u^b

_b_
UNIVERSITÄT
BERN

Graduate School for Cellular and Biomedical Sciences

University of Bern

Learning to Dream, Dreaming to Learn

PhD Thesis submitted by

Nicolas Rouben Pascal Deperrois

for the degree of

PhD in Neuroscience

Supervisor
Prof. Dr. Walter Senn
Department of Physiology
University of Bern, Switzerland

Co-advisor
Prof. Dr. Paolo Favaro
Institute of Computer Science
University of Bern, Switzerland

Accepted by the Faculty of Medicine, the Faculty of Science and the Vetsuisse Faculty of the University of Bern at the request of the Graduate School for Cellular and Biomedical Sciences

Bern, Dean of the Faculty of Medicine

Bern, Dean of the Faculty of Science

Bern, Dean of the Vetsuisse Faculty Bern

Acknowledgements

After graduating from my Master of neuroscience in 2018, I did not have any concrete plan for a PhD. Intrigued by the world of artificial intelligence, I decided to stay for another year in ENS Paris to take classes in machine learning and prepare a transition to industry. In October 2018, all these plans got ruined after receiving an email from Walter Senn, inviting me to visit his lab in Bern. Four months later, I would leave Paris to open a new chapter of my life, "a PhD story from Bern". Thank you Walter, to have offered me the opportunity to work in this amazing lab, that merges both of my favorite fields, neuroscience and machine learning. Without this email, I would probably be an eternal unfit 30-year old Parisian smoking rolled cigarettes on a café terrace.

Surviving as a PhD student was not a small feat. I am immensely grateful to Jakob Jordan whose weekly guidance, support and writing helped to turn this dreamed project into reality. And I am sure that the discussions we had about the model while swimming in the Aare played a lot too.

I am greatly thankful to Mihai A. Petrovici, a redoutable scientist, athlete and board game player, who made sure of the successful completion of different steps of my PhD adventure.

Furthermore, I thank all past and present group members of the Senn and Petrovici groups, both for their scientific input and for great moments spent outside the lab. In particular, I thank Elena Kreutzer and Camille Gontier, for their instructions about the thesis writing, and Laura Kriener and Jakob Jordan, for their feedback on the thesis manuscript.

Last but not least, I would like to thank my mother Nathalie, my father Hervé, my sister Elvire and my cousin Gilles, for their lifelong support, as well as my friends (in particular from the Iron Paradise) and my girlfriend Nina (who is also part of the Iron Paradise since recently).

Abstract

The importance of sleep for healthy brain function is widely acknowledged. However, it remains mysterious how the sleeping brain, disconnected from the outside world and plunged into the fantastic experiences of dreams, is actively learning. A main feature of dreams is the generation of new realistic sensory experiences in absence of external input, from the combination of diverse memory elements. How do cortical networks host the generation of these sensory experiences during sleep? What function could these generated experiences serve?

In this thesis, we attempt to answer these questions using an original, computational approach inspired by modern artificial intelligence. In light of existing cognitive theories and experimental data, we suggest that cortical networks implement a generative model of the sensorium that is systematically optimized during wakefulness and sleep states. By performing network simulations on datasets of natural images, our results not only propose potential mechanisms for dream generation during sleep states, but suggest that dreaming is an essential feature for learning semantic representations throughout mammalian development.

Contents

List of Figures

List of Abbreviations

ACC	Anterior Cingulate Cortex
ACh	Acetylcholine
AE	Autoencoder
AI	Artificial Intelligence
ANNs	Artificial Neural Networks
CIFAR	Canadian Institute for Advanced Research
CLS	Complementary Learning Systems
CNNs	Convolutional Neural Networks
DCGANs	Deep Convolutional Generative Adversarial Networks
DRM	Deese-Roediger-McDermott
EEG	Electro-Encephalogram
ELBO	Evidence Lower Bound
FID	Fréchet Inception Distance
fMRI	functional Magnetic Resonance Imaging
GANs	Generative Adversarial Networks
HBP	Human Brain Project
IT	Inferior-Temporal
JS	Jensen-Shannon
KL	Kullback-Leibler
MLP	Multi-Layer Perceptron
mPFC	medial Prefrontal Cortex
NA	noradrenaline
NREM	Non-rapid-eye-movement
PAD	Perturbed and Adversarial Dreaming
PC	Predictive coding
PCA	Principal Component Analysis
PGO	Ponto-Geniculo Occipital
RBM	Restricted Boltzmann Machine
REM	Rapid-eye-movement
SCT	Standard consolidation theory
SVHN	Street View House Numbers
SWR	Sharp-Wave Ripples
SWS	Slow-Wave Sleep
TTT	Trace Transformation Theory
VAE	Variational Autoencoder
WS	Wake-Sleep

To my mum, who will win the fight...

Chapter 1

Introduction

What I cannot create, I do not understand

— Richard Feynman

1.1 General introduction

Right from birth, humans and animals access the amazing and unique experience of life. Plunging into this gigantic adventure requires to properly perceive and interact with the environment. To this aim, animals are equipped with sensory abilities, such as vision, hearing or touch, allowing them to correctly interpret the upcoming sensory signals from the outside world and perform the appropriate actions.

Quite surprisingly, these sensory abilities seem to develop by themselves. Throughout the days, animals naturally learn to make sense of what they perceive, by discovering, organizing and connecting concepts, without being systematically guided to do so. The brain, with its centralized control over our body, processing sensory signals and ordering motor commands, lies at the heart of these abilities. In fact, already early in life, neuronal activities in sensory cortical areas tend to extract high-level semantic concepts such as objects, faces or voices from raw sensory signals (Ito et al., 1995; Hung et al., 2005; Formisano et al., 2008). However, how these well organized cortical representations are learned throughout development remains mysterious.

Another mystery is that we all spend an important fraction of our lifetime asleep, immobile and disconnected from this world. Notably, this black-out state often hosts fantastic experiences, or dreams, merging diverse elements from our waking life into a whole new story, full of colors, sounds and emotions, and that strikingly appear realistic (Nir and Tononi, 2010). Possibly, from the hard labour of interacting with the environment while awake, the brain finally deserves a moment of peace where its energy and learning abilities are recovered, leaving dreams as a by-product. But what if sleep, similarly to wakefulness, plays an active role in learning? What if, during the so-far unexplained occurrence of dreams, the brain is actually rearranging its sensory representations to construct a better understanding of its environment?

Answering these questions is not a trivial task. The brain contains billions of neurons, each connected by synapses to several thousands of other neurons. Moreover, the connectivity and the organization of these neurons differ across brain areas, and evolve every single second. To understand how the brain constructs neuronal representations of the environment, one should record the stimulus-evoked neuronal activity from

many neurons of sensory areas, the plasticity changes occurring during wakefulness and sleep in millions of synapses, and the consequent behavior, throughout many days of development. Unfortunately, in spite of the exciting development of experimental techniques allowing to record many neurons at a time, such as multi-array electrodes or optogenetical stimulation, this type of experimental set-up is still hardly feasible. Furthermore, even if one had access to this data, it would not be sufficient to infer what learning principles the brain is provided with.

Complementary to experimental neuroscience, for over a century, computational neuroscience has attempted to connect biological form to function by describing neuronal activity through mathematical models, abstracting away from certain biological details. These models have been successful at characterizing dynamical properties of neuronal activity and their consequence on behavior. For the past decade, taking inspiration from the recent revolution of deep artificial neural networks (ANNs, LeCun et al., 2015), computational models have attempted to explain how cortical neurons perform complex tasks such as inferring the semantic properties of real-world sensory inputs. Here, we aim to take part in this scientific endeavour by hypothesizing whether deep generative modeling (Rao and Ballard, 1999; Kingma and Welling, 2013; Goodfellow et al., 2014; Bond-Taylor et al., 2021) could provide hypotheses about the mechanisms underlying the generation of dreams during sleep and their role in learning cortical representations.

In the next sections, we review the neurobiological and computational principles underlying this work, covering the functional organization of the sensory cortex, the view of the brain as a generative model and the current theories about the role of sleep and dreams.

1.2   Learning in the brain: from sensory inputs to semantic knowledge

1.2.1   What do we perceive from our sensorium?

Throughout their life, animals are daily exposed to various sensory (visual, auditory, tactile) stimuli that constitute the basis of their actions in the environment: walking, finding food, interacting socially, escaping danger. For instance, in vision, the reception of photons on the retina and the interpretation of their patterns into shapes and objects allow an animal to understand what its surrounding environment is made of. Indeed, animals effortlessly recognize objects and complex shapes in a fraction of a second over multiple viewpoints (Thorpe et al., 1996), and can generalize them over different instances and contexts (DiCarlo et al., 2012).

A proper perception of visual inputs is mainly relevant to perform appropriate actions in the environment. This is observed throughout development, where complexity and accuracy of perception evolve in parallel with the complexity of performed actions, e.g., from being fed to driving a bicycle. Indeed, newborns are already familiar with visual stimuli with face-like structure (Johnson et al., 1991; Farroni Teresa et al., 2005) and by three-four months can recognize three-dimensional shapes (Nishimura et al., 2009). These visual abilities keep improving such that by six years of age, their grating acuity and contrast sensitivity are adult-like (Ellemberg et al., 1999). The ability to name more complex objects (bicycles, cars, abstract 3-D shapes) then

improves from young chilhood to adolescence (Bova et al., 2007; Nishimura et al., 2009).

It thus seems that what we perceive from the sensorium are statistical regularities, such as shapes, objects or faces. Discerning these semantic features may underly the ability to perform appropriate actions, e.g., walking, interacting with other animals, or avoiding danger, as it allows us to delimit our surrounding environment into discrete instances from which we decide if a given action is possible (Hafner et al., 2020). An enigmatic question remains - how does the brain transform retinal signal into this semantically interpretable information?

1.2.2   Hierarchical processing of sensory inputs into semantic representations

The efficient perception observed in animals is likely to be supported by neurons from the visual areas of the brain that detect regularities within visual inputs.

