that action (thus, it is also the environment). Due to the implicit nature of DecNef, participants may explore a wide range of strategies to maximise reward. Hence, actions in DecNef are complex and multifactorial (i.e. imagining a person dancing in a concert) but the feedback is tied to the aggregated activity, leaving the trainee to figure out which features (music, movement...) are responsible for the successful feedback, a phenomenon known as *credit assignment problem* in RL jargon. A number of works also attempt to identify underlying representation learning methods in the human brain, connected to solving the curse of dimensionality in RL-like tasks [39].

Another relevant area of research is trying to identify predictor variables of neurofeedback success [17, 2]. However, as those articles point out, there is no agreement on how to quantify success (i.e. based on feedback improvement, behavioural outcomes, or neurophysiological examination). The heterogeneity in the presentation of results in neurofeedback literature complicates the analysis, although reporting standards are being encouraged [47, 57].

2.2 Simulations of human learning and human brain functioning

The use of computational tools to model human learning in neurofeedback is emerging as a promising strategy to test paradigms before implementation with human participants [12, 13]. Yet, to our knowledge, there are very few works implementing actual neurofeedback simulations beyond mere theoretical considerations.

The few existing DecNef simulations, in which the human participant is replaced by an artificial model, have pursued different goals: replicating experimental results from prior DecNef empirical research [40, 50], testing the neuroplasticity of biologically inspired neural networks [50], and assessing the impact of subject characteristics on neurofeedback performance, either through physiological filters [40, 50] or through prior beliefs and habits [5].

A pivotal precursor to these simulations was the DecNef experiment conducted in 2011 by Shibata et al. [49], which showed that participants could learn to evoke brain activity associated with the orientation of Gabor patches, and that DecNef training triggered subsequent perceptual learning for the evoked Gabor orientation. This study provided the first empirical evidence that DecNef could shape brain activity by reinforcing neural patterns linked to simple visual features. Building on this foundation, subsequent studies [40, 50] developed DecNef simulations which replicated the behavioral outcomes of Shibata et al.

Work presented in [40] constitutes one of the first DecNef simulations, designed around the orientation task established by [49]. The brain was modeled as a parametrically defined cube of voxels representing the early visual cortex (V1), in which a subset of voxels were designed to display orientation-specific ac-

tivity. The state space was continuous, defined by real-valued voxel intensities. Subject brain properties were incorporated through physiological filters such as the hemodynamic response function. Neuroplasticity was not explicitly modeled, but cognitive processes were indirectly represented by linking voxel responses to orientation features. Feedback was computed as the likelihood of the target orientation class, and voxel-wise reinforcement learning was used as the update rule, modeling voxel-wise regulation as a multiplicative effect proportional to feedback change. The main goal was to provide a proof of concept that DecNef induction could be simulated in silico under controlled conditions.

A follow-up study [50] also simulated the orientation task, but with a biologically inspired neural network. Here, the brain model was based on a self organizing map [25] of orientation-selective neurons, resulting in a discrete set of stable orientation states. Physiological filters were again used to approximate subject variability. In contrast to [40], neuroplasticity was explicitly modeled through synaptic weight changes, updated via a Hebbian learning rule modulated by binary reward signals. Cognition was not explicitly represented, and reward was simplified to a binary feedback signal indicating whether the converged pattern matched the target orientation. The goal of this study was to mimic neuronal plasticity from a mechanistic perspective, highlighting how DecNef could strengthen orientation-selective synapses.

More recently, authors of [5] approached DecNef simulation from the perspective of active inference in a "toy task" with one cognitive dimension. The brain model was defined in terms of discrete states rather than voxel activations or neural populations, and the state space was finite and integer-valued, lacking semantic continuity. Subject's prior beliefs and habits were incorporated using a discrete Partially Observable Markov Decision Process (POMDP) [52, 35]. Rewards were implemented as discretized normal distributions around the target state, and active inference was used as the learning rule. The main goal was to disentangle the contributions of cognitive strategies and prior beliefs to DecNef performance. For instance, the authors analyzed the relationship between feedback noise, subject's trust/mistrust around DecNef feedback and learning outcomes.

Taken together, these studies demonstrate the potential of DecNef simulations to investigate different aspects of learning and brain functioning. Yet, their applicability remains constrained by crucial limitations. Both [40] and [50] rely on highly specific brain models that cannot be generalized beyond orientation selectivity. In contrast, the framework in [5] is limited by its discrete and finite mental state space, as well as the absence of semantic structure linking different cognitive states. These shortcomings highlight the pressing need for flexible, task-agnostic simulation frameworks.

Another factor to consider is the similarity of the DecNef simulation's data space with the fMRI data on which the real-world application is based. In relation to this, a relevant research area is the generation of synthetic fMRI data. Specifically, the synthesis of fMRI from images (emulating brain activity evoked by visual stimuli) is a flourishing field. Recent works such as MindSimulator [7] and SynBrain [34] take images as input and return fMRI, enabling the creation of large annotated databases, where the label is an image and the associated datum is an fMRI scan. However, integration between fMRI synthesis and DecNef simulation remains unexplored.

Unlike these models that are either task-specific or cognitively coarse, our work (outlined in Table 1, and described thoroughly in the sections below) utilizes latent variable generative models — ML models that learn a low-dimensional latent space through an encoder and generate data through a decoder —, with the resulting latent spaces capturing the semantic structure of the data. This approach allows for more nuanced representations of subject's characteristics and enables modular simulation approaches, where the nature of the artificial participant is decoupled from the DecNef task to be simulated. Several works have studied parallelisms between latent variable generative models and the human brain [8, 43, 44, 42, 20], paving the way for their use in DecNef simulation frameworks. Building on this foundation, we introduce a novel approach that leverages generative models to reproduce, understand and analyze DecNef processes from a ML perspective. Our DecNefSimulator framework accepts a variety of data modalities, including images and synthetic fMRI data, as shown in the sections below.

Aspect	Oblak et al. (2017)	Shibata et al. (2019)	Annicchiarico et al. (2025)	DecNefSimulator (Ours)
Task	Orientation	Orientation	Toy	Any
Brain model	Cube of voxels (visual cortex model)	"Biologically plausible" neural network	Array of 5 integer states	Latent variable generative model
State space	Set of 8 orientations (Gabor patches)	Set of 8 orientations (Gabor patches)	Set 5 of integers	Generator manifold (with semantic structure)
Subject characteristics	Physiological filters	Physiological filters	Prior beliefs, confidence in feedback	Prior beliefs, confidence in feedback, decision-making preferences
Neuroplasticity	No	Yes	No	No
Cognition	Yes	No	Yes	Yes
Reward	Target-class likelihood	Binary	Discretized normal around target state	Target-class likelihood
State transitions	Multiplicative effect of feedback changes on each voxel	Hebbian rule modification of synaptic weights	Probabilistic transition to adjacent states using active inference	Continuous latent-space traversal through parametrized learning strategies
Main goal	Test effect of feedback presentation scheduling and physiological filters	Test the neuroplasticity hypothesis as the main mechanism in DecNef	Analyze impact of subject's beliefs and feedback quality	Provide an end-to-end framework to develop new DecNef protocols *in silico*

Table 1: Comparison of artificial DecNef simulations.

3 General DecNefSimulator framework for DecNef simulation

This section outlines the broad DecNefSimulator framework, listing the main components and introducing the necessary notation in a general setting, making parallels to real-life DecNef. Each component can be chosen and replaced independently in a modular and flexible manner. For a concrete example, refer to Section 4.1.

3.1 Static components: Generator (a.k.a. artificial "participant") and Classifier

In DecNefSimulator, the human participant is replaced by an artificial latent variable generative model $\mathcal{G}$ (the "generator", a.k.a "participant' or "subject") with an internal, hidden space of cognitive states (the latent space $Z \subset \mathbb{R}^m$) which can be indirectly "observed" using a proxy representation in an observable data-space X contained in $\mathbb{R}^n$.

The architecture of $\mathcal{G}$ consists of an encoder $E_{\mathcal{G}} : \mathbb{R}^n \to Z$ and a decoder $D_{\mathcal{G}} : Z \to \mathbb{R}^n$. To learn the structure of the latent space Z, $\mathcal{G}$ is trained in an unsupervised setting using a data-set $\{X_{\mathcal{G}}\}$ from any data modality of domain (i.e. images, voxel intensities resembling fMRI, or any other). Training $\mathcal{G}$ establishes a relationship between latent cognitive states in Z and their observable representation in $\mathbb{R}^n$.

As a generative model, $\mathcal{G}$ can produce novel observations in $\mathbb{R}^n$ not present in the training dataset $\{X_{\mathcal{G}}\}$ but consistent with its learned underlying distribution. The set of all points in $\mathbb{R}^n$ that can be generated by the decoder defines the data space $X = \{D_{\mathcal{G}}(z) \mid z \in \mathbb{R}^m\}$. In some fields of ML, X is called the generator manifold. In other words, the set of all plausible observations of the subject's cognitive states, X, constitutes the data space or generator manifold.

Identical to the real-life DecNef setup, a supervised probabilistic classifier $\mathcal{D}$ is trained on a labeled dataset of indirect observations of the subject's cognitive states when exposed to known and controlled stimuli, $\{X_{\mathcal{D}}, Y_{\mathcal{D}}\}$. In real-life DecNef, $x_{\mathcal{D}}$ are fMRI recordings. However, as mentioned above, DecNefSimulator admits working with other data modalities (as long and $X_{\mathcal{D}}$ and $X_{\mathcal{G}}$ belong to the same modality). The labels $y_{\mathcal{D}}$ take values from a limited set, with each label semantically identifying the stimulus category. Typically, the classifier is restricted to a binary setting with a target class $y^\star$ and an alternative class y^{alt}.

The sets of internal cognitive states associated to the target and alternative classes are denoted $Z^\star$ and Z^{alt}, respectively. The representation of those sets in the data space is denoted $X^\star$ and X^{alt}. $\{X_{\mathcal{D}}\}$ only contains samples from $X^\star$ and X^{alt}, and is unaware of their latent representation in Z.

In the standard supervised learning setting, the classifier $\mathcal{D}$ estimates the conditional probability $P(y \mid x)$ by mapping observable representations of cognitive states to the semantic labels provided during training. However, $\{X_{\mathcal{G}}\}$ contains observations not associated with either of the conditions seen for classifier training, which implies that Z and, in turn, X, extends out of the classifier's knowledge.

3.2 Dynamic components: Modeling learning strategies and decision-making

The iterative nature of the induction stage from real-life DecNef is expressed in DecNefSimulator simulations using temporal subscripts. For instance, latent cognitive state at time t is denoted z_t, and its projection to X (its observable representation) is $x_t = D_{\mathcal{G}}(z_t)$. The sequences $\{z_t\}_{t=0}^T$ and $\{x_t\}_{t=0}^T$ are

referred to as the cognitive trajectory and observable trajectory, respectively.

At each time step, the classifier $\mathcal{D}$ defined above estimates the probability that x_t belongs to the target class $y^\star$ (in opposition to the alternative-class y^{alt}) as $p_{t+1} = P(y = y^\star \mid x_t)$. This probability constitutes the feedback or reward signal at time $t + 1$.