1.2.2.1   The neuron doctrine

Historically, the neuron doctrine initially assumed that individual neurons constitute the structural and functional unit of the nervous system (Barlow, 1995). In this line, for the visual system, it was believed that a specific object is represented by a single neuron, also known as the "grand-mother cell", that gets activated when this object is present (Gross, 2002). This assumption is based on the observation that individual neurons in temporal cortex of monkeys humans fired selectively to the perception of particular instances, such as faces or other complex shapes (Booth and Rolls, 1998; Kreiman et al., 2000; Quiroga et al., 2005).

1.2.2.2   Neuronal representations

However, the neuron doctrine could be partially attributed to the extensive use of single-neuron recording methods at the time. In fact, it is difficult to believe that one neuron encodes for a particular face or object. First, such encoding would be lost if this specific neuron were to die. Second, statistically, it is unlikely that a recorded neuron, out of millions, responds to a specific instance that an animal observes. It is more likely that this encoding is distributed across a population of many neurons, which could simultaneously encode for multiple objects, depending on its activity pattern (distributed representation, Ishai et al., 1999). Indeed, with the development of new recording methods in neuroscience, notably allowing to record multineuronal activities, experimental work has revealed that populations of neurons, rather than individual cells, carry object information (Ishai et al., 1999; Yuste, 2015). A particular visual scene would give rise to a specific population activity pattern among the ensemble of neurons, that could then be easily read out by downstream brain areas in order to perform appropriate actions. This form of coding is found in inferior-temporal (IT) cortex of monkeys (Grill-Spector et al., 2001; Hung et al., 2005) or humans (Ishai et al., 1999; Haxby James V. et al., 2001; Majaj et al., 2015) where a simple weighted-sum of firing rates from many IT neurons is enough to separate representations according to object category.

1.2.2.3   Cortical processing of visual information

From the observation of stimulus-evoked activity and receptive fields from different brain areas of the cortex, it is widely believed that the transformation of a retinal

input into a high-level, semantic representation is performed through a succession of neuronal computations along the cortical hierarchy. Indeed, in the lowest visual area (V1), neurons encode low-level properties of the observed image, acting as Gabor-like edge detectors (Carandini, 2005). It then continues through a series of brain areas (V2, V4), where neurons become tuned to object features of intermediate complexity (Rust and DiCarlo, 2010; Connor et al., 2007) and eventually reaches the IT cortex where neuronal representations have large receptive fields and are sensitive to global shapes (objects, faces) involved in invariant object recognition (Hung et al., 2005; DiCarlo et al., 2012).

Conceptually, one can consider that initially, the retinal image is tangled and hardly delineate object information. Indeed, at the pixel level, the representations of all exemplar of a same category under different identity-preserving transformations form a low-dimensional manifold in a high-dimensional space that is highly curved and tangled (Fig. 1.1a), difficult to separate from other category-manifolds (DiCarlo et al., 2012). The brain might then apply a succession of processing steps that gradually transforms this representation into a disentangled one where object manifolds are more easily separable by a linear plane, i.e., where object concepts are easily readable by downstream areas (Fig. 1.1b). This gradual transformation could be performed by layers of interconnected neurons forming a network whose output displays disentangled representations of the perceived sensory input.

1.2.2.4   The role of feedback pathways in cortical processing

In parallel to this bottom-up feedforward process (from low to high visual cortical areas), there exists a large amount of descending top-down connections constituting feedback pathways that reversely send higher-order information to lower cortical areas (Fig. 1.1c, Gilbert and Li, 2013). These are thought to be involved in spatial attention, i.e., to select behaviorally relevant stimuli or focus on specific parts of the visual field (Moore and Zirnsak, 2017), or object expectation, to make neurons become selective for shapes of expected objects by creating a set of specific low-level filters toward this object (McManus et al., 2011). Certain theories of neuronal computation (Rao and Ballard, 1999) propose that feedback pathways carry predictions of the upcoming sensory input (see Section 1.3.3.1). As mentioned later (Section 1.4.4.1), feedback is also believed to be involved in the generation of sensory experience during mental imagery or dreaming, initiated in high cortical areas and descending in reverse direction of feedforward flow to create low-level representations in V1 (Nir and Tononi, 2010; Pearson, 2019). Finally, as we discuss next, feedback connections are also proposed to send backpropagated errors to update feedforward connections (Whittington and Bogacz, 2019; Lillicrap et al., 2020).

1.2.3   Computational theory of sensory processing: the deep learning framework

The ability of animals to perform efficient object recognition might be a consequence of a disentanglement of sensory representations along the ventral stream. How can neurons perform such operations? As stated above, neurons are organized into a network of successive areas that progressively extract high-level features along the hierarchy. There must be a computational principle that explains how interconnected neurons among this network altogether compute these complex features, abandonning the initial neuron doctrine to make way for neural network models paradigms (Yuste, 2015).
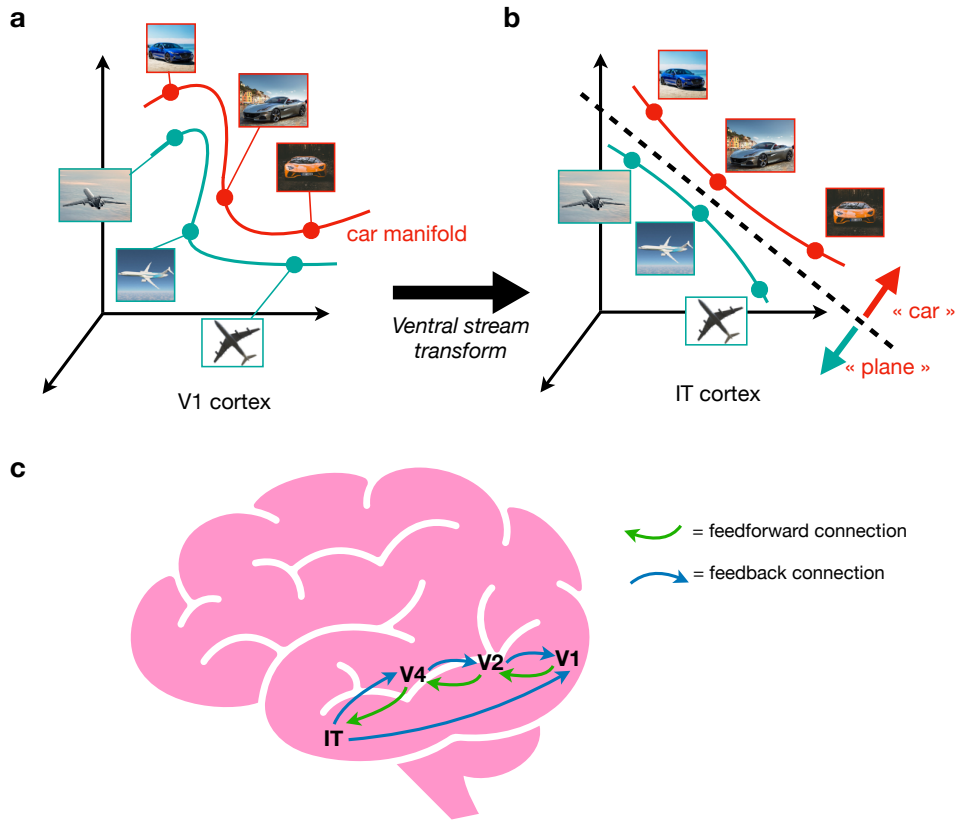
FIGURE 1.1: **Feedforward and feedback processing in the visual cortex. (a, b)**
We represent the activity pattern of population of visual neurons to each image as a point
in a high-dimensional space where each axis is the activity level of each neuron. All possi-
ble instances of an object form a low-dimensional manifold in the population vector space
(turquoise for plane, red for car). **(a)** In early visual areas (V1), object identity manifolds
are highly curved and tangled together. **(b)** The series of successive steps along the cortical
hierarchy allows to represent object manifolds in higher areas (IT) such that it can be sepa-
rated by a simple weighted summation rule (i.e., a hyperplane, black dashed line). Adapted
from DiCarlo et al. (2012). **(c)** In the ventral stream of the visual cortex, feedforward con-
nections (green) project from lower (V1) to higher (IT) areas. Matching these feedforward
connections are a series of reciprocal feedback connections (blue arrows). Diverse informa-
tion is conveyed across these feedback pathways, including attention, expectation and mental
imagery. Adapted from Gilbert and Li (2013).

1.2.3.1 The multi-layer perceptron

In the past decade, deep learning models using ANNs have proved success in various
tasks like object recognition (Krizhevsky et al., 2012), generative modeling (Good-
fellow et al., 2014; Karras et al., 2018) or reinforcement learning (Mnih et al., 2013).
These models are made of small elements called units, that can be seen as a high-level
abstraction of biological neurons: they integrate multiple inputs from other units by
performing a weighted sum, apply a non-linear transformation, and produce a single
scalar output that will be sent to other units (Fig. 1.2a). This is modelled via the
following equation:

$$x^{(1)} = f\left(\sum_{j=1}^{m} w_j^{(0)} x_j^{(0)}\right) \qquad (1.1)$$
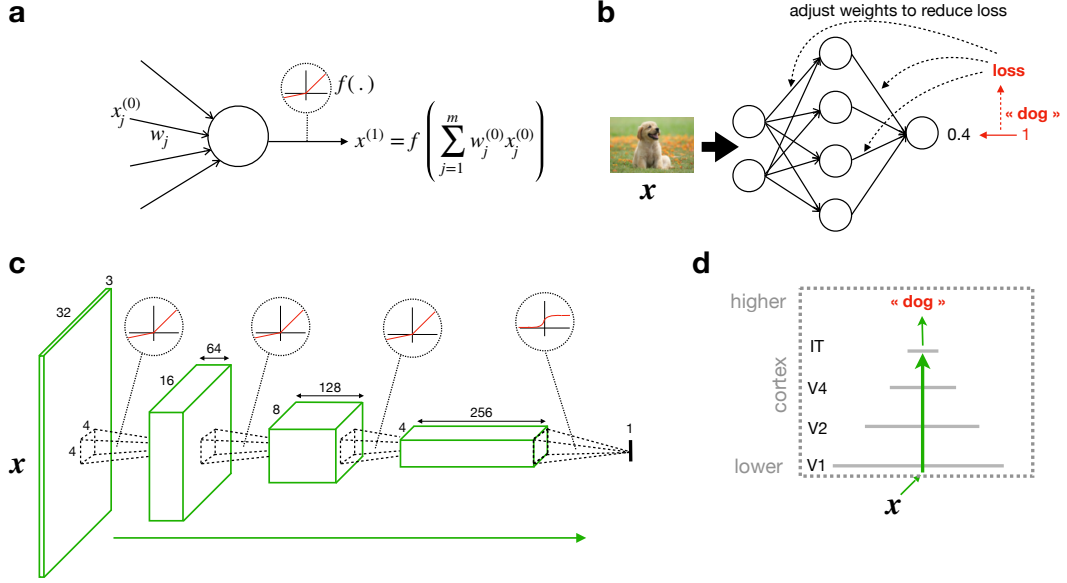
FIGURE 1.2: **The deep learning framework. (a)** Deep learning models are composed of units that receive a weighted sum of multiple inputs x_j and apply a non-linear transformation $f(.)$, here a LeakyReLU, peace-wise linear function (Maas et al., 2013). **(b)** By stacking several layers of such units together, we obtain a multi-layer perceptron (MLP). Here, the MLP receives an image as input and outputs the probability that it belongs to a certain category ("dog"). The classification loss, i.e., the mismatch between output prediction and actual target, is backpropagated through the network to improve the network performance at this task. **(c)** A convolutional neural network (CNN) is made of convolutional layers where each unit is locally connected to the previous layer by a kernel that slides along the layer. **(d)** Goal-driven modeling approaches (Yamins et al., 2014; Yamins and DiCarlo, 2016) propose that CNNs can be mapped to the ventral stream where each layer corresponds to a cortical area.

where $w_j^{(0)}$ is the weight of the jth synaptic input $x_j^{(0)}$, $x^{(1)}$ is the activation obtained from a non-linear function $f(.)$ applied to the weighted-sum of synaptic inputs.

Several units $(x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)})$ can then be aligned to form a layer, where each unit receives input from the neurons of another layer $(x_1^{(0)}, x_2^{(0)}, \ldots, x_m^{(0)})$ following Eq. 1.1. By stacking these several layers together, we obtain a multilayer perceptron (MLP, Fig. 1.2b, Rosenblatt, 1961) which can be trained end-to-end (i.e., by only using an input and its associated target output) using the backpropagation algorithm (Rumelhart et al., 1986; LeCun et al., 1989, 2015). This consists of minimizing a loss function $\mathcal{L}(\boldsymbol{w})$, defined on the network output layer, and measuring the task-specific performance of the network, via gradient descent, i.e., by modifying the network weights in the opposite direction of the gradient of the loss:

$$\boldsymbol{w} := \boldsymbol{w} - \alpha \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}) \tag{1.2}$$

where $\alpha > 0$ is the learning rate. The backpropagation algorithm computes the gradient $\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w})$ with respect to each network weight by the chain rule, iterating backward from the last layer (Fig. 1.2b).

For example, we consider the image classification between images of cats and dogs with a feedforward network containing three layers: an input layer receiving the pixel image, one hidden layer processing the output from the input layer, and an output

unit that returns the probability that the image is a dog (Fig. 1.2b). The loss measures how well we performed in this prediction by comparing the output unit to the label information ("this is a dog" if it is a dog image). The goal of training consists of minimizing this loss such that for the next presented image of a dog, the output will predict the dog category. Backpropagation (Rumelhart et al., 1986; LeCun et al., 1989, 2015) aims to minimize this loss by adjusting the network's weights for each layer. These adjustments are obtained from the computation of the gradient of the loss function with respect to each weight by applying the chain rule. Through this simple principle, backpropagation is so far the most successful way to learn in deep networks (Lillicrap et al., 2020).

1.2.3.2   A deep learning framework for neuroscience

ANNs and the backpropagation algorithm present an efficient way to train an entire network to satisfy an objective function (i.e., minimize a loss function). Recent computational neuroscience research suggested that cortical networks could be trained through a similar principle, based on three components: objective functions, learning rules and architectures (Richards et al., 2019). The objective function describes the goal of the task to learn (measured by a loss function in ANNs), the learning rule describes how synapses are adapted to improve the objective function (for ANNs, via backpropagation) and the architecture describes how units are connected and which operation they perform within the network. This goal-driven approach allows to develop new functional theories on how the observed millions of neurons in brain structure coordinate to achieve a complex task, through individual synaptic changes governed by a global objective.

The main flaws of this approach is that in order to learn an objective function efficiently, cortical learning rules should implement the backpropagation algorithm to transport errors, which was for a long time considered as biologically questionable (Whittington and Bogacz, 2019). However, recent work has suggested cortical models rendering backpropagation more biologically plausible (reviewed in Whittington and Bogacz, 2019; Lillicrap et al., 2020).

To give an example of the issue with error-backpropagation, let's consider the error $\boldsymbol{\delta}_l$ at layer l given by the following recursive formula:

$$\boldsymbol{\delta}_l = (W_{l+1}^T \boldsymbol{\delta}_{l+1}) \circ f'(\boldsymbol{a}_l), \tag{1.3}$$

where $\boldsymbol{a}_l$ is the vector of activations of the layer l (before applying the non-linearity f in Eq. 1.1), and W_{l+1}^T is the transpose of the weights projecting to the layer $l +$ 1. As shown in Eq. 1.3, to backpropagate errors one uses the same weights in the backward pass as in the forward pass, imposing identical synaptic connections in both directions between cortical neurons, that is biologically unplausible. However, it has been shown that networks with fixed random feedback connections could also backpropagate errors (Lillicrap et al., 2016; Guerguiev et al., 2017).

Together, these recent developments arguing that backpropagation could be biologically plausible build a foundation to a new computational approach that uses the deep learning framework to study complex behaviors in neural circuits.

1.2.3.3   Convolutional neural network as a model of the visual cortex

Even though successful on simple tasks, MLPs are fully-connected, i.e., each unit in one layer is connected to all units in the next layer, which makes training inefficient

when trained on complex and high-dimensional data such as images, as they contain too many parameters. To circumvent this problem, convolutional neural networks (CNNs, LeCun et al., 2015) highly reduce the number of connections as layers are locally connected by a convolution kernel that only connects a subset of units from the layer below (Fig. 1.2c). As the kernel slides along the layer, the convolution operation produces a feature map that contributes to the input for the next layer. Through this specific architecture, CNNs are tolerant to image translations and contain much fewer parameters than fully-connected networks, and consequently marked the beginning of the deep learning revolution by beating records in image classification (Krizhevsky et al., 2012).

CNNs were initially inspired by biological processes in that cortical neurons from the visual cortex respond to stimuli only in a restricted region of the visual field, the receptive field (Fukushima, 1980). For instance, in order to detect local edges or shapes, each neuron from one layer only needs to be connected to local patches of the image. The receptive field size however increases as the signal passes through successive convolutional layers.

Moreover, CNNs trained on natural images tend to produce features that are qualitatively similar to those found in the ventral stream of the visual cortex. Early layers develop Gabor-like features as in V1, while higher layers respond to partial object features and eventually global features such as faces as in the IT cortex (Lindsay, 2021). The CNN structure can thus be mapped to the architecture of visual cortex: each convolution-nonlinearity motif can be considered as an approximation to a single visual area (Fig. 1.2d). This motivated a few studies to quantify the similarities between CNNs and biological networks, by learning a linear mapping between the activity of artificial units of CNNs to the activity of real neurons in the visual cortex. One main finding is that CNNs which internal representations best predicted IT activity tend to perform better at object recognition (Yamins et al., 2014). Moreover, the top hidden layers of these models turned out to be the most accurate model of neural responses in IT cortex. Together, these results suggested that CNNs are strong candidates for a computational framework of sensory learning (Yamins and DiCarlo, 2016).

Considering recent evidence showing that cortical representations present similarities with learned features from artificial networks, that backpropagation could be biologically plausible, and the convenience of the deep learning framework for learning complex tasks in the biological system, it is of our crucial interest to suggest new hypotheses based on these principles to explain how the brain constructs sensory representations throughout development.

1.3   Unsupervised learning in the brain: learning by generating sensory inputs

Despite their ability to describe cortical function, most AI-inspired brain models presented so far rely heavily on labelled data during training, i.e., each sensory input requires an additional teaching signal (e.g., explicitly indicating the category of the observed input). In the natural world, these "supervision" signals are scarce, and human and other animals do not receive millions of labels during development (Bergelson and Swingley, 2012; Bergelson and Aslin, 2017; Slone and Johnson, 2015; Lindsay, 2021). Infants are not systematically taught that the object they observe

belongs to a certain category, but only get this information occasionally, from which they easily generalize to other instances. Therefore, to better characterize animal learning, deep learning theories of cortical function should implement objectives that do not require a huge amount of labeled data.

1.3.1 Unsupervised learning in the brain

As an alternative to supervised learning, unsupervised learning algorithms have drawn increasing attention for their ability to learn without human labelling, i.e., only from statistics of the data (Liu et al., 2021). Similarly to supervised algorithms, most implementations use of a feedforward process, or encoder, transforming a high-dimensional input into a low-dimensional output (Goodfellow et al., 2016). However, in contrast with supervised algorithms, this output, or "latent representation" is not made to match specific targets (ex: "dog", "cat" categories). It aims to discover patterns and statistical regularities within the data by itself, i.e., from objectives that do not use human annotations.

What are these representations useful for? In machine learning, they can be used to group unlabeled data based on their similarities or differences (unsupervised clustering, Caron et al., 2018). They can also serve as a pre-trained basis to make subsequent (supervised) tasks easier, such as object classification or detection (Bengio et al., 2013). For instance, a supervised linear classifier trained on top of these representations could benefit from the unsupervised learning process that made object categories or attributes more easily separable.

In the ventral stream of the visual cortex, sensory representations might not have been acquired from specific supervised tasks (e.g., classifying between dog and cats), but from a self-organizing process disentangling high-level information from perceived sensory inputs (Zhuang et al., 2021). The brain might thus make use of unsupervised learning objectives to construct these representations. These can be used to learn new tasks quickly with simple neuronal read-outs in high cortical areas, instead of re-training the whole cortical structure for each task from scratch. For instance, the facility by which infants learn to speak might depend on pre-learned cortical representations that already disentangle voice frequencies into syllables (Gennari et al., 2021). These representations would have been learned in an unsupervised fashion since birth, through the constant exposure of spoken words from their parents (Bergelson and Swingley, 2012).