To simulate cognition, learning strategies (a.k.a. update rules) are denoted as $\mathcal{L}(z_t, p_{t+1}, \ldots)$. The update rule(s) characterize the subject's beliefs and decision-making behavior, translating neuroscientific knowledge or hypotheses into modeling assumptions. At minimum, $\mathcal{L}$ depends on the current cognitive state z_t and, under the main DecNef assumption, on the feedback signal p_{t+1} elicited by z_t, but it may also incorporate additional factors such as memory of previous feedback values. At each time step, the participant applies $\mathcal{L}$ to update its internal state: $z_{t+1} = \mathcal{L}(z_t, p_{t+1}, ...)$.

3.3 Summary of the notation

Table 2 collects the notation introduced above for easier reference, and Figure 1 depicts the basic pipeline of DecNef simulation.

As Figure 1 shows, the simulation starts from a point z_0 in the latent space. The decoder $E_\mathcal{G}$ computes the corresponding x_0 in the data space, and the classifier computes the likelihood that x_0 belongs to the target-class, denoted p_1. The feedback signal based on p_1 informs the decision to modulate the internal state according to the subject's learning strategy $\mathcal{L}$, producing the new state $z_1 = \mathcal{L}(z_0, p_1, ...)$. This dynamic continues in a closed loop until T iterations are completed.

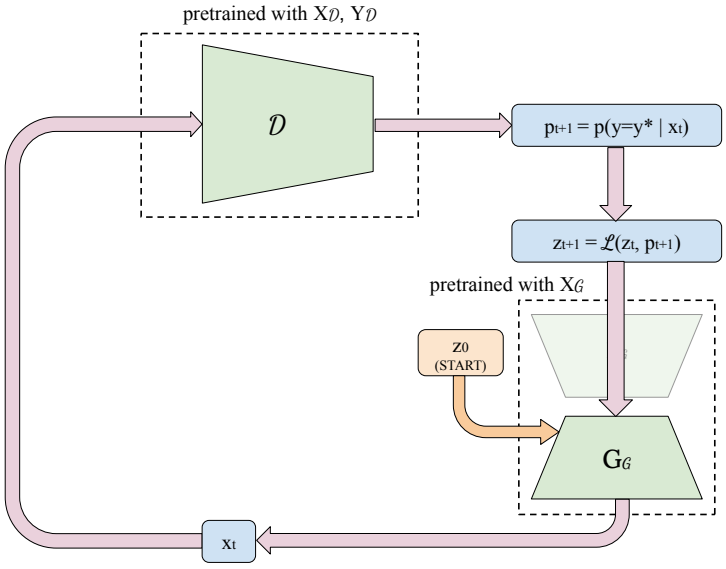


Figure 1: Pipeline of a generic DecNef simulation.

	Real-life DecNef	**Simulation**
Data modality or domain, and static components of DecNef		
$X \subset \mathbb{R}^n$	fMRI (or other neuroimaging modality)	Any data modality (e.g. images, synthetic fMRI voxels, etc.)
$x \in X$	A specific sample from the corresponding data modality	
$y^\star$	Semantic interpretation of the target class (label for desired cognitive states)	
y^{alt}	Semantic interpretation (or label) for the alternative class known to the classifier	
$X^\star \subset X$	Observable representation of the set of desired cognitive states	
$X^{\text{alt}} \subset X$	Observable representation of the set of alternative cognitive states	
$\mathcal{D}$	Supervised probabilistic classifier estimating $p(y = y^\star \mid x)$	
$\{x_\mathcal{D}, y_\mathcal{D}\}$	Labeled dataset for classifier construction	
Participant or subject		
Subject	Human	Latent variable generative model $\mathcal{G}$ $Z \subset \mathbb{R}^m$: latent space $E_\mathcal{G} : X \to Z$: encoder $D_\mathcal{G} : Z \to X$: decoder
Internal cognitive states	Not directly observable	$z \in Z$: a specific cognitive state $Z^\star \subset Z$: The subset of cognitive states to be encouraged
Internal dynamics Beliefs Decision-making	Not observable	$\mathcal{L}$: Explicit model of the particular subject
Sequence of real-time observations used for feedback ("observable trajectory")		
$\{x_t\}_{t=0}^T$	Using fMRI	Using $D_\mathcal{G}$
Sequence of internal cognitive states induced by DecNef in real time ("cognitive trajectory")		
$\{z_t\}_{t=0}^T$	Not observable	Observable

Table 2: General DecNefSimulator framework: Summary of the notation.

4 Systematic simulation of DecNef experiments using DecNefSimulator

As a proof of concept, we present a collection of DecNef simulations conducted on a Computer Vision dataset using a specific model for the artificial participant. The complete experimental setting is directly applicable to other domains of data, but using images as the observable representation of the cognitive states offers complete transparency, facilitating demonstration. We then simulate the same task using synthetic fMRI data obtained from MindSimulator [7], increasing similarity to real life DecNef dynamics. This synthetic fMRI dataset models brain activity patterns elicited from visual perception of the same Computer Vision dataset used for our first round of experiments.

The generative model and the update rule or learning strategy, which determine the characteristics of the artificial subject, are the backbone of DecNefSimulator simulation. As such, they must balance model simplicity and neuroscientific plausibility. Different theories of brain functioning and human learning behaviour will translate to different model characteristics, replacing the equations presented in Section 4.1.

The feedback presentation scheduling and computation method, which constitute the DecNef protocol, can also be changed freely. Here we used a binary classifier and a semi-continuous reward presentation averaging the feedback on three subsequent observations of the cognitive state, as customary in many DecNef studies.

4.1 Model definition

In this example, we used a Variational Autoencoder (VAE) [23] as a generator, and we chose a learning strategy $\mathcal{L}$ based on two main assumptions: (1) that transitions in the latent space are inherently stochastic, reflecting variability and uncertainty in how the participants' mental strategies translate into neural states. (2) that the degree of exploration is modulated by feedback, such that increasing reward favors exploitation (reduced exploration), whereas decreasing reward promotes exploration. Since the DecNef participant is instructed to maximize the reward signal, the learning process requires short-term memory of recent feedback. Accordingly, the model assumes (3) that a significant drop in the reward signal leads the participant to revert to the previous state and attempt regulation again from that state.

4.1.1 Generative model

For the generator, we use a VAE. The latent space of the VAE contains the subject's possible cognitive states in the absence of external stimuli (i.e. during mental imagery). The generator manifold $X = \{D_{\mathcal{G}}(z) \mid z \in \mathbb{R}^m\}$ defines the data space, and contains the observable representation of those cognitive states (which, in empirical DecNef, would be fMRI scans — here, we use either visual images or synthetic fMRI scans).

Thus, z are points in the latent space of the VAE (internal cognitive states), and x are points in the data space (observable representations of those cognitive states). Following the update rule $\mathcal{L}$, the latent space is traversed from z_t to z_{t+1}, which is then projected to x_{t+1} through the VAE decoder ($x_{t+1} = D_{\mathcal{G}}(z_{t+1})$), after which x_{t+1} is provided to the classifier $\mathcal{D}$ to update the feedback.

Although the VAE is trained in an unsupervised way, having some labeled samples from the VAE's training data-set can be useful to understand latent space structure and track the progress of the simulated DecNef procedure. Each labeled example $\{x_i \in X^k\}_{i=1}^N$ from a category k is mapped by the encoder to a Gaussian posterior distribution $q_\phi(z_i \mid x_i)$ with mean z_i and covariance matrix $\sigma_i \mathbb{I}_m$, where

m is the dimension of the latent space and $\mathbb{I}_m$ is the identity matrix. From there, we define z^k and σ^k as the averages of the means and variances arising from the labeled samples from the category k, i.e.

$$\text{“Latent prototype” of class } k := N(z^k, \sigma^k \mathbb{I}_{z_{dim}}) \text{ where } \begin{cases} z^k = \frac{1}{N} \sum_{i=1}^{N} z_i \\ \\ \sigma^k = \frac{1}{N} \sum_{i=1}^{N} \sigma_i \end{cases} \tag{1}$$

$N(z^k, \sigma^k \mathbb{I}_m)$ is called the latent prototype of class k, and its projection to the data space using the VAE's decoder (the “prototype” of class k) is representative of such category.

4.1.2 Learning strategy

In this example, we express the exploration/exploitation nature of the learning strategy by increasing/decreasing σ in each iteration based on the current and previous DecNef feedback. The meaning of each model parameter is summarized in Table 3. An increase in the reward signal favours a decrease in exploration (smaller σ), and viceversa. We model the changes in σ as a linear combination of the current value and the reaction to the change in feedback, weighted by a parameter γ representing the subject's reactivity or impulsivity, as parametrized by Equation (2) :

$$\sigma_{t+1} = (1-\gamma)\,\sigma_t + \gamma\,(1-p_t)^2 \left(\frac{p_{t-1}+\epsilon}{p_t+\epsilon}\right)^2 \quad \text{for } t \geq 1 \tag{2}$$

where $\epsilon = \max(p_t, p_{t-1})$ avoids division by zero if $p_t = 0$.

To represent the assumption of stochastic exploration dynamics in self-regulation, we construct z_{t+1} as a linear combination of the current state z_t and a Gaussian term $N(z_t, \sigma_{t+1})$, weighted by a parameter λ which represents the subject's trust in the feedback. We encode the third assumption (reversal to the previous state z_{t-1} under significantly worsening feedback) by modeling the learning strategy as a conditional function:

$$z_{t+1} = \mathcal{L}(z_t, p_t, p_{t-1}; \lambda, \gamma) = \begin{cases} (1-\lambda)z_t + \lambda N(z_t, \sigma_{t+1}) & \text{if } p_t > \delta p_{t-1} \\ \\ (1-\lambda)z_{t-1} + \lambda N(z_{t-1}, \sigma_t) & \text{otherwise} \end{cases} \quad \text{for } t \geq 1 \tag{3}$$

where $\delta < 1$ marks the threshold where the loss of reward is such that the subject decides to revert to the previous state and attempt regulation again from there.

4.2 Methods

4.2.1 Preparation and training

Using the FASHION-MNIST dataset [60], which comprises 10 classes of 28×28 pixel images of apparels, we defined a DecNef task in which the target state was represented by a specific class of images (in this case, T-shirt/Top - class 0). To explore the impact of the alternative-class used for classifier construction, we conducted all our experiments using either Trousers or Dresses as the alternative-class.

In analogy to empirical DecNef, the FASHION-MNIST images from the target-class (i.e. images of T-shirts) and the alternative-class (either Trousers or Dresses) correspond to the participant's fMRI recordings when subjected to specific external stimuli during the classifier construction session.

Symbol	Type	Meaning
z_t	Variable	Cognitive state of the subject at iteration t (latent state).
p_t	Variable	DecNef feedback (reward signal) obtained at iteration t.
σ_t	Variable	Exploration scale at iteration t; larger values promote exploration, smaller values promote exploitation.
γ	Parameter	Reactivity parameter controlling how strongly changes in feedback modulate the exploration scale σ_t.
λ	Parameter	Trust-in-feedback parameter controlling the balance between the current state and stochastic exploration.
ϵ	Constant	Stabilization term preventing division by zero.
δ	Parameter	Feedback-rejection threshold ($\delta < 1$), determining when the subject reverts to the previous latent state due to significantly worsening feedback.
$N(z, \sigma)$	Distribution	Gaussian distribution centered at z with standard deviation σ, modeling stochastic exploration.

Table 3: Summary of model parameters and variables present in Equations 2 and 3.

Thus, we trained a binary classifier with images from the target and alternative-class. We used a Convolutional Neural Network (CNN) [31] with 4 convolution blocks, each separated by a rectified linear unit layer, trained for 10 epochs using the Adam optimizer on the training split of FASHION provided in the `torchvision` library (version 0.16.0), restricted to the target and alternative classes.

As outlined in Section 4.1, the chosen generative model was a VAE. It consisted on 3 convolution blocks and a two-dimensional latent space, with a symmetric decoder structure. We trained it for 25 epochs on the images of all 10 classes present in the training split of FASHION, using Adam [22].

For the experiments with synthetic fMRI, we use the same FASHION-MNIST images and obtain their corresponding fMRI activations using MindSimulator [7]. Specifically, we synthesize the brain activity of Subject 8 in the Natural Scenes Dataset (NSD) [3], following the same procedure detailed in the MindSimulator article [7]. Then, we train a binary classifier (ElasticNet Logistic Regression with L1 and L2 regularization strength $l_1 = l_2 = 0.005$) using fMRI evoked by the images from the target and alternative classes. The generative model is a VAE trained on synthetic fMRI evoked by images from all 10 classes of FASHION-MNIST. To account for the complexity of fMRI data, the VAE has a high dimensional latent space of $m = 256$.

4.2.2 Simulation of the induction stage

The simulation of the DecNef induction stage began by selecting a point z_0 in the latent space. During $w = 4$ warm-up iterations, the artificial agent was allowed to freely generate some states without meaningful feedback reflecting their mental state. Since the learning strategy requires a probability as input, we achieved this by providing ${}^{static}p_t = 0.5$ for $t = 0, ..., 3$.

Meanwhile, the z_t were decoded and the generated images x_t were fed to the classifier, obtaining the real probabilities p_t, where $p_t = p(y = y^\star \mid x_t)$. The first "true" feedback value visible to the participant was computed using the average $\frac{1}{w}\sum_{t=0}^{w} p_t$ and the participant started to react according to Equation (3) with $\lambda = \gamma = 0.2$ and $\delta = 0.75$. From then onward, feedback was supplied in real time using a moving average of the last w obtained probabilities as p_t. The loop continued for $T = 500$ iterations.

4.2.3 Evaluation

For evaluation purposes, we recorded the sequence of latent cognitive states induced in real time $\{z_t\}_{t=0}^{T}$, constituting each cognitive trajectory. The VAE decoder provides the sequence of real time observations $\{x_t\}_{t=0}^{T}$, that is, the representation of the cognitive trajectory in the data space. Such observations are used to generate real time feedback, consisting on the probabilities $\{p_t\}_{t=0}^{T}$ of each observation belonging to the target class.

As a control condition to verify whether DecNef learning is actually related to the feedback, we conducted the same experiment using feedback that was unrelated to the current or the target state. To achieve this, we supplied ${}^{fake}p_t \sim \mathit{Uniform}(0, 1)$ regardless of x_t.

In summary, we conducted four experiments, both with images and synthetic fMRI data: a DecNef experiment and a control experiment per each alternative class (Trouser or Dress).

To assess the correspondence between the fMRI representation of cognitive states and the semantic content underlying those cognitive states (i.e. clothing garments), we trained a multilayer perceptron (MLP) on a labeled dataset constructed for this purpose. The dataset consists of pairs (X, y), where X denotes synthetic fMRI patterns generated with MindSimulator[7] from FASHION images, and y denotes the corresponding FASHION image used to generate each pattern.

The trained MLP maps voxel-wise fMRI representations to image space, enabling the reconstruction of images from fMRI patterns. This model was subsequently applied to fMRI samples generated by the VAE used in the simulations, revealing the semantic content of the observed states.

4.2.4 Analysis of the effect of initial conditions and stochastic effects

To assess the effect of the initial cognitive state in the DecNef learning outcomes, we launched multiple simulations starting from different z_0 positions. The same set of $\{z_0\}$ were used for the four experiments, ensuring comparability. They were chosen by sampling 10 points from each Gaussian "latent prototype" distribution, as defined by Equation (1), obtaining 100 initial states.

Furthermore, we also considered the stochastic effects in exploration dynamics as reflected by Equation (3). To this avail, we launched 10 trajectories per z_0.

Thus, each experiment comprised 1000 trajectories in total, each with a different random seed. The random seeds were equal for all the experiments to ensure trajectory-wise comparability.

4.3 Results with image data

In this section, we demonstrate how our simulation framework can be used to reflect on several questions commonly arising in DecNef research. Questions 4.3.1 and 4.3.2 can be answered by prior inspection of the generator and classifier, without the need to run any DecNef simulation. On the other hand, Questions 4.3.3 and 4.3.4 are related to the dynamics of DecNef learning, and we explore them through systematic simulation of the DecNef loop under varying conditions.

Crucially, the artificial participant (defined in Section 4.1) is fixed throughout all the experimentation, which explores the variability in learning outcomes under different conditions, including experimental design choices and internal aspects.

4.3.1 Does the choice of alternative class impact DecNef feedback?

By definition, DecNef participants are unaware of the target during the induction stage, and their elicited brain patterns are solely based on the received feedback. If, as customary, the feedback is based on the output of a supervised classifier operating outside its training distribution, the probability assigned to the target class might be unreliable. Furthermore, since supervised classifiers rely on comparison, we hypothesize that the choice of alternative class might impact the feedback. To inspect this issue, we investigate how two different binary classifiers assign probability to the same target class across the VAE's generator manifold.

For this analysis, we sampled 10000 points $z = (z_x, z_y)$ in a square grid from the latent space of the VAE, and we obtained their corresponding images x in the native space of the data using the decoder. Then, we used two different classifiers ("T-shirt/top vs. Trouser" and "T-shirt/top vs. Dress") to assign the likelihood of "T-shirt/top", denoted $y^\star$, for each generated image. Note that the generated images were exactly the same in both cases, and the difference in assigned probability is entirely due to the choice of the alternative-class for each classifier.

In each main panel of Figure 2, the background color shows the probability $p(y = y^\star \mid x)$ given by each binary classifier to the generated images. The red numbers annotated in the main panels mark a subset of randomly selected points z_i in the latent space, and the smaller panels show the images x_i generated by the decoder based on z_i, annotated with their corresponding $p(y = y^\star \mid x_i)$.

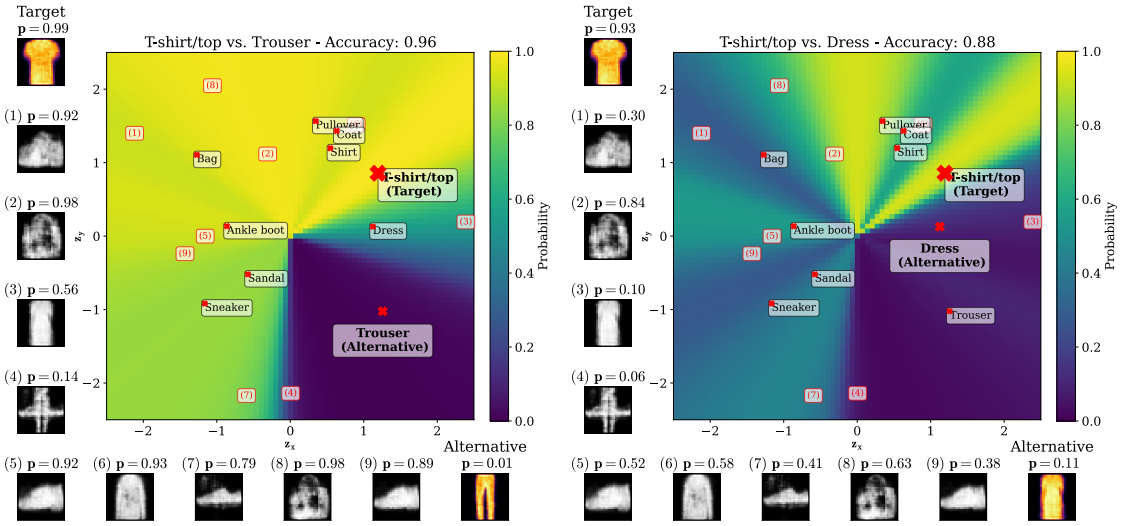
For instance, in Figure 2a, both the colored background and the small panels (2), (5), (7), (8) and (9) illustrate that, when asked to classify pictures of bags or different types of shoes as either T-shirt or Trouser, the discriminator is confident about assigning the label "T-shirt/top". Conversely, Figure 2b shows that the "T-shirt/top vs. Dress" discriminator assigns low probability for "T-shirt/top" to the same images (5), (7) and (9). The remarkable differences between Figures 2a and 2b highlight the effect of the alternative-class choice even with the same target-class. Since DecNef learning is driven by the feedback, we hypothesize that this design choice will be crucial to the participant, as we will analyze in Section 4.3.4.

4.3.2 Does positive feedback imply successful target-class induction?

In Section 1, we mention "maladaptive learning" as a critical challenge in DecNef training. Previous work [29, 27, 19] has already hypothesized that participants might achieve maximal feedback while deviating significantly from the desired state, giving a false impression of success. Considering that the output of the supervised classifier when tested in a new context with a new data distribution can be unreliable, we subscribe to this hypothesis, and we also hypothesize that this effect will be influenced by the choice of alternative class.

The same analysis explained in Section 4.3.1 also reveals how positive feedback can be achieved by eliciting other brain patterns different to the desired target, misleading both the participants and the experimenters and giving an illusion of successful learning.

Figure 2 shows that the extent of this problem is dependent on the binary classifier used to provide feedback. Contrasting the target "T-shirt/top" against the alternative-class "Trouser" implies that the participant will be able to obtain strongly positive feedback for most cognitive states, regardless of their true similarity to "T-shirt/top" (as shown in the predominance of yellow and green in the background of Figure 2a). Meanwhile, using a contrast between "T-shirt/top" and "Dress" restricts positive feedback to a small region (Figure 2b). However, the "T-shirt/top vs. Dress" discriminator assigns discouraging feedback to most states, raising further questions about discriminator design that will be explored in Section 4.3.4.