Together, these considerations encourage us to explore unsupervised learning mechanisms that the brain could implement.

1.3.2 The brain as a generative model

In this section, we will present the general principles of generative models, an important sub-class of unsupervised algorithms (Bond-Taylor et al., 2021). The central idea is to learn a model $p_\theta(\boldsymbol{x})$ whose samples $\boldsymbol{x} \sim p_\theta(\boldsymbol{x})$ belong to the same distribution as the data distribution $p(\boldsymbol{x})$ that is usually unknown. This consists of first collecting a large amount of data $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\} \sim p(\boldsymbol{x})$ from some domain (images, sentences or sounds) and then training the model to generate samples that resemble this data.

1.3.2.1   Maximizing the likelihood of sensory data

Learning a generative model usually relies on the principle of maximum likelihood which consists of choosing the parameters θ that maximize the likelihood of the model under the data. For the dataset $\mathcal{D}$ of independent and identically distributed inputs, this consists of maximizing the product of probabilities

$$\theta = \arg\max_{\theta} \prod_{i=1}^{N} p_\theta(\boldsymbol{x}_i) = \arg\max_{\theta} \sum_{i=1}^{N} \log p_\theta(\boldsymbol{x}_i), \qquad (1.4)$$

where we applied the monotically increasing log operator to turn the product into a sum over examples. We thus attempt to find the parameters θ that maximize the sum of the log-probabilities assigned to the data by the model (Kingma and Welling, 2019).

However, it is almost impossible to define a simple model $p_\theta(\boldsymbol{x})$ that could capture the complex distribution of high-dimensional data like images. To alleviate this issue, latent variable models (Bishop, 1998) supplement the model distribution $p_\theta(\boldsymbol{x})$ with an additional latent, or hidden, distribution $p(\boldsymbol{z})$ from which data can be generated via the likelihood distribution $p_\theta(\boldsymbol{x}|\boldsymbol{z})$. The model distribution over the observed variables is then obtained by marginalizing over the latent variables $\boldsymbol{z}$:

$$p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} = \int p_\theta(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z}. \qquad (1.5)$$

Thus, $p_\theta(\boldsymbol{x})$ is a mixture distribution where each component $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ weighted according to $p(\boldsymbol{z})$. The main goal is then to find the likelihood distribution $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ that can generate a realistic sample $\boldsymbol{x}$ out of a variable $\boldsymbol{z}$. This model is usually implemented by a deep neural network, or generator network, whose structure inverts the feedforward's architecture previously introduced (Fig. 1.3a). For instance, in the case of CNNs, a feedback architecture would consist of stacked transposed convolution layers (Dumoulin and Visin, 2016), that increase the spatial dimensions of intermediate feature maps at each level, until reaching the image dimensions (Fig. 1.3b). The main advantage of deep latent variable models is that while the marginal distribution $p_\theta(\boldsymbol{x})$ can be arbitrarily complex, the prior $p(\boldsymbol{z})$ and likelihood $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ can be relatively simple. For instance, the prior is often assumed to follow a Gaussian unit distribution $p(\boldsymbol{z}) \sim \mathcal{N}(0, 1)$ and the likelihood can be modeled by a Gaussian distribution whose mean and variance are parametrized by a deep generative network taking $\boldsymbol{z}$ as input, i.e., $p_\theta(\boldsymbol{x}|\boldsymbol{z}) \sim \mathcal{N}(\boldsymbol{\mu}_{G(\boldsymbol{z})}, \mathrm{diag}(\boldsymbol{\sigma}^2_{G(\boldsymbol{z})}))$ (Kingma and Welling, 2019). All the complexity of the data distribution is then captured within the generative network.

1.3.2.2   Variational inference

Maximizing the likelihood of data under the model requires to evaluate $p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) dz$ which is generally intractable because it requires an integration over all latent variables. A solution is to find the $\boldsymbol{z}$ that yields $\boldsymbol{x}$ through the generative distribution $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ by finding the posterior distribution $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ via the Bayes rule

$$p_\theta(\boldsymbol{z}|\boldsymbol{x}) = \frac{p_\theta(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z})}{p_\theta(\boldsymbol{x})} \qquad (1.6)$$

but because it contains the marginal $p_\theta(\boldsymbol{x})$ in the denominator, this distribution is also intractable. The idea of variational inference is to find an approximation of $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ with an approximate posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ with parameters ϕ, such that
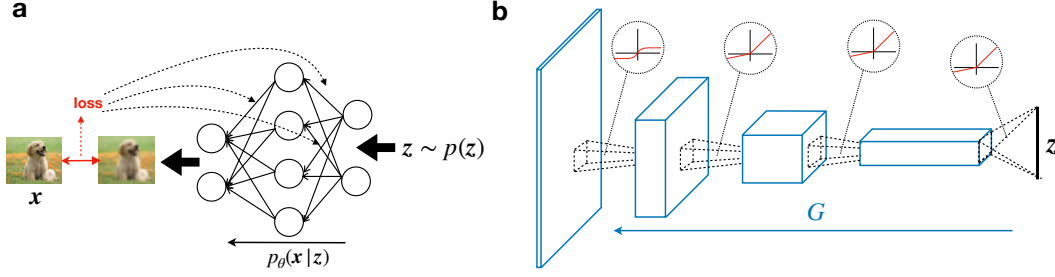
FIGURE 1.3: **Generative network. (a)** A generative model requires to specify how latent variables $\boldsymbol{z}$ are related to observations $\boldsymbol{x}$. The function that links both variables can be implemented by a deep network that takes a latent variable $\boldsymbol{z}$ as input and outputs a data point x that is made to resemble a data point from the training set. This is learned for instance via a reconstruction loss (red double arrow) backpropagated through the generative network. **(b)** This network can have a deconvolutional structure, composed of stacked convolution layers and non-linearities, that mirrors the architecture of a feedforward CNN, increasing spatial dimension of feature maps at each level until reaching the image dimension.

$q_\phi(\boldsymbol{z}|\boldsymbol{x}) \approx p_\theta(\boldsymbol{z}|\boldsymbol{x})$, turning an integration problem into an optimization problem. This can be done by minimizing the Kullback-Leibler distance (Joyce, 2011) that measures how the two distributions are different from each other, defined by:

$$D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p_\theta(\boldsymbol{z}|\boldsymbol{x})) = \int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log \frac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})} dz \tag{1.7}$$

that still contains the intractable posterior $p_\theta(\boldsymbol{z}|\boldsymbol{x})$. This divergence can then be re-written as

$$D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p_\theta(\boldsymbol{z}|\boldsymbol{x})) = \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log q_\phi(\boldsymbol{z}|\boldsymbol{x}) - \log p_\theta(\boldsymbol{x}|\boldsymbol{z}) - \log p(\boldsymbol{z})] + \log p_\theta(\boldsymbol{x}) \tag{1.8}$$

by applying Eq. 1.6. By re-arranging the terms in Eq. 1.8, we obtain:

$$\log p_\theta(\boldsymbol{x}) - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p_\theta(\boldsymbol{z}|\boldsymbol{x})) = \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z})). \tag{1.9}$$

As the KL-divergence term $D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p_\theta(\boldsymbol{z}|\boldsymbol{x}))$ is non-negative, we obtain the evidence lower bound (ELBO) on the log-likelihood of the data:

$$\log p_\theta(\boldsymbol{x}) \geq \mathcal{L}_{\theta,\phi}(\boldsymbol{x}) \tag{1.10}$$

where

$$\begin{aligned} \mathcal{L}_{\theta,\phi}(\boldsymbol{x}) &= \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z})) \\ &= \log p_\theta(\boldsymbol{x}) - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p_\theta(\boldsymbol{z}|\boldsymbol{x})). \end{aligned} \tag{1.11}$$

The goal of variational inference is to maximize the ELBO (Eq. 1.11), which in turn approximately maximizes the marginal likelihood $\log p_\theta(\boldsymbol{x})$ due to the inequality in Eq. 1.10, and minimizes the KL-divergence between the approximate posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and the true posterior $p_\theta(\boldsymbol{z}|\boldsymbol{x})$.

1.3.2.3   Representation learning via approximate inference

Generative models can be used for unsupervised learning by leveraging the inference process $q_\phi(z|x)$ that associates a data point x to its latent representation z (Bond-Taylor et al., 2021). As these representations z are tuned to generate an input x through the generative process $p_\theta(x|z)$, they potentially capture factors of variation (e.g., having two legs, ears shape, etc.) on data points x that would facilitate the learning of subsequent tasks (e.g., object classification or detection) (Hinton et al., 1995; Rao and Ballard, 1999; Donahue et al., 2016). For instance, the image of a dog x could be generated by an internal representation z whose elements contain semantic attributes about the dog (four legs, fur, etc.). Inferring such a representation would then facilitate the classification task.

In neuroscience, the idea that the brain learns internal representations of the world through a generative model of the sensorium has been around for decades (Barlow et al., 1961; Gregory, 1980; Rao and Ballard, 1999; Friston, 2010; Keller and Mrsic-Flogel, 2018; Gershman, 2019). The core idea is that feedback pathways implement a generative process predicting the upcoming inputs from latent representations. In parallel, the feedforward pathway, as described earlier, performs inference of the latent activity z associated to the observed input x. Following this general principle, we next highlight the different models that were proposed to explain sensory learning in the brain.

1.3.3   Explicit generative models: reconstructing sensory inputs

1.3.3.1   Predictive coding

The predictive coding (PC) framework (Rao and Ballard, 1999) has long been considered as the reference for generative learning in the brain and has served as a theoretical foundation of major neuroscience theories (Friston, 2005; Clark, 2013), even though it received criticism (Koch and Poggio, 1999; Murray et al., 2004). The idea is that the brain systematically minimizes prediction errors between top-down generated inputs and input actually received (Fig. 1.2c). Mechanistically, feedback connections from higher areas are thought to carry predictions of lower-level activities. In parallel, bottom-up connections send the prediction errors to adjust generative connections and to find a high-level activity compatible with the observed input (Fig. 1.4a).

The PC framework has been considered as a special case of variational inference (Friston, 2005; Millidge et al., 2022; Marino, 2022), by assuming that likelihood and prior distributions have gaussian densities

$$p_\theta(x|z) = \mathcal{N}(G(z), \text{diag}(\sigma_x^2)) \tag{1.12}$$

$$p(z) = \mathcal{N}(\mu_z, \text{diag}(\sigma_z^2)), \tag{1.13}$$

where G is a deep generative network, μ_z is the prior mean, and σ_x^2 and σ_z^2 are vectors of variance. Predictive coding aims to infer the z^* that maximizes the posterior distribution $p_\theta(z|x)$, which corresponds to maximizing the ELBO (Eq. 1.11) by choosing a delta distribution centered in z^* for the approximate posterior:

$$z^* = \arg\max_z p_\theta(z|x) = \arg\max_z \frac{p_\theta(z,x)}{p_\theta(x)} = \arg\max_z p_\theta(x,z) \tag{1.14}$$

$$= \arg\max_z p_\theta(x|z)p(z). \tag{1.15}$$
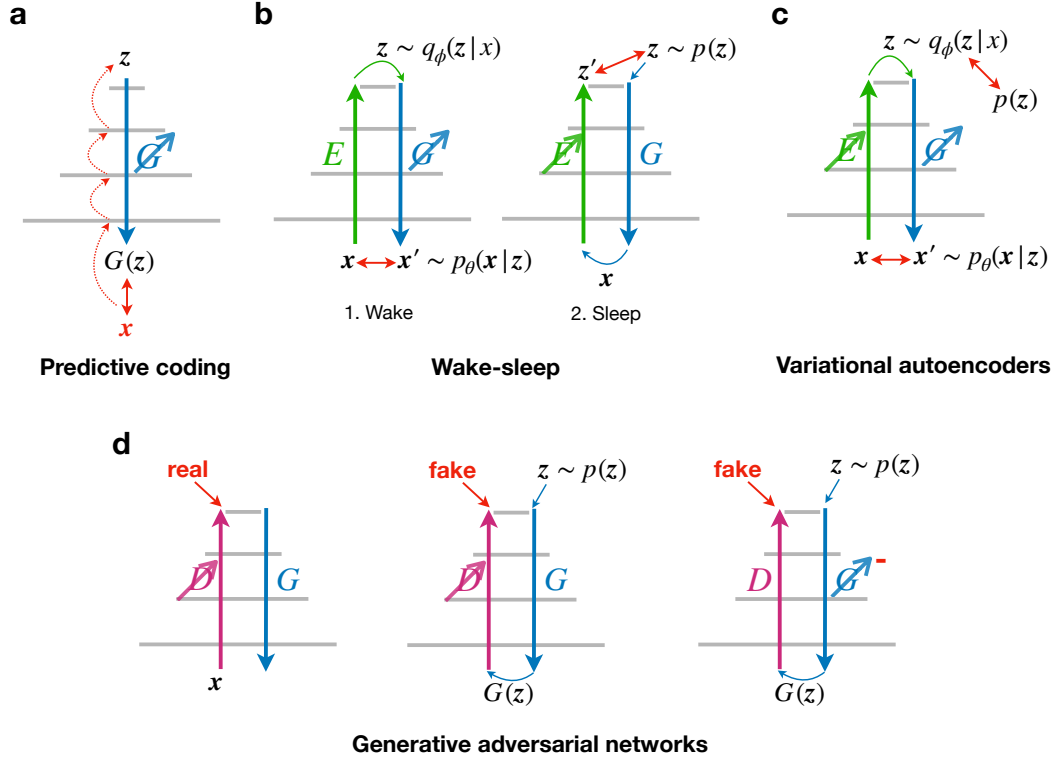
FIGURE 1.4: **Potential generative models for learning in the brain.** Here, we review the main generative models and their potential arrangement into cortical architecture (grey horizontal bars indicate cortical areas across the hierarchy). An oblique arrow ($\nearrow$) indicates that learning occurs in a network. **(a)** Predictive coding (Rao and Ballard, 1999) propose a multi-layer, hierarchical architecture of the cortex where feedback connections (G, blue) carry predictions of neural activity at the lower level, whereas feedforward pathways (red dotted arrows) carry prediction errors between top-down prediction and actual activity. **(b)** The wake-sleep algorithm (Hinton et al., 1995) introduces an inference network E that learn to invert the generator G. Wake: data $\boldsymbol{x}$ is fed through E and G tries to reconstruct x from the inferred latent activity $\boldsymbol{z}$. Sleep: latent activity $\boldsymbol{z}$ is sampled from the prior distribution and G generate the associated samples. E is trained to reproduce the latent activity. **(c)** Variational autoencoders (VAEs) (Kingma and Welling, 2013) train E and G simultaneously to reproduce input $\boldsymbol{x}$ using the reparametrization trick making the latent sampling operation differentiable. The learned, approximate posterior distribution is forced to match the prior $p(\boldsymbol{z})$. **(d)** Generative adversarial networks (GANs) (Goodfellow et al., 2014) train a discriminator (D) to distinguish between real data $\boldsymbol{x}$ ("real", left) and generated samples $G(z)$ ("fake", middle) against a generator (G) that tries to fool D into believing that these samples are real (right).

By applying the monotically increasing log(.) function, we obtain

$$\boldsymbol{z}^* = \arg\max_{\boldsymbol{z}} \left[\log \mathcal{N}(G(\boldsymbol{z}), \mathrm{diag}(\boldsymbol{\sigma}_{\boldsymbol{x}}^2)) + \log \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{z}}, \mathrm{diag}(\boldsymbol{\sigma}_{\boldsymbol{z}}^2)) \right] \qquad (1.16)$$

$$= \arg\min_{\boldsymbol{z}} \left[\frac{1}{2} \left\| \frac{\boldsymbol{x} - G(\boldsymbol{z})}{\boldsymbol{\sigma}_{\boldsymbol{x}}} \right\|^2 + \frac{1}{2} \left\| \frac{\boldsymbol{z} - \boldsymbol{\mu}_{\boldsymbol{z}}}{\boldsymbol{\sigma}_{\boldsymbol{z}}} \right\|^2 \right]. \qquad (1.17)$$

The predictive coding model thus requires to infer a latent activity $\boldsymbol{z}^*$ by minimizing the reconstruction error between the generated prediction $G(\boldsymbol{z})$ and the actual sensory input $\boldsymbol{x}$ (first term of Eq. 1.17). The second term is a regularization term that constrains the activity to match the prior mean.

Once latent activities have converged, a gradient step is taken on the generator weights to optimize the following objective:

$$\max_G p_\theta(\boldsymbol{x}|\boldsymbol{z}^*) = \min_G \frac{1}{2}\left\|\frac{\boldsymbol{x} - G(\boldsymbol{z}^*)}{\boldsymbol{\sigma_x}}\right\|^2 \qquad (1.18)$$

minimizing the reconstruction error between predicted and actual inputs.

Conceptually, predictive coding claims that both perceptual inference and learning in the brain are operationalized via the minimization of reconstruction errors, first via an optimization of neuronal firing rates on a fast timescale (Eq. 1.17) and then by the optimization of synaptic weights on a slow timescale (Eq. 1.18) (Rao and Ballard, 1999; Millidge et al., 2022). Due to the multi-layer (deep) structure of G, the latent representations learned from this model could extract high-level features from the sensorium, and were shown to present similarities with biological receptive fields (Rao and Ballard, 1999). However, this model relies on gradient-based optimization to perform inference of the latent activity $\boldsymbol{z}^*$ (or, equivalently, the parameters of the approximate posterior) for each sensory input $\boldsymbol{x}$ (Eq. 1.17), that might be too slow considering how fast animals infer sensory inputs (Kingma and Welling, 2019; Marino, 2022).

1.3.3.2   Wake-sleep algorithm

To alleviate the need to perform gradient descent during inference, Helmholtz machines (Dayan et al., 1995), introduce a feedforward recognition, or encoder network to amortize inference, i.e., to model the approximate posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$. For instance, in the case of a Gaussian approximate posterior,

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}_{E(\boldsymbol{x})}, \text{diag}(\boldsymbol{\sigma}^2_{E(\boldsymbol{x})}))\,, \qquad (1.19)$$

where E is a deep neural network with two heads, one for the mean $\boldsymbol{\mu}_{E(\boldsymbol{x})}$ and the other for the variance $\boldsymbol{\sigma}^2_{E(\boldsymbol{x})}$ of the approximate posterior (Fig. 1.4b-c). In parallel, the likelihood distribution $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ is still parametrized by a generative network G. The goal the Wake-Sleep (WS) algorithm (Hinton et al., 1995) is to learn these two distributions by splitting the training of the recognition (E) and generative (G) networks into two phases (Fig. 1.4b).

Wake In the Wake phase (Fig. 1.4b, left), the encoder weights ϕ are fixed. Suppose that we sample a minibatch of inputs $\boldsymbol{x} \sim p(\boldsymbol{x})$. The goal is to maximize the ELBO (Eq. 1.11) according to the generative weights θ

$$\max_\theta \mathcal{L}_{\theta,\phi}(\boldsymbol{x}) = \max_\theta \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] \qquad (1.20)$$

For a Gaussian likelihood with fixed variance, this corresponds to minimizing the following loss:

$$\min_G \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\|G(\boldsymbol{z}) - \boldsymbol{x}\|^2\,. \qquad (1.21)$$

This objective is optimized by passing sensory inputs $\boldsymbol{x}$ through the encoder network E, sampling latent activities $\boldsymbol{z}$ from the encoder output, and training the generative network G to reconstruct the sensory inputs $\boldsymbol{x}$ from the latent activities $\boldsymbol{z}$ (note the similarity with PC, Eq. 1.18).