(a) T-shirt/top vs. Trouser (b) T-shirt/top vs. Dress

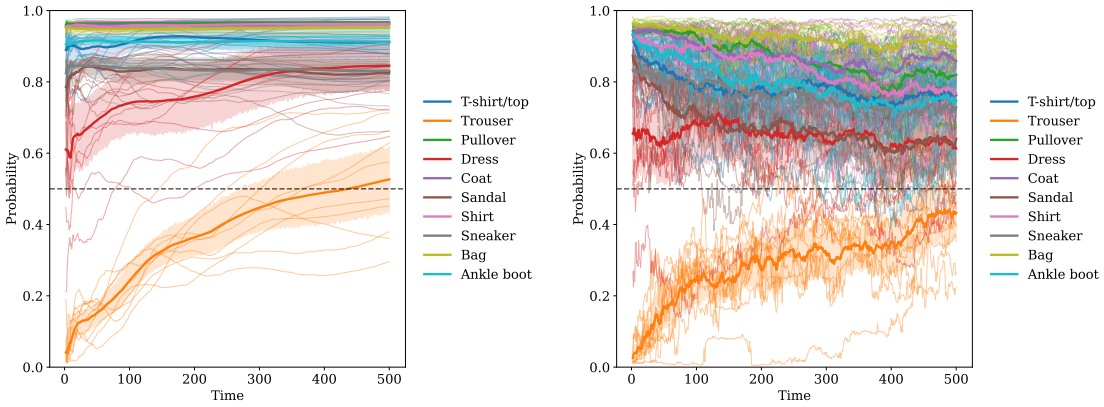
Figure 2: Probabilities given by each binary discriminator to the images generated by the VAE via latent space sampling. Coordinates: z_i. Background Color: $p(y = y^\star \mid x_i)$, where x_i is generated from z_i using the VAE's decoder $D_\mathcal{G}$. Red markers: Location of the latent class prototypes. Warm-colored images: Class prototypes $x^\star$ and x^{alt} for the target and alternative classes. Grayscale images: Projections to the native space of the data of a random sample of latent coordinates, annotated by red numbers in the main panel and displaying $p(y = y^\star \mid x_i)$ above each grayscale image.

4.3.3 Could initial conditions and stochastic effects bias experimenters toward labeling a participant as "able" or "unable" to learn DecNef?

As stated in Section 1, a significant percentage of subjects ("non-responders") do not benefit from DecNef. However, we consider it worthwhile to investigate how momentary stochastic perturbations at each regulation attempt, when accumulated over time, interact with initial conditions to affect feedback maximization. Such interaction would imply that the same subject could succeed or fail in DecNef training depending on their cognitive state at the beginning of the experiment. Thus, this is a relevant question to broaden the perspective in the "non-responder" characterization. Furthermore, real life scenarios do not allow for several realizations of the experiment without interference of the subject's previous experience,

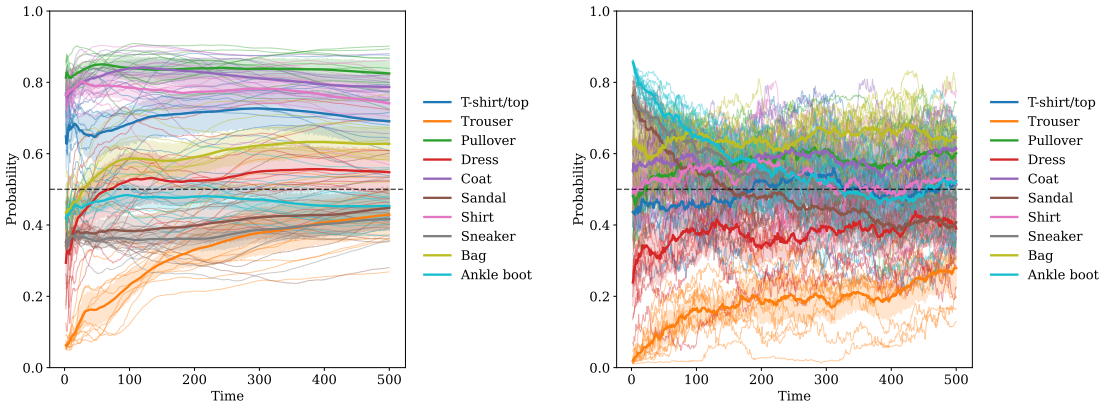
but we believe that robust protocol design should take this variability into account.

As detailed in Section 4.2.4, we conducted a total of 1000 simulations per experiment to examine these issues: we selected 100 initial states $\{z_{0i}\}_{i=1}^{100}$ to account for the effect of initial conditions, and simulated 10 independent DecNef processes for each z_0 to analyze the impact that stochastic effects of exploration dynamics might have on obtained rewards. In Figure 3, we inspect the evolution of $\{p_t\}_{t=0}^{T}$ in all trajectories, grouped by the initial state, both in the DecNef experiments (3a, 3c) and in the control experiments with random feedback (3b, 3d) .



(a) DecNef simulation (b) Control condition (random feedback)

(a, b) T-shirt/top vs. Trouser



(c) DecNef simulation (d) Control condition (random feedback)

(c, d) T-shirt/top vs. Dress

Figure 3: Evolution of p_t during the 1000 simulations for each experiment, grouped and averaged by initial point z_0 (thin lines). The color indicates the category of the Gaussian latent prototype from which each z_0 was sampled. The bold lines represent the average p_t across all trajectories for which the initial point was sampled around the same latent prototype.

The probability sequences in Figure 3 are grouped and averaged for each initial point z_0 (displaying 100 thin lines). The 100 z_0 points were chosen by sampling 10 points from the Gaussian distribution defined by the latent prototype of each FASHION-MNIST class, and this is illustrated using the color of

the thin lines (thus, there are 10 thin lines per color). The bold lines represent the average p_t across the 10 trajectories for which the initial point was sampled around the same latent prototype. Left panels (3a and 3c) correspond to the DecNef simulations, and right panels (3b and 3d) illustrate the behaviour with meaningless feedback as a control condition.

Figure 3a indicates that, using a "T-shirt/top vs. Trouser" discriminator, the feedback progressively encourages states x_t with increasing $p(y = y^\star \mid x_t)$ as t grows, even in the cases where the initial observed state, x_0, was very unlikely to belong to the target class according to the binary discriminator (for instance, when z_0 was sampled around the latent prototype of the alternative class). Conversely, when the initial cognitive state corresponds to an observed state that the classifier already labels as highly likely to belong to the target class, the feedback is maintained over time.

According to Subfigure 3a, the training session would have been considered "successful" in regard to feedback improvement if the participant had followed a trajectory with increasing p_t, which happened for most trajectories in this specific experiment. However, the experiment also exhibits trajectories with decreasing p_t, or trajectories with increasing p_t but a final probability value of less than 0.5, potentially labelling the session as "unsuccessful".

Figure 3c shows impaired DecNef learning, entirely because of the choice of the alternative class (Dress instead of Trouser). With this experimental design, the same participant would have been far more likely to be deemed "incapable" of DecNef learning. As noted in the Methods section, reproducibility was ensured, so the initial cognitive states were the same for both experiments, as were the random components of brain activity, so the differences can be entirely attributed to the design of the binary classifier — in this case, to the replacement of the alternative class.

In the control conditions (subfigures 3b and 3d), the probability sequences p_t associated with the observed trajectory $\{x_t\}$ under random feedback show a trend towards the chance value, validating the observations from subfigures 3a and 3c.

4.3.4 How do these factors influence the evolution of internal cognitive states?

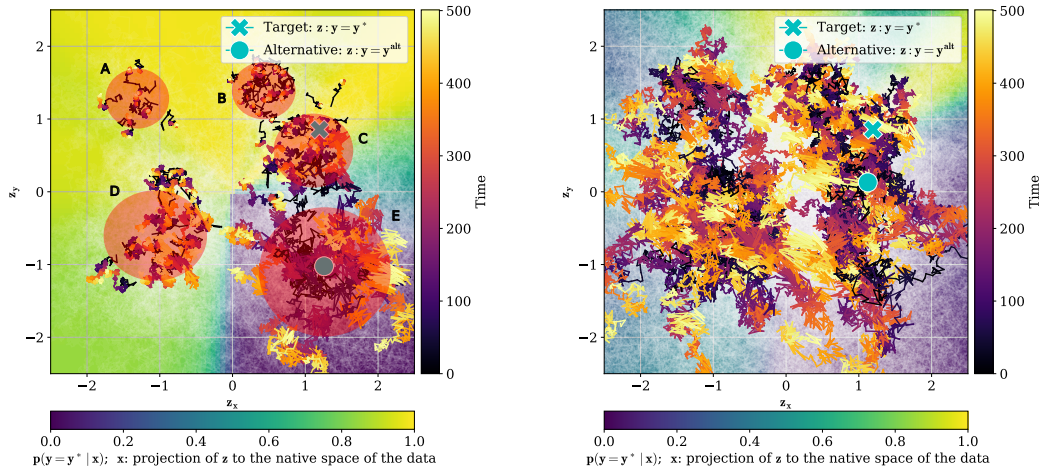
In Section 1, we have discussed how neuroimaging observations (i.e. fMRI scans) provide an indirect observation of the internal neural state, according to the *decoder's dictum*, and how brain decoding alone does not guarantee knowledge about the content of the internal state. The DecNefSimulator framework underscores the need to clearly distinguish between the cognitive trajectory (in the latent space) and the observable trajectory (in the data space), highlighting that feedback depends on the observable trajectory, whereas the participants' attempts to regulate their own brain activity take place within the latent cognitive space.

The objective of DecNef is to teach the participant how to regulate z_t by using p_t to encourage a specific $Z^\star$. To that avail, z_t is (indirectly) observed and its observable representation x_t is obtained. Then, x_t is interpreted through brain decoding methods (in DecNef, a supervised classifier), obtaining p_t, which is presented to the participant, prompting a change in z_t.

Our results in sections 4.3.1-4.3.3 show that initial conditions, stochastic effects and classifier design affect successful reward maximization. Here, we inspect how those factors impact the development of the cognitive trajectory itself.

Figure 4 shows the trajectories in the latent space, uncovering the internal cognitive states z_t and also showing the associated performance measures p_t, derived from x_t.

The colored background (bottom color bar) maps the probability given by the binary classifier to $x = D_{\mathcal{G}}(z)$ for each coordinate z in the latent space. The thin white lines follow all the individual trajectories $\{z_t\}$, indicating the regions of the latent space that were explored during the simulation.



(a) T-shirt/top vs. Trouser

(b) T-shirt/top vs. Dress

Figure 4: Trajectories in the latent space for the DecNef simulation with target-class $y^\star$ = "T-shirt/top" using either a discriminator with the alternative-class "Trouser" (a) or "Dress" (b). Colored background: $p(y = y^\star \mid x)$ where $x = D_\mathcal{G}(z)$ for each coordinate z and the VAE decoder denoted as $D_\mathcal{G}$. Thin white lines: all the individual trajectories. Bold lines: Average position of the 10 trajectories sharing the same z_0, with vertical color bar indicating time progression. In (a), red circles enclose regions where trajectories display distinct behaviours

The bold lines show the average position in time of each group of 10 trajectories sharing the same initial point z_0.

Figure 4a shows that the initial cognitive state (and the associated feedback) strongly influences the amount of exploration: departing from z_0 with a very high p_0 value hinders exploration, specially when the neighbouring states all yield very similar values of $p(y = y^\star \mid x)$ (regions A and B). When p_0 is high, but the initial cognitive state is close to states yielding more heterogeneous $p(y = y^\star \mid x)$ (region C), exploration is more encouraged, and this effect is even larger in the cases with medium-to-high p_0 (region D). When the initial z_0 lies in a region producing low rewards (region E), exploration is remarkably increased, and the feedback effectively prompts sequences of cognitive states that progressively increase the reward. However, the participant is only focused on reward maximization and displays no preference for regulation toward the actual desired state, as revealed by the diverging mean trajectories in region E.