Sleep In the Sleep phase (Fig. 1.4b, right), the generative weights θ are fixed. The goal is to maximize the ELBO (Eq. 1.11) according to the encoder (recognition) weights ϕ. This corresponds to minimizing the KL-divergence between the approximate and the true posterior:

$$\min_{\phi} \mathcal{L}_{\theta,\phi}(\boldsymbol{x}) = \min_{\phi} D_{KL}(q_{\phi}(\boldsymbol{z}|\boldsymbol{x}) \parallel p_{\theta}(\boldsymbol{z}|\boldsymbol{x})). \tag{1.22}$$

However, this minimization is generally intractable so instead, the sleep phase minimizes the KL-divergence the other way around

$$\begin{aligned} \min_{\phi} D_{KL}(p_{\theta}(\boldsymbol{z}|\boldsymbol{x}) \parallel q_{\phi}(\boldsymbol{z}|\boldsymbol{x})) &= \min_{\phi} \mathbb{E}_{\boldsymbol{z} \sim p_{\theta}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\theta}(\boldsymbol{z}|\boldsymbol{x}) - \log q_{\phi}(\boldsymbol{z}|\boldsymbol{x})] \\ &= \max_{\phi} \mathbb{E}_{\boldsymbol{x},\boldsymbol{z} \sim p_{\theta}(\boldsymbol{x},\boldsymbol{z})}[\log q_{\phi}(\boldsymbol{z}|\boldsymbol{x})] \end{aligned} \tag{1.23}$$

supposing that we sample $\boldsymbol{z} \sim p(\boldsymbol{z})$ and $\boldsymbol{x} \sim p_{\theta}(\boldsymbol{x}|\boldsymbol{z})$. For a Gaussian approximate posterior with fixed variance, this corresponds to minimizing the reconstruction loss

$$\min_{E} \mathbb{E}_{\boldsymbol{x} \sim p_{\theta}(\boldsymbol{x}|\boldsymbol{z})} \|\boldsymbol{z} - E(\boldsymbol{x})\|^2. \tag{1.24}$$

This objective is optimized by passing the sampled latent activities $\boldsymbol{z}$, or "fantasies" through the generative network and training the encoder E to reconstruct this latent activity.

Computationally speaking, this process allows to learn an amortized approximate posterior $q_{\phi}(\boldsymbol{z}|\boldsymbol{x})$ of the generative model $p_{\theta}(\boldsymbol{x}|\boldsymbol{z})$ by training the inference network E in the sleep phase, and thus allows direct inference when a sensory input $\boldsymbol{x}$ is presented, in contrast to predictive coding. However, WS only trains the inference network on generated samples $G(\boldsymbol{z})$ that might not resemble the actual data $\boldsymbol{x}$, which can be detrimental especially at the beginning of training where where the model does not produce samples that resemble those from the data distribution.

1.3.3.3   Variational autoencoders

Similar to Wake-Sleep, variational autoencoders (VAE) (Kingma and Welling, 2013) also introduce an encoder network to parametrize the approximate posterior distribution $q_{\phi}(\boldsymbol{z}|\boldsymbol{x})$ to maximize the ELBO from Eq. 1.9

$$\mathcal{L}_{\theta,\phi} = \mathbb{E}_{\boldsymbol{x} \sim q_{\phi}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\theta}(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}(q_{\phi}(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z})), \tag{1.25}$$

except that both encoder and generator networks are trained jointly via stochastic gradient descent. The idea is to make the latent variable $\boldsymbol{z}$ differentiable by introducing an auxiliary random variable $\boldsymbol{\epsilon}$ that does not depend on $\boldsymbol{x}$ or ϕ, referred to as the reparametrization trick (Kingma and Welling, 2013, 2019). For instance, in the case of a Gaussian posterior, $\boldsymbol{z} \sim q_{\phi}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}_{E(\boldsymbol{x})}, \text{diag}(\boldsymbol{\sigma}^2_{E(\boldsymbol{x})}))$ can be rewritten as $\boldsymbol{z} = \boldsymbol{\mu}_{E(\boldsymbol{x})} + \boldsymbol{\sigma}^2_{E(\boldsymbol{x})}\boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$. By considering a Gaussian likelihood distribution with fixed variance, the VAE objectives becomes

$$\min_{E,G} \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x}), \boldsymbol{z} \sim q_{\phi}(\boldsymbol{z}|\boldsymbol{x})} \|G(\boldsymbol{z}) - \boldsymbol{x}\|^2 + D_{KL}(q_{\phi}(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z})), \tag{1.26}$$

where $q_{\phi}(\boldsymbol{z}|\boldsymbol{x})$ is encoded by the network E.

The first term of Eq. 1.26 is reminiscent of the autoencoder (AE) loss function (Rumelhart and McClelland, 1987)

$$\min_{E,G} \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\|G(E(\boldsymbol{x})) - \boldsymbol{x}\|^2 \,, \tag{1.27}$$

explaining its name. Thus, VAEs can be seen as AEs but with the encoder output regularized to match the prior (second term of Eq. 1.26).

Compared to models introduced above, VAEs were shown to perform relatively well at reproducing data distribution and have recently been considered as potential brain models (van de Ven et al., 2020; Higgins et al., 2021; Marino, 2022). However, despite their success on simple datasets, generated samples tend to be blurry when trained on more complex datasets such as natural images, and do not perform as well as the so-called implicit generative models for representation learning, that we will consider next (Dosovitskiy and Brox, 2016; Berthelot et al., 2018; Bond-Taylor et al., 2021).

1.3.4   Implicit generative models: generative adversarial networks

1.3.4.1   Likelihood-free learning

The previously introduced models are explicit: they aim to maximize the likelihood of data under the model by learning to generate data points $\boldsymbol{x}$ from latent representations $\boldsymbol{z}$ via (if assuming Gaussian likelihood) element-wise reconstruction errors, potentially making the network sensitive to irrelevant variations from the sensorium, and limiting the possibility to generate new samples (Goodfellow, 2016). Implicit generative models, in contrast, do not specify the distribution of the data itself, but rather define a stochastic procedure that directly generates data without maximizing the likelihood or any derived quantities, i.e., without training the generator to reproduce a particular input.

1.3.4.2   Generative adversarial networks

Among implicit models, Generative Adversarial Networks (GANs, Goodfellow et al., 2014) introduce a binary classifier, or discriminator (D), that distinguishes between real data and generated samples from the generator network (Fig. 1.4e). The generator G is trained to fool the discriminator D into believing that its generated samples are real, i.e., to create samples that belong to the real data distribution. This can be formalized as the optimization of the following mini-max game:

$$V(D,G) = \min_{G} \max_{D} \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \,, \tag{1.28}$$

where $p(\boldsymbol{x})$ is the real data distribution, $p(\boldsymbol{z})$ is a prior distribution, e.g., $p(\boldsymbol{z}) \sim \mathcal{N}(0, I)$. This equation defines the cross-entropy loss for a binary classifier (D) with a sigmoid output, where the label is for instance 1 for all dataset examples $\boldsymbol{x}$, and 0 for all generated samples $G(\boldsymbol{z})$. The generator is trained adversarially to generate samples that would be mis-classified by the discriminator by minimizing the discriminator objective. By alternating the training of each network, the discriminator improves its ability to discern real from generated images, while the generator improves the quality of its generated samples so it can fool the (improved) discriminator classification.

1.3.4.3   Global optimality of GANs

Following Eq. 1.28, the optimal discriminator D for a fixed generator G is given by

$$
\begin{aligned}
D^*(\boldsymbol{x}) &= \arg\max_D V(D, G) \\
&= \arg\max_D \int_{\boldsymbol{x}} p(\boldsymbol{x}) \log(D(\boldsymbol{x})) + p_\theta(\boldsymbol{x}) \log(1 - D(\boldsymbol{x})) dx \,.
\end{aligned}
\tag{1.29}
$$

The maximum of this expression can be found by computing the derivative with respect to D:

$$
\nabla_D = \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x})} \left[\frac{1}{D(\boldsymbol{x})} \right] - \mathbb{E}_{\boldsymbol{x} \sim p_\theta(\boldsymbol{x})} \left[\frac{1}{1 - D(\boldsymbol{x})} \right]
\tag{1.30}
$$

that is canceled when

$$
D^*(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{p(\boldsymbol{x}) + p_\theta(\boldsymbol{x})} \,.
\tag{1.31}
$$

By plugging-in the optimal discriminator into the overall objective $V(D, G)$, we obtain the objective for G:

$$
\begin{aligned}
V(D^*, G) &= \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x})}[\log D^*(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x} \sim p_\theta(\boldsymbol{x})}[\log(1 - D^*(\boldsymbol{x}))] \\
&= \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x})}[\log \frac{p(\boldsymbol{x})}{p(\boldsymbol{x}) + p_\theta(\boldsymbol{x})}] + \mathbb{E}_{\boldsymbol{x} \sim p_\theta(\boldsymbol{x})}[\log \frac{p_\theta(\boldsymbol{x})}{p(\boldsymbol{x}) + p_\theta(\boldsymbol{x})}] - 2\log 2 \\
&= 2JS(p \parallel p_\theta) - \log 4
\end{aligned}
\tag{1.32}
$$

where

$$
JS(p \parallel q) = \frac{1}{2} KL(p \parallel \frac{p+q}{2}) + \frac{1}{2} KL(q \parallel \frac{p+q}{2}) \,.
\tag{1.33}
$$

is the Jensen-Shannon (JS) divergence between the probability distribution of real and fake data (Menéndez et al., 1997). As the KL-divergence, JS-divergence measures how distinguishable two probability distributions are, but is symmetric as it contains both KL-divergence and reverse-KL divergence terms. The global minimum is attained if this divergence is minimized (Eq. 1.32), i.e., if $p_\theta(\boldsymbol{x}) = p(\boldsymbol{x})$, in which case $V(D, G) = -\log 4$. In other words, this result shows that even though not defined on maximizing the data likelihood, the global optimality of GANs is attained when the generator perfectly replicates the data distribution.

However, optimizing such an objective is not straightforward as each player's objective is the opposite of the other player's objective. The goal is thus to find a local Nash equilibrium (Ratliff et al., 2013): a point that is local minimum of each player's objective with respect to that player's parameters. With local moves, no player can maximize its objective further, assuming that the other's players parameters are fixed (Goodfellow et al., 2020).

1.3.4.4   Training procedure

To guarantee a possible equilibrium, the most common training procedure for GANs consists of optimizing each player's objective alternately on a small subset of data, or minibatch, ensuring that one player does not win over the other one.