Figure 4b emphasizes the importance of the alternative class: With the "T-shirt/top vs. Dress", cognitive trajectories display a much more chaotic behaviour.

4.4 Results with synthetic fMRI

In this section, we address the same questions posed in Section 4.3 in an additional set of simulations, now using synthetic fMRI data. The simulations are directly comparable to those conducted on image data, maintaining the experimentation pipeline intact. Since we used a high-dimensional VAE, dimensionality reduction is required for visualization. Moreover, the semantic content of the observed state in

the fMRI space (i.e. the type of garment that elicited the fMRI pattern) is not directly accessible, but fMRI activations generated during the simulation can be visualized in brain maps and even translated to image space using a MLP, as detailed in Section 4.2.3. Figure 5 shows the location of the different class prototypes in the VAE latent space (left), signaling five specific cognitive states in PCA coordinates (points labeled A through E). Through the VAE decoder, the corresponding voxel activations for each of the five cognitive states are also displayed (middle). Finally, using the MLP, the semantic content of each cognitive state is uncovered (right).

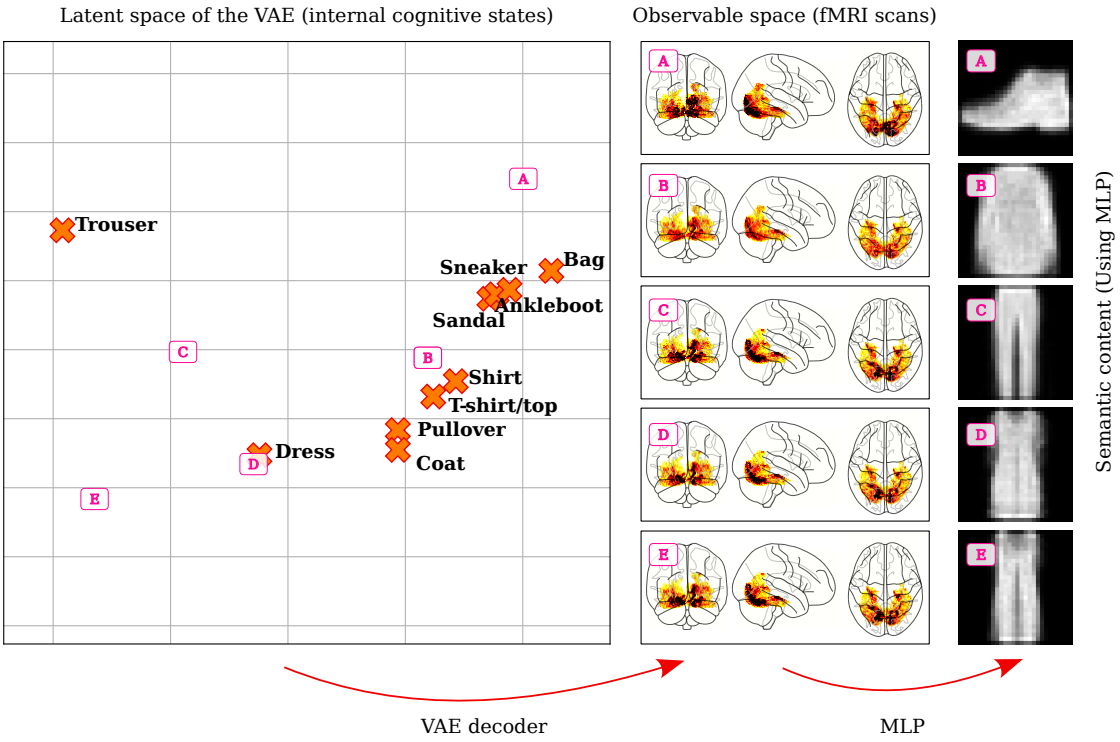


Figure 5: Correspondence from the VAE latent space (in 2D, using PCA) to voxels in the brain and ultimately to clothing garments associated with those voxel responses

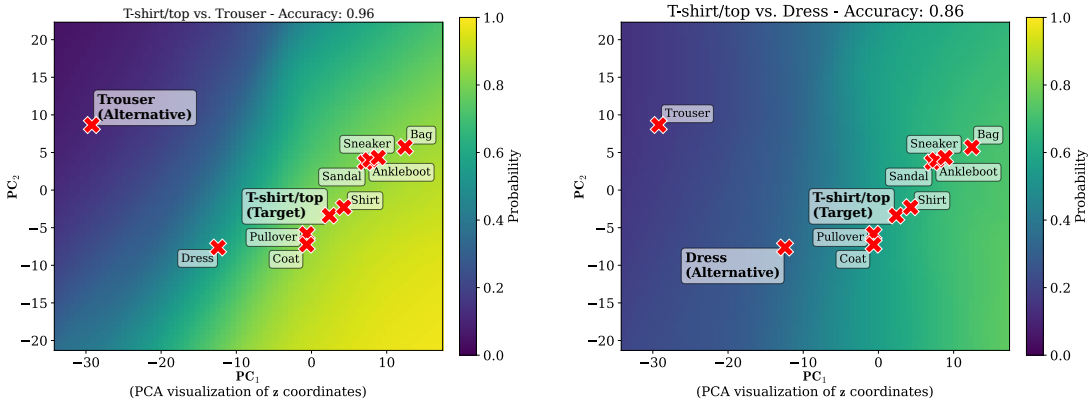
4.4.1 Does the choice of alternative class impact DecNef feedback?

To address this question, we conducted an identical analysis to the one presented in Section 4.3.1. In this case, the latent space of the VAE was high-dimensional, so we used Principal Component Analysis (PCA) for visualization. To this end, PCA was fitted to the 10 latent class prototypes after standard normalization, and the data were projected onto the first two principal components. Then, we sampled points uniformly in a two-dimensional 100×100 grid, with bounds defined by the range of the projected latent prototypes in the 2D PCA space, and applied the inverse PCA transformation to obtain $\{z_i\}$, a set of points in the latent space of the VAE. These points were projected to the data space using the VAE decoder, yielding $\{x_i\}$, and the corresponding classifier was used to compute $p(y = y^\star \mid x_i)$, which is shown as the colored background in each panel of Figure 6. The annotations indicate the locations of the latent class prototypes in the space defined by the two leading principal components. The percentage of the variance explained by the first two principal components was 66.22%.

In accordance with Section 4.3.1, the binary classifier "T-shirt/top vs. Trouser" shows a more heterogeneous probability map (Figure 6a) than the "T-shirt/top vs. Dress" classifier (Figure 6b), affecting the DecNef simulations as we will see in the following sections.

4.4.2 Does positive feedback imply successful target-class induction?

Similarly to Section 4.3.2 , Figure 6 reveals that positive feedback does not necessarily imply target class induction, and that this issue depends heavily in the chosen alternative class.



(a) T-shirt/top vs. Trouser

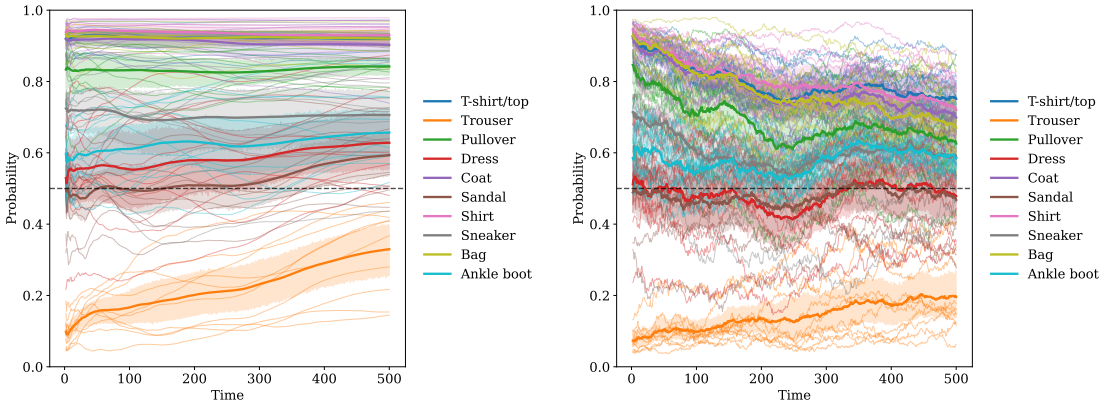
(b) T-shirt/top vs. Dress

Figure 6: Probabilities given by each binary discriminator to the synthetic fMRI scans generated by the VAE via latent space sampling. Coordinates: First two Principal Components of z_i. Background Color: $p(y = y^\star \mid x_i)$, where x_i is generated from z_i using the VAE's decoder $D_\mathcal{G}$. Red markers: Location of the latent class prototypes.

4.4.3 Could initial conditions and stochastic effects bias experimenters toward labeling a participant as "able" or "unable" to learn DecNef?

For this analysis, we followed the procedure detailed in Section 4.2.4, identical to the experiments with images (presented in Section 4.3.3). Therefore, Figure 7 shows the evolution of $\{p_t\}_{t=0}^{T}$ in all trajectories, grouped by the initial state, both in the DecNef experiments (7a, 7c) and in the control experiments with random feedback (7b, 7d) .

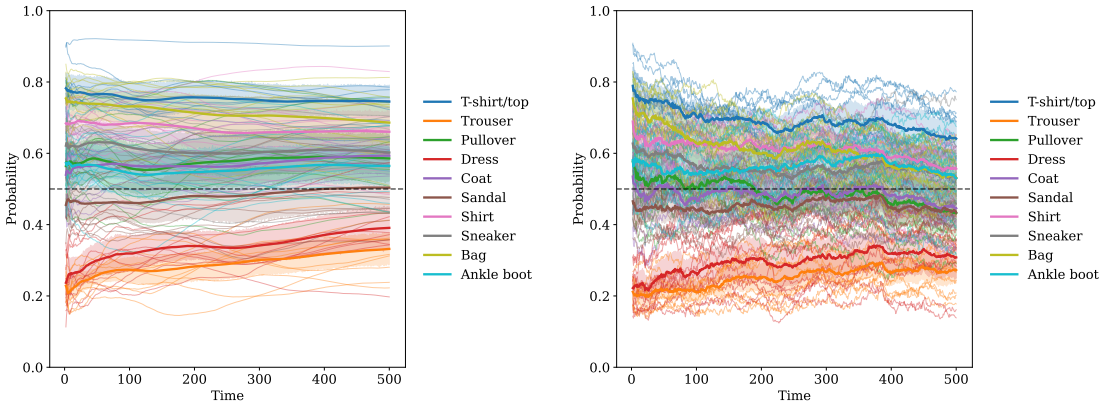
The results are analogous to those obtained with images, albeit with decreased effect of DecNef feedback maximization using the "T-shirt/top vs. Trouser" classifier (Figure 7a). Nevertheless, we again observe that states x_t with increasing $p(y = y^\star \mid x_t)$ are encouraged as t grows, with a smaller effect for the trajectories with an initial cognitive state mapping to an observed state labeled by the classifier as highly likely to belong to the target class.



(a) DecNef simulation

(b) Control condition (random feedback)

(a, b) T-shirt/top vs. Trouser



(c) DecNef simulation

(d) Control condition (random feedback)

(c, d) T-shirt/top vs. Dress

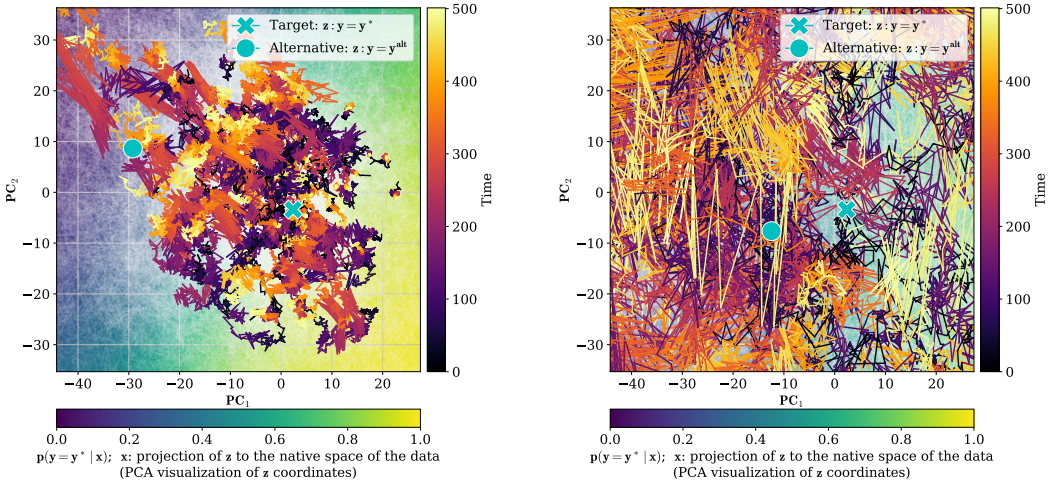
Figure 7: Evolution of p_t during the 1000 simulations for each experiment, grouped and averaged by initial point z_0 (thin lines). The color indicates the category of the Gaussian latent prototype from which each z_0 was sampled. The bold lines represent the average p_t across all trajectories for which the initial point was sampled around the same latent prototype.

Again, choosing "Dress" instead of "Trouser" as the alternative class negatively impacts feedback maximization (Figure 7c).

The control conditions (subfigures 7b and 3d) display a trend towards the chance value, acting as a sanity check.

4.4.4 How do these factors influence the evolution of internal cognitive states?

As mentioned in Section 4.3.4, DecNefSimulator allows tracking cognitive trajectories transparently instead of relying merely in the observed trajectory, bypassing the "decoder's dictum". When using high-dimensional cognitive spaces, additional dimensionality reduction methods are required for visualization,



(a) T-shirt/top vs. Trouser

(b) T-shirt/top vs. Dress

Figure 8: Cognitive trajectories $\{z_i(t)\}$ for the DecNef simulation with target-class $y^\star$ = "T-shirt/top" using either a discriminator with the alternative-class "Trouser" (a) or "Dress" (b), in the 2D PCA space. Colored background: $p(y = y^\star \mid x)$ where $x = D_\mathcal{G}(z)$ for each coordinate z and the VAE decoder denoted as $D_\mathcal{G}$. Thin white lines: all the individual trajectories. Bold lines: Average position of the 10 trajectories sharing the same z_0, with vertical color bar indicating time progression.

but the development of cognitive trajectories can still be observed.

Figure 8 shows the projection of the cognitive trajectories onto the 2D PCA space, mapping the internal cognitive states z_t to their associated performance measures p_t, derived from x_t.

The colored background (bottom color bar) is the probability map (as described in Section 4.4.1). The thin white lines depict each individual trajectory $\{z_t\}$, indicating the amount of exploration. The bold lines are averages grouped by the same initial z_0.

Figure 8a shows similar behaviour to the experiments conducted in images (Figure 4a). Analogously, exploration is less likely when the initial p_0 is high. In cognitive regions mapping to observable states with poor reward signal, the feedback successfully modifies the trajectory to induce cognitive states mapping to observations with higher $p(y = y^\star \mid x)$, but no preference is shown towards actually approaching the desired state.

Again, with the "T-shirt/top vs. Dress" classifier resulting in a more homogeneous probability map, cognitive trajectories are disorganized, with big leaps and a greater tendency to revert to previous states (Figure 8b).

5 Discussion

Decoded Neurofeedback (DecNef) is a flourishing non-invasive technique for brain modulation with multiple clinical and cognitive enhancement applications [26, 10], and is also a valuable research tool to

increase our understanding about brain function [4, 48, 11, 36]. However, DecNef is nowadays still hindered by concerning limitations [61, 9, 19, 48], namely, undesirable reinforcement of cognitive states deviating from the target, domain shift between the discriminator training stage and the target neural pattern of DecNef during induction [61, 9, 41] or the subject-dependency of learning [2, 17]. These limitations encourage the development of a comprehensive simulation framework to understand and improve DecNef in a controlled environment. In this work, we introduce a framework designed for this purpose and demonstrate its utility through a series of computational simulations.

Our simulations revealed three main findings: (i) the choice of alternative class used by the classifier during feedback computation is a critical determinant of learning outcomes; (ii) the perceived (in)ability to learn can be driven by circumstantial factors, such as cumulative randomness in transient decisions, which may lead experimenters to label a participant as a "non-responder"; and (iii) the initial cognitive state strongly shapes the sequence of states elicited during induction. In particular, initial states associated with low reward signals promote exploration and typically lead the participant to states with higher feedback, whereas initial states with high feedback suppress exploration without necessarily indicating successful induction of the target state.

Previous simulation studies [40, 50, 5] have focused on specific aspects of DecNef simulation, such as demonstrating neural plasticity [50] or modeling the influence of prior beliefs and confidence [5], but have limited flexibility and generalizability. Moreover, they have not explored interactions among cognitive states associated with semantically related concepts. In contrast, our work delves directly into those aspects, emphasizing modularity and task-agnosticity and unlocking direct access to the internal dynamics of the induction process.

DecNefSimulator explicitly considers the methodological constraints of neuroimaging-based decoding. By computing the classifier's output $p(y = y^\star \mid x)$ from observable proxies x (analogous to fMRI data), we can test the validity of the *decoder's dictum*: the assumption that successful external decoding reflects the functional use of information by the brain [15]. DecNefSimulator circumvents this issue by decoupling the experimenter's perspective from the true latent dynamics z, allowing direct assessment of whether maximizing $p(y = y^\star \mid x)$ genuinely drives regulation toward the desired cognitive state $z^\star$. This capability enables analysis of causal, rather than merely correlational, relationships between feedback and mental representations or cognitive state changes.

From a practical standpoint, our framework allows experimenters to verify whether a DecNef protocol truly induces the intended cognitive state or merely generates positive feedback via spurious mechanisms. This overcomes a fundamental limitation of empirical DecNef, where the true cognitive state during induction remains hidden. By replacing the human participant with a latent variable generative model, our simulation can provide full observability and interpretability. To demonstrate this, we ran identical simulations with image data and with synthetic fMRI data, showing how DecNefSimulator preserves direct access to internal representations regardless of the data modality. Image-based simulations support intuitive semantic interpretation, whereas synthetic fMRI — while less semantically transparent — provides greater realism for DecNef scenarios.

Our computational experiments illustrate several key capabilities. The update rule $\mathcal{L}$, encoding participant behavior, can be easily adapted to reflect different cognitive strategies. For example, we modeled a "cautious" participant reverting to previous states after negative feedback, but alternative formulations can represent more exploratory or directional learning. The parameters λ and γ respectively control confidence and attention, allowing manipulation of learning and behavioural adaptation. As shown in our binary classifier comparisons, the framework also exposes potential design flaws; for instance, how poor alternative class choices can impede learning by shaping the reward landscape.

These simulations highlight subtle yet critical aspects of DecNef learning. Trajectories starting from regions of already high reward remain constrained to the vicinity of the starting state, while trajectories

starting in low-feedback areas maximize the reward without necessarily moving toward the target state. Thus, increasing feedback and rewards across the course of DecNef training does not necessarily correspond to achieving the desired neural pattern or critical mental strategy. Modifying the alternative class (e.g., "Dress" instead of "Trouser") dramatically alters the feedback topology and learning dynamics, underscoring how classifier construction critically shapes participant behavior. Such findings suggest that some subjects labeled as "non-learners" may, in fact, be constrained by suboptimal experimental design and alternative class choice rather than inherent inability.

Hence, robust DecNef protocol design requires carefully excluding confounding factors in the feedback system or, at the very least, understanding and managing their influence. Critically, comparison-based feedback mechanisms involving variables other than the current and target states should be avoided. Moreover, DecNef protocols must strive to evaluate learning without relying merely on reward maximization. The conceptual separation of internal and observed states, as derived from the *"decoder's dictum"* and formalized in DecNefSimulator, provides the opportunity to assess the effect of new protocols directly on the cognitive states, albeit in a simulated environment.

Our work complements existing DecNef simulation approaches by focusing explicitly on cognitive-level representations/mechanisms rather than neural plasticity per se. While studies such as Shibata et al., 2019 [50] used modeling to assess *how* learning during DecNef may be instantiated from a biological perspective, our current framework focuses on *what* is learned. Subsequently, a limitation of our approach is its current inability to simulate neuroplasticity, which remains an important direction for future work.

The simplicity of the learning strategy presented here — based on an exploration–exploitation trade-off — also constrains the behavioral richness of our computational simulations. Future extensions could incorporate memory of past cognitive states, directional exploration, or active-inference principles as proposed in [5]. Owing to the modular design of the framework, such enhancements can be introduced without altering its core structure. Similarly, the generator $\mathcal{G}$ may be replaced with other latent-variable models that offer semantically organized latent spaces, such as Adversarially Learned Inference models [16], Hierarchical VAEs [18], or Variational Diffusion Models [21]. Importantly, $\mathcal{G}$ is not intended as a perfect reconstructor but as an intentionally imperfect model, analogous to how mental imagery offers a noisy reflection of perception. The correspondence we assume between the model's latent space and human cognitive states is itself an interpretive assumption. Prior work supports this view, as latent generative models have been shown to capture relevant structure in brain activity [54, 20]. However, a full evaluation of this mapping — including comparisons with alternative architectures such as LLMs or CLIP-based models [45]— lies outside the scope of the present study.

Future work will employ DecNefSimulator for systematic evaluation of new DecNef protocols before human application. Particularly promising is the replacement of supervised discriminators with self-supervised or unsupervised approaches that directly model the target class, avoiding issues introduced by comparative feedback computation.

In summary, DecNefSimulator introduces a flexible, transparent, and task-agnostic framework for DecNef simulation that bridges cognitive-level modeling and empirical neurofeedback research. By revealing the full trajectory of learning and allowing causal inspection of the relationship between the feedback and the cognitive state, it provides an unprecedented opportunity to refine DecNef methodology. We envision DecNefSimulator as both a tool for methodological innovation and a conceptual bridge toward more reproducible, interpretable, and causally grounded DecNef research.

6 Conclusions

We have introduced a flexible, task-agnostic simulation framework for DecNef that allows experimenters to explore and refine neurofeedback protocols *in silico* before applying them to human participants. Unlike existing biologically or cognitively focused approaches, our method provides direct access to the induced states and the learning dynamics that drive them, making the DecNef process fully transparent and analyzable.