First, a minibatch $\boldsymbol{x}$ of inputs from the dataset is sampled and D is trained to classify these as real (Fig. 1.4e, left), i.e., maximizing the objective: $\max_D \frac{1}{B} \sum_{i=1}^{B} \log D(\boldsymbol{x}^{(i)})$, where B is the minibatch size.

Then, a minibatch of $\boldsymbol{z}$ latent vectors drawn from the generator's prior $p(\boldsymbol{z}) = \mathcal{N}(0, I)$ is sampled and fed to G. From the minibatch of generated images $G(\boldsymbol{z})$, D is trained to classify them as fake (Fig. 1.4e, middle), by maximizing the objective: $\max_D \frac{1}{B} \sum_{i=1}^{B} \log(D(1 - G(\boldsymbol{z}^{(i)})))$.

Finally, G is trained to generate an image that D would mistakenly classify as real (Fig. 1.4e, right) by minimizing D's objective: $\min_G \frac{1}{B} \sum_{i=1}^{B} \log(D(1 - G(\boldsymbol{z}^{(i)})))$. However, this objective might suffer from vanishing gradient when D is too close to its optimal value. Therefore, in practice the following objective is preferred for the generator: $\max_G \frac{1}{B} \sum_{i=1}^{B} \log D(G(\boldsymbol{z}^{(i)}))$. This corresponds to maximizing the probability that the discriminator classifies a generated input as real.

1.3.4.5 GANs in practice

The training steps described above are repeated many times until the generator produces samples that resembles the ones from the dataset. At the end of training, GANs are often able to generate realistic samples, even for complex datasets containing high-resolution images (Radford et al., 2015; Karras et al., 2018; Brock et al., 2019). Generated samples are more realistic, novel, sharper and usually match better the real data distribution than samples from other generative models, such as reported by Fréchet Inception Distance measurements (FID, Heusel et al., 2018). Furthermore, unlike explicit models, GANs do not leverage element-wise objectives to train the generator. The generator has no direct access to real data, and its training signal comes only through what the discriminator has learned, making it more resistant to overfitting and prone to synthesize novel data (Goodfellow, 2016).

However, GANs suffer from training issues such as non-convergence or mode collapse (Goodfellow, 2016; Bond-Taylor et al., 2021). Non-convergence emerges from the inability to finding an equilibrium to the two-player's game. Even if both players move downhill their respective loss functions, the same update might undo the other player's progress. Mode collapse occurs when the generator only learns to generate one mode of the probability distribution, for instance different views of the same cat. However, over the past years, many GANs-variants were proposed to overcome these issues, such as Wasserstein-GANs (WGANs, Arjovsky et al., 2017; Gulrajani et al., 2017), that minimize the Wasserstein distance, measuring how much "probability mass" should be moved to turn the generated distribution into the real one. Other techniques were also developed to stabilize training and improve generated samples, such as spectral normalization (Miyato et al., 2018) or data augmentations (Karras et al., 2020).

1.3.4.6 GANs and representation learning

Beside their ability to generate realistic samples, GANs are also known to extract useful representations from data. This was first observed within the discriminator features, on which a supervised linear classifier performed better than on other generative models' representations (Radford et al., 2015). This ability to extract semantic attributes from data is also reflected within the generator's latent space, where interpolations often lead to semantically-meaningful interpolations in the data space, such as creating objects combining features from two different objects (e.g., a dog and a

car) in a realistic manner (Radford et al., 2015; Brock et al., 2019). Furthermore, certain directions in this space correspond to particular semantic attributes (e.g., for human faces, gender, presence of eyeglasses, etc.).

However, unlike VAEs or WS, GANs lack an inference mechanism $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ mapping data $\boldsymbol{x}$ to latent representations $\boldsymbol{z}$ and thus not benefit from the learned latent space structure. Several models attempted to train an additional encoder to invert the generator of GANs and were shown to make GANs applicable to representation learning (Makhzani et al., 2015; Donahue et al., 2016; Dumoulin et al., 2017; Brock et al., 2019; Chen and Wilson, 2017; Ulyanov et al., 2017). For instance, combining VAE (or AE) and GANs objectives (Dosovitskiy and Brox, 2016; Brock et al., 2017) could benefit from the inference mechanism of the former and the generation quality of the latter to learn good representations from data.

1.3.4.7 GANs as a brain model?

A few authors suggested that the brain might learn from implicit principles such as adversarial learning rather than explicit, likelihood-based objectives (Gershman, 2019; Benjamin and Kording, 2021a). Their architecture rely on two networks where information flows in opposite direction, as for PC, WS and VAE. One could thus consider the discriminator network as part of the feedforward pathway of the visual cortex, inferring whether an input is externally driven or internally generated. In parallel, as for other generative brain models, the generator could be implemented by feedback pathways to generate internal inputs based on the discriminator backpropagated error (Fig. 1.4d).

Together, the presented models introduce computational principles that the brain could implement to learn representations of the world in an unsupervised fashion, ranging from old (Hinton, 1984; Rumelhart and McClelland, 1987; Hinton et al., 1995) to recent (Kingma and Welling, 2013; Goodfellow et al., 2014) models. We will see in Section 1.4 how these principles could be exploited in order to explain the role of sleep in learning.

1.3.5 A note on cognitive science – extracting semantic concepts from episodic memories

An interesting parallel can be made between representation learning and cognitive concepts of memory. In psychology, Tulving introduced in 1972 a distinction of declarative memory between episodic and semantic memory (Tulving, 1972). Episodic memory relates to personal experiences bound to a spatio-temporal context, about what happened, when and where (Tulving, 2002). In contrast, semantic memory refers to a general world knowledge about facts, objects, words or beliefs, independent of specific experiences or contextual information and arises from regularities and repetition in our experience. As an analogy, when we remember that yesterday I saw a yellow bicycle parked in front my house, we are drawing on episodic memory. However, when we state that bicycles are two-wheeled, with pedals and handle bars, we are drawing on semantic memory (Greenberg and Verfaellie, 2010).

While initially thought to be encoded by similar brain structures (McClelland et al., 1995), experimental evidence suggests that episodic memory is encoded in the hippocampus and semantic is neocortically represented (Nadel and Moscovitch, 1997; Burgess et al., 2002; Winocur and Moscovitch, 2011). As we will discuss below, the

creation of semantic memories is thought to involve episodic memories according to "transformation theories", or "semantization" (Nadel and Moscovitch, 1997; Rosenbaum et al., 2001; Meeter and Murre, 2004; Winocur et al., 2010). In this process, semantic memory can emerge from the gradual extraction of statistical regularities across distinct episodic memories, abstracting away from specific context and details, through a hippocampo-cortical dialogue.

We here note that the notion of semantic knowledge could be interpreted in terms of learning deep cortical representations as previously introduced. For instance, we consider that we form distinct episodic memories about the moments we observed cats in different context (street, garden, etc.). The idea of semantization is that the cortex gradually builds a semantic representation from these diverse episodes. This representation will keep being enriched by the accumulation of new episodes involving cats. Similarly, deep learning algorithms tend to learn representations that contains the "cat" information from the observation of multiple examples of cats.

1.4   Learning during sleep: reactivation of memories and dream generation

So far, we observed that the brain organizes its knowledge into high-level, semantic concepts, both at behavioral and neuronal levels. In accordance with cortical structure, deep generative models suggest learning principles that could potentially explain how semantic knowledge emerges through the development in an unsupervised manner. However, most of these ideas assume that sensory inputs are constantly perceived, and hardly consider their offline internal generation. Indeed, as soon as we lose attention, our mind starts to shift towards internal memories and self-generated experiences, the extreme of this occurring while we are asleep and dreaming.

1.4.1   What is sleep?

After a long day of work, followed by our favorite sport activity, ending up washing the dishes from dinner and reading the newspaper, we usually get invaded by a feeling of drowsiness that inevitably guides us to bed. Both our mind and body have interacted enough with the external world and deserve a peaceful time of inactivity, in perfect silence, lights off, so they can be fully recovered for the next long day to come. This daily ritual defines sleep: "a natural and reversible state of reduced responsiveness to external stimuli and relative inactivity, accompanied by a loss of consciousness" (Rasch and Born, 2013).

While during wakefulness, our experience is mainly influenced by external stimuli, upon falling asleep perception is reduced and action is ceased. Our body becomes only responsive to sensory stimuli that are strong, sudden or salient enough, in which case sleep is disrupted and we wake up. This strong disconnection with the environment is accompanied by loss of consciousness, although sleep still hosts conscious experiences that are mainly internally generated, ranging from spontaneous thoughts to hallucinations that characterize dreams (Nir and Tononi, 2010; Aru et al., 2020).