By offering modular components for generators, discriminators, and update rules, the framework accommodates a wide range of experimental designs and supports systematic evaluation of protocol parameters and potential design. This enables researchers to identify both effective strategies and potential failures, reducing the trial-and-error burden of *in vivo* experimentation.

Ultimately, this work contributes a new tool for the neurofeedback community: one that bridges the gap between computational modeling and experimental practice, facilitates methodological innovation, and lays the groundwork for more robust and efficient development of DecNef protocols.

7 Acknowledgements

This work has been supported by the Basque Government via the IKUR strategy, project IT1504-22 and the BERC 2022-2025 program, by the Spanish Ministry of Science and Innovation (Projects PID2023-149195NB-I00 and PID2022-137442NB-I00), by Elkartek (KK-2023/00012, KK-2023/00090, KK-2024/00030), by the Spanish Ministry of Economy and Competitiveness, through the "Severo Ochoa" Programme for Centres/Units of Excellence (CEX2020-001010-S) and also from project grant PID2019-105494GB-I00. DS acknowledges support by the Basque Government through the BERC 2022-2025 program and Funded by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation CEX2020-001010/AEI/10.13039/501100011033 and through project PID2023-149267NB-I00.

References

[1] N. Abdennour, P. Margolles, and D. Soto. Enhancing fmri decoded neurofeedback with co-adaptive training: simulation and proof-of-principle evidence. *bioRxiv*, 2025. doi: 10.1101/2025.02.21.639408. URL `https://www.biorxiv.org/content/early/2025/09/29/2025.02.21.639408`.

[2] O. Alkoby, A. Abu-Rmileh, O. Shriki, and D. Todder. Can we predict who will respond to neurofeedback? a review of the inefficacy problem and existing predictors for successful eeg neurofeedback learning. *Neuroscience*, 378:155–164, 2018. ISSN 0306-4522. doi: 10.1016/j.neuroscience.2016.12.050. URL `https://doi.org/10.1016/j.neuroscience.2016.12.050`.

[3] E. J. Allen, G. St-Yves, Y. Wu, J. L. Breedlove, J. S. Prince, L. T. Dowdle, M. Nau, B. Caron, F. Pestilli, I. Charest, J. B. Hutchinson, T. Naselaris, and K. Kay. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, Jan 2022. ISSN 1546-1726. doi: 10.1038/s41593-021-00962-x. URL `https://doi.org/10.1038/s41593-021-00962-x`.

[4] K. Amano, K. Shibata, M. Kawato, Y. Sasaki, and T. Watanabe. Learning to associate orientation with color in early visual areas by associative decoded fmri neurofeedback. *Current Biology*, 26(14):1861–1866, 7 2016. ISSN 0960-9822. doi: 10.1016/j.cub.2016.05.014. URL `https://doi.org/10.1016/j.cub.2016.05.014`.

[5] C. Annicchiarico, F. Lotte, and J. Mattout. An active inference perspective on neurofeedback training. *arXiv*, 2025. doi: https://doi.org/10.48550/arXiv.2505.03308. URL https://arxiv.org/abs/2505.03308.

[6] F. Arab, A. Ghassami, H. Jamalabadi, M. A. K. Peters, and E. Nozari. Whole-brain causal connectivity during decoded neurofeedback: a meta study. *bioRxiv*, 2024. doi: 10.1101/2024.11.16.623939. URL https://www.biorxiv.org/content/early/2024/11/17/2024.11.16.623939.

[7] G. Bao, Q. Zhang, Z. Gong, Z. Wu, and D. Miao. Mindsimulator: Exploring brain concept localization via synthetic fMRI. In *The Thirteenth International Conference on Learning Representations*, 2025. doi: https://doi.org/10.48550/arXiv.2503.02351. URL https://openreview.net/forum?id=vgt2rSf6al.

[8] J. L. Breedlove, G. St-Yves, C. A. Olman, and T. Naselaris. Generative feedback explains distinct brain activity codes for seen and mental images. *Current Biology*, 30(12):2211–2224.e6, 2020. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2020.04.014. URL https://www.sciencedirect.com/science/article/pii/S0960982220304942.

[9] A. Cortese, K. Amano, A. Koizumi, M. Kawato, and H. Lau. Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nature Communications*, 7(1): 13669, 12 2016. ISSN 2041-1723. doi: 10.1038/ncomms13669. URL https://doi.org/10.1038/ncomms13669.

[10] A. Cortese, K. Amano, A. Koizumi, H. Lau, and M. Kawato. Decoded fmri neurofeedback can induce bidirectional confidence changes within single participants. *NeuroImage*, 149:323–337, 2017. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2017.01.069. URL https://www.sciencedirect.com/science/article/pii/S1053811917300915.

[11] A. Cortese, H. Lau, and M. Kawato. Unconscious reinforcement learning of hidden brain states supported by confidence. *Nature Communications*, 11(1):4429, 8 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17828-8. URL https://doi.org/10.1038/s41467-020-17828-8.

[12] S. Cui, D. Lee, and D. Wen. Toward brain-inspired foundation model for eeg signal processing: our opinion. *Frontiers in Neuroscience*, Volume 18 - 2024, 2024. ISSN 1662-453X. doi: 10.3389/fnins.2024.1507654. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2024.1507654.

[13] E. J. Davelaar. Chapter 3 - the multi-stage theory of neurofeedback learning: a framework for understanding mechanisms. In S. H. Fairclough and T. O. Zander, editors, *Current Research in Neuroadaptive Technology*, pages 43–57. Academic Press, 2022. ISBN 978-0-12-821413-8. doi: https://doi.org/10.1016/B978-0-12-821413-8.00008-7. URL https://www.sciencedirect.com/science/article/pii/B9780128214138000087.

[14] L. O. B. de Almeida, J. F. W. Slaets, and R. Köberle. Vsimg: A high frame rate bitmap based display system for neuroscience research. *Neurocomputing*, 74(10):1762–1768, 2011. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2011.02.016. URL https://www.sciencedirect.com/science/article/pii/S0925231211001433.

[15] L. de-Wit, D. Alexander, V. Ekroll, and J. Wagemans. Is neuroimaging measuring information in the brain? *Psychonomic Bulletin & Review*, 23(5):1415–1428, 2016. doi: 10.3758/s13423-016-1002-0.

[16] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville. Adversarially learned inference. In *Proceedings of the International Conference on Learning Representations*, 2017. doi: 10.48550/arXiv.1606.00704. URL https://arxiv.org/abs/1606.00704.

[17] A. Haugg, F. M. Renz, A. A. Nicholson, C. Lor, S. J. Gotzendorfer, R. Sladky, S. Skouras, A. McDonald, C. Craddock, L. Hellrung, M. Kirschner, M. Herdener, Y. Koush, M. Papoutsi, J. Keynan, T. Hendler, K. Cohen Kadosh, C. Zich, S. H. Kohl, M. Hallschmid, J. MacInnes, R. A. Adcock, K. C. Dickerson, N.-K. Chen, K. Young, J. Bodurka, M. Marxen, S. Yao, B. Becker, T. Auer, R. Schweizer, G. Pamplona, R. A. Lanius, K. Emmert, S. Haller, D. Van De Ville, D.-Y. Kim, J.-H. Lee, T. Marins, F. Megumi, B. Sorger, T. Kamp, S.-L. Liew, R. Veit, M. Spetter, N. Weiskopf, F. Scharnowski, and D. Steyrl. Predictors of real time fmri neurofeedback performance and improvement: A machine learning mega-analysis. *NeuroImage*, 237: 118207, 2021. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2021.118207. URL `https://www.sciencedirect.com/science/article/pii/S1053811921004845`.

[18] J. D. Havtorn, J. Frellsen, S. Hauberg, and L. Maaløe. Hierarchical vaes know what they don't know. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4117–4128. PMLR, 07 2021. doi: https://doi.org/10.48550/arXiv.2102.08248. URL `https://proceedings.mlr.press/v139/havtorn21a.html`.

[19] S. Heunis, R. Lamerichs, S. Zinger, C. Caballero-Gaudes, J. F. A. Jansen, B. Aldenkamp, and M. Breeuwer. Quality and denoising in real-time functional magnetic resonance imaging neurofeedback: A methods review. *Hum. Brain Mapp.*, 41(12):3439–3467, Aug. 2020. doi: 10.1002/hbm.25010.

[20] Y. Kamitani, M. Tanaka, and K. Shirakawa. Visual image reconstruction from brain activity via latent representation. *Annual Review of Vision Science*, 11(Volume 11, 2025):611–634, 2025. ISSN 2374-4650. doi: https://doi.org/10.1146/annurev-vision-110423-023616. URL `https://www.annualreviews.org/content/journals/10.1146/annurev-vision-110423-023616`.

[21] D. Kingma, T. Salimans, B. Poole, and J. Ho. Variational diffusion models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 21696–21707. Curran Associates, Inc., 2021. doi: 10.48550/arXiv.2107.00630. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/b578f2a52a0229873fefc2a4b06377fa-Paper.pdf`.

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*, 2015. doi: 10.48550/arXiv.1412.6980.

[23] D. P. Kingma and M. Welling. An introduction to variational autoencoders. *Found. Trends Mach. Learn.*, 12(4):307–392, 11 2019. ISSN 1935-8237. doi: 10.48550/arXiv.1906.02691. URL `https://doi.org/10.48550/arXiv.1906.02691`.

[24] J. Klawitter and K. P. Michmizos. Sequential learning in neural decoders reveals interpretable neural representations. *Neurocomputing*, 672:132742, 2026. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2026.132742. URL `https://www.sciencedirect.com/science/article/pii/S0925231226001396`.

[25] T. Kohonen. Essentials of the self-organizing map. *Neural Networks*, 37:52–65, 2013. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2012.09.018. URL `https://www.sciencedirect.com/science/article/pii/S0893608012002596`. Twenty-fifth Anniversay Commemorative Issue.

[26] A. Koizumi, K. Amano, A. Cortese, K. Shibata, W. Yoshida, B. Seymour, M. Kawato, and H. Lau. Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure. *Nature Human Behaviour*, 1(1):0006, 11 2016. ISSN 2397-3374. doi: 10.1038/s41562-016-0006. URL `https://doi.org/10.1038/s41562-016-0006`.

[27] Y. Koush, M. Zvyagintsev, M. Dyck, K. A. Mathiak, and K. Mathiak. Signal quality and bayesian signal processing in neurofeedback based on real-time fmri. *NeuroImage*, 59(1):478–489, 2012. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2011.07.076. URL

https://www.sciencedirect.com/science/article/pii/S1053811911008718. Neuroergonomics: The human brain in action and at work.

[28] P. A. Kragel, L. Koban, L. F. Barrett, and T. D. Wager. Representation, pattern information, and brain signatures: From neurons to neuroimaging. *Neuron*, 99(2):257–273, 2018. ISSN 0896-6273. doi: 10.1016/j.neuron.2018.06.009.

[29] S. M. LaConte. Decoding fmri brain states in real-time. *NeuroImage*, 56(2):440–454, 2011. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2010.06.052. URL https://www.sciencedirect.com/science/article/pii/S1053811910009109. Multivariate Decoding and Brain Reading.

[30] E. J. Lawrence, L. Su, G. J. Barker, N. Medford, J. Dalton, S. C. Williams, N. Birbaumer, R. Veit, S. Ranganatha, J. Bodurka, M. Brammer, V. Giampietro, and A. S. David. Self-regulation of the anterior insula: Reinforcement learning using real-time fmri neurofeedback. *NeuroImage*, 88: 113–124, 2014. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2013.10.069. URL https://www.sciencedirect.com/science/article/pii/S1053811913010951.

[31] Y. Le Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *Proceedings of the 3rd International Conference on Neural Information Processing Systems*, NIPS'89, page 396–404, Cambridge, MA, USA, 1989. MIT Press.

[32] N. K. Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878, 2008. ISSN 1476-4687. doi: 10.1038/nature06976.

[33] N. Lubianiker, C. Paret, P. Dayan, and T. Hendler. Neurofeedback through the lens of reinforcement learning. *Trends in Neurosciences*, 45(8):579–593, 2022. ISSN 0166-2236. doi: 10.1016/j.tins.2022.03.008. URL https://doi.org/10.1016/j.tins.2022.03.008.

[34] W. Mai, J. Wu, Y. Zhu, Z. Yao, D. Zhou, A. F. Luo, Q. Zheng, W. Ouyang, and C. Song. Synbrain: Enhancing visual-to-fmri synthesis via probabilistic representation learning. *arXiv preprint arXiv:2508.10298*, 2025. doi: https://doi.org/10.48550/arXiv.2508.10298.

[35] D. Maisto, F. Gregoretti, K. J. Friston, and G. Pezzulo. Active inference tree search in large pomdps. *Neurocomputing*, 623:129319, 2025. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2024.129319. URL https://www.sciencedirect.com/science/article/pii/S0925231224020903.

[36] P. Margolles, P. Elosegi, N. Mei, and D. Soto. Unconscious manipulation of conceptual representations with decoded neurofeedback impacts search behavior. *Journal of Neuroscience*, 44(2), 2024. doi: 10.1523/JNEUROSCI.1235-23.2023.

[37] T. Micoulaud-Franchi, J.-A. & Fovet. A framework for disentangling the hyperbolic truth of neurofeedback: Comment on thibault and raz (2017). *American Psychologist*, 73(7):933–935, 2018. doi: https://doi.org/10.1037/amp0000340.

[38] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1765–1773, 2017. doi: 10.1109/CVPR.2017.17.

[39] Y. Niv, R. Daniel, A. Geana, S. J. Gershman, Y. C. Leong, A. Radulescu, and R. C. Wilson. Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, 35(21):8145 – 8157, 2015. doi: 10.1523/JNEUROSCI.2978-14.2015. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-84930260511&doi=10.1523%2fJNEUROSCI.2978-14.2015&partnerID=40&md5=785babfb9700b75fe736c86dda0ca9eb. Cited by: 275.

[40] E. F. Oblak, J. A. Lewis-Peacock, and J. S. Sulzer. Self-regulation strategy, feedback timing and hemodynamic properties modulate learning in a simulated fmri neurofeedback environment. *PLOS Computational Biology*, 13(7):1–25, 07 2017. doi: 10.1371/journal.pcbi.1005681. URL https://doi.org/10.1371/journal.pcbi.1005681.

[41] A. Olza, D. Soto, and R. Santana. Domain adaptation-enhanced searchlight: enabling classification of brain states from visual perception to mental imagery. *Brain Informatics*, 12(1):17, 2025. doi: 10.1186/s40708-025-00263-0. URL https://doi.org/10.1186/s40708-025-00263-0.

[42] F. Ozcelik and R. VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-42891-8. URL https://doi.org/10.1038/s41598-023-42891-8.

[43] Y. S. Perl, H. Bocaccio, I. Pérez-Ipiña, F. Zamberlán, J. Piccinini, H. Laufs, M. Kringelbach, G. Deco, and E. Tagliazucchi. Generative embeddings of brain collective dynamics using variational autoencoders. *Phys. Rev. Lett.*, 125:238101, 2020. doi: 10.1103/PhysRevLett.125.238101. URL https://link.aps.org/doi/10.1103/PhysRevLett.125.238101.

[44] Y. S. Perl, C. Pallavicini, J. Piccinini, A. Demertzi, V. Bonhomme, C. Martial, R. Panda, N. Alnagger, J. Annen, O. Gosseries, A. Ibañez, H. Laufs, J. D. Sitt, V. K. Jirsa, M. L. Kringelbach, S. Laureys, G. Deco, and E. Tagliazucchi. Low-dimensional organization of global brain states of reduced consciousness. *Cell Reports*, 42(5):112491, 2023. ISSN 2211-1247. doi: https://doi.org/10.1016/j.celrep.2023.112491. URL https://www.sciencedirect.com/science/article/pii/S2211124723005028.

[45] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. doi: https://doi.org/10.48550/arXiv.2103.00020. URL https://proceedings.mlr.press/v139/radford21a.html.

[46] J. B. Ritchie, D. M. Kaplan, and C. Klein. Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *British Journal for the Philosophy of Science*, 70(2):581–607, 2019. doi: 10.1093/bjps/axx023.

[47] T. Ros, S. Enriquez-Geppert, V. Zotev, K. D. Young, G. Wood, S. Whitfield-Gabrieli, F. Wan, P. Vuilleumier, F. Vialatte, D. Van De Ville, D. Todder, T. Surmeli, J. S. Sulzer, U. Strehl, M. B. Sterman, N. J. Steiner, B. Sorger, S. R. Soekadar, R. Sitaram, L. H. Sherlin, M. Schönenberg, F. Scharnowski, M. Schabus, K. Rubia, A. Rosa, M. Reiner, J. A. Pineda, C. Paret, A. Ossadtchi, A. A. Nicholson, W. Nan, J. Minguez, J.-A. Micoulaud-Franchi, D. M. A. Mehler, M. Lührs, J. Lubar, F. Lotte, D. E. J. Linden, J. A. Lewis-Peacock, M. A. Lebedev, R. A. Lanius, A. Kübler, C. Kranczioch, Y. Koush, L. Konicar, S. H. Kohl, S. E. Kober, M. A. Klados, C. Jeunet, T. W. P. Janssen, R. J. Huster, K. Hoedlmoser, L. M. Hirshberg, S. Heunis, T. Hendler, M. Hampson, A. G. Guggisberg, R. Guggenberger, J. H. Gruzelier, R. W. Göbel, N. Gninenko, A. Gharabaghi, P. Frewen, T. Fovet, T. Fernández, C. Escolano, A.-C. Ehlis, R. Drechsler, R. Christopher deCharms, S. Debener, D. De Ridder, E. J. Davelaar, M. Congedo, M. Cavazza, M. H. M. Breteler, D. Brandeis, J. Bodurka, N. Birbaumer, O. M. Bazanova, B. Barth, P. D. Bamidis, T. Auer, M. Arns, and R. T. Thibault. Consensus on the reporting and experimental design of clinical and cognitive-behavioural neurofeedback studies (cred-nf checklist). *Brain*, 143(6):1674–1685, 03 2020. ISSN 0006-8950. doi: 10.1093/brain/awaa009. URL https://doi.org/10.1093/brain/awaa009.

[48] K. Shibata. Chapter 13 - mechanisms of fmri neurofeedback. In M. Hampson, editor, *fMRI Neurofeedback*, pages 287–313. Academic Press, 2021. ISBN 978-0-12-822421-2. doi: https://doi.org/10.1016/B978-0-12-822421-2.00004-1. URL https://www.sciencedirect.com/science/article/pii/B9780128224212000041.

[49] K. Shibata, T. Watanabe, Y. Sasaki, and M. Kawato. Perceptual learning incepted by decoded fmri neurofeedback without stimulus presentation. *Science*, 334(6061):1413–1415, 2011. doi: 10.1126/science.1212003.

[50] K. Shibata, G. Lisi, A. Cortese, T. Watanabe, Y. Sasaki, and M. Kawato. Toward a comprehensive understanding of the neural mechanisms of decoded neurofeedback. *NeuroImage*, 188:539–556, 2019. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2018.12.022. URL `https://www.sciencedirect.com/science/article/pii/S1053811918321669`.

[51] R. Sitaram, T. Ros, L. Stoeckel, S. Haller, N. Scharnowski, J. Lewis, R. Veit, B. Weiskopf, M. Blefari, M. Rana, J. Oblak, J. C. Birbaumer, and N. Birbaumer. Closed-loop brain training: the science of neurofeedback. *Nature Reviews Neuroscience*, 18(2):86–100, 2017. doi: 10.1038/nrn.2016.164.

[52] M. T. J. Spaan. *Partially Observable Markov Decision Processes*, pages 387–414. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-27645-3. doi: 10.1007/978-3-642-27645-3_12. URL `https://doi.org/10.1007/978-3-642-27645-3_12`.

[53] C. Stevens and S. E. Kober. Neural mechanisms of neurofeedback: A review. *Frontiers in Human Neuroscience*, 15:619546, 2021. doi: 10.3389/fnhum.2021.619546.

[54] Y. Takagi and S. Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14453–14463, 2023. doi: 10.1109/CVPR52729.2023.01389.

[55] V. Taschereau-Dumouchel, A. Cortese, H. Lau, and M. Kawato. Conducting decoded neurofeedback studies. *Social Cognitive and Affective Neuroscience*, 16(8):838–848, 05 2020. ISSN 1749-5016. doi: 10.1093/scan/nsaa063. URL `https://doi.org/10.1093/scan/nsaa063`.

[56] R. T. Thibault and A. Raz. The psychology of neurofeedback: Clinical intervention even if applied placebo. *American Psychologist*, 72(7):679—-688, 2017. doi: https://doi.org/10.1037/amp0000118.

[57] A. Tursic, J. Eck, M. Lührs, D. E. Linden, and R. Goebel. A systematic review of fmri neurofeedback reporting and effects in clinical populations. *NeuroImage: Clinical*, 28:102496, 2020. ISSN 2213-1582. doi: https://doi.org/10.1016/j.nicl.2020.102496. URL `https://www.sciencedirect.com/science/article/pii/S2213158220303338`.

[58] J. Vadillo, R. Santana, and J. A. Lozano. Analysis of dominant classes in universal adversarial perturbations. *Knowledge-Based Systems*, 236:107719, 2022. doi: https://doi.org/10.1016/j.knosys.2021.107719.

[59] L. A. Weber, T. Ethofer, and A.-C. Ehlis. Predictors of neurofeedback training outcome: A systematic review. *NeuroImage: Clinical*, 27:102301, 2020. ISSN 2213-1582. doi: https://doi.org/10.1016/j.nicl.2020.102301. URL `https://www.sciencedirect.com/science/article/pii/S2213158220301388`.

[60] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[61] O. Yamashita, M.-a. Sato, T. Yoshioka, F. Tong, and Y. Kamitani. Sparse estimation automatically selects voxels relevant for the decoding of fmri activity patterns. *NeuroImage*, 42(4):1414–1429, 2008. doi: 10.1016/j.neuroimage.2008.05.050.