We all are aware of the importance of a good night of sleep in our daily life. The only observation that car accidents due to sleep deprivation, leading to drowsy driving, exceed by those by alcohol or other drugs demonstrate that sleep plays a crucial role in our waking functions (Tefft, 2018; Walker, 2017). Moreover, the lack of sleep
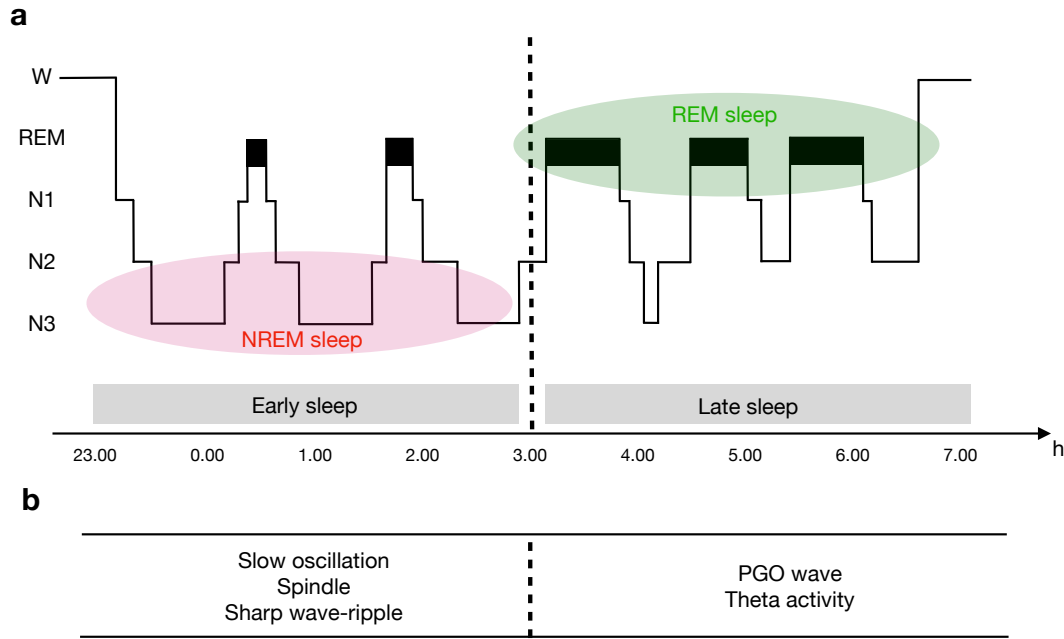
FIGURE 1.5: **Sleep phases and related EEG signatures.** Alternation between rapid-eye-movement (REM) sleep and non-REM (NREM) throughout the night. NREM sleep contains three sub-stages, slow-wave sleep or N3, and lighter stages N1 and N2. NREM sleep tends to be prominent during the first half of the night, and REM during the second half. **(b)** Sleep-related EEG signals. SWS is mainly characterized by neocortical slow oscillations (~ 0.8 Hz), thalamocortical spindles ($10 - 15$ Hz) and hippocampal sharp-wave ripples ($100 - 300$ Hz). REM sleep is hallmarked by ponto-geneiculo-occipital (PGO) waves and hippocampal theta ($4 - 8$ Hz) activity. Figure adapted from Rasch and Born (2013).

has been shown to have serious consequences on emotional functioning (Goldstein and Walker, 2014), metabolic regulation and obesity (Knutson et al., 2007), immune functions (Lange et al., 2010), restoration of energy (Benington and Craig Heller, 1995) and memory (Diekelmann and Born, 2010; Klinzing et al., 2019).

1.4.2 Physiological features of NREM and REM sleep

Sleep in mammals is subdivided into two essential stages: rapid-eye-movement (REM) sleep and non-rapid-eye-movement (NREM) sleep and alternate in a cyclic manner. NREM sleep tends to be dominant during the first half of the night and give way to REM sleep for the second half (Fig. 1.5a). In particular, within NREM sleep, a phase of deep sleep, or Slow-Wave Sleep (SWS) is characterized by slow and synchronized high-amplitude electro-encephalogram (EEG) oscillations. REM sleep, or paradoxical sleep, is marked by wake-like activity patterns, with fast and low-amplitude oscillatory brain activity, while muscle tone over the body is completely inhibited (muscle atonia), that led to its other designation "paradoxical sleep" (Rasch and Born, 2013).

Slow oscillations during NREM sleep originate in the neocortex at a frequency of ~ 0.8 Hz (Fig. 1.5b). They synchronize neuronal activity into down-states where neurons are globally hyperpolarized and silent, and up-states where neurons get depolarized and fire together (Buzsáki and Draguhn, 2004; Steriade, 2006). NREM is also characterized by 10-15 Hz oscillations called spindles in both N2 and N3 stages, originating in the thalamus, and sharp wave-ripples (SWR), which are high-frequency oscillations (100-300 Hz) originating in the hippocampus. SWR usually accompany

the reactivation of hippocampal neuronal representations that were active during the preceding waking experience and are thought to mediate memory consolidation (Nadasdy et al., 1999; O'Neill et al., 2010).

Besides its highly asynchronous, wake-like activity patterns, REM sleep presents typical EEG signatures such as Ponto-Geniculo Occipital (PGO) waves, which are internally triggered bursts of synchronized activity propagating from the pontine brainstem to the lateral geniculate nucleus and visual cortex, that occur concurrently with rapid eye movements in rats. PGO waves have been proposed to promote synaptic plasticity in the regions they reach (Datta, 1999) and to be associated with the internal generation of visual imagery during dreaming (Hobson et al., 2000; Gott et al., 2017). Theta (4-8 Hz) oscillations are also found in the hippocampus during REM sleep and are also thought to contribute to memory consolidation (Diekelmann and Born, 2010; Boyce et al., 2017).

Sleep stages are also distinguished by dramatic changes in activity levels of different neuromodulators. For instance, as compared to wakefulness, acetylcholine (ACh) levels strongly decrease during NREM and increase again to its waking levels during REM sleep (Rasch and Born, 2013).

Out of these physiological features, two main differences seem to stand out between these two stages. While neuronal replay is mainly observed during NREM sleep, REM sleep activity patterns are more random and less stereotyped, comparable to waking activity. This latter observation might coincide with the high prevalence of dreams during the REM state that rarely faithfully replay past memories.

1.4.3   Sleep and memory consolidation

Besides its importance for our survival, sleep has long been known to favour the consolidation of memories, process that transforms new and initially labile memories acquired during wakefulness into more stable representations integrated with pre-existing long-term memories (Diekelmann and Born, 2010). This idea was initiated in 1924 when Jenkins and Dallenbach (1924) tested two participants after learning lists of verbal facts. They observed that the recall of these memories was distinctly better when participants had slept after learning than if they had stayed awake. Since then, a vast amount of studies reported the importance of sleep for the retention of memories (reviewed in Diekelmann and Born, 2010; Rasch and Born, 2013; Dudai et al., 2015).

1.4.3.1   Sleep for consolidating memories via a complementary learning system

Initially, consolidation was attributed to the protective effect of sleep on newly encoded and still fragile memories, preventing them from being overwritten by new information (retroactive interference). However, the observation of neuronal replay during sleep suggested an active role of sleep in memory consolidation (McClelland et al., 1995; Diekelmann and Born, 2010; Rasch and Born, 2013; Klinzing et al., 2019). Following this active consolidation hypothesis, the "standard consolidation theory" (SCT) proposes that consolidation relies on complementary learning systems (CLS) where the hippocampus is a fast-encoding store that quickly encodes memories from the day, and the neocortex is a slow-learning store that retains memories for long-term. According to the SCT, the reactivation of hippocampal memories, that mainly occurs within hippocampal SWR during NREM sleep, mediates the redistribution and
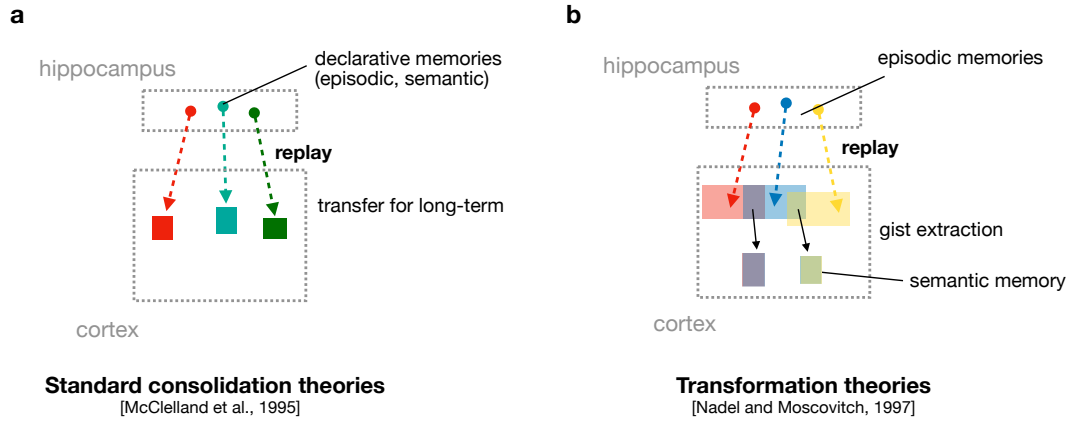
FIGURE 1.6: **Memory consolidation theories.** (a) Standard consolidation theory (McClelland et al., 1995) assumes two distinct memory stores, a fast-learning store, the hippocampus, and a slow-learning long-term store, the cortex. The hippocampus initially encodes memories during wakefulness, and replay them during subsequent NREM sleep within SWR. This replay leads to the gradual transfer of hippocampal memories to the neocortex for long-term storage. This transfer applies for both types of declarative memories (episodic and semantic) (b) Transformation theories (Nadel and Moscovitch, 1997; Winocur et al., 2010; Lewis and Durrant, 2011) proposes that the hippocampus mostly stores episodic memories, and that their reactivation during sleep leads to the extraction of the semantic overlapping "gist", thus transforming episodic hippocampal memories into semantic cortical representations for long-term.

consolidation of these memories in the neocortex for long-term. The hippocampus is then made free to encode new memories from future waking experiences.

The idea of consolidating memories through CLS was further explored by computational models proposing that memory replay during sleep prevents catastrophic forgetting (Shin et al., 2017; van de Ven et al., 2020). This issue characterizes the tendency of a neural network to forget previously learned tasks upon learning new tasks in a continual learning setting where tasks are presented sequentially. A trivial solution to overcome catastrophic forgetting would consist of storing previous experiences in the hippocampus, viewed as a memory buffer, and interleave new learning with the exact replay of these experiences. However, using stored exact data like pixels of an image is biologically implausible and requires an increasing amount of memory (Quiroga et al., 2008). As an alternative to storing data, some computational models instead propose that the hippocampus generate the data to be replayed with a generative network model sampling from random activity that has learned from past observations, using a GAN (Shin et al., 2017) or a VAE (van de Ven et al., 2020) paradigm. When a new task is learned, the hippocampus replays altered versions of previous experiences through this generative network along with new data. Even though relatively efficient at preventing catastrophic forgetting (Shin et al., 2017; van de Ven et al., 2020) these approaches assume that the hippocampus stores all past experiences, which contradicts with the general view of the hippocampus as a temporary store only encoding recent experiences.

1.4.3.2 Sleep for abstracting semantic concepts from episodic memories

Another caveat of the SCT (McClelland et al., 1995) is that it treats both episodic and semantic memories equivalently. Indeed, in this view, both types of declarative

memories (episodic or semantic) are initially encoded in the hippocampus and subsequently transferred in the neocortex for long-term. However, this theory has been challenged by evidence that preserved memories following hippocampal damage were more semantic in nature, indicating that the hippocampus mostly retains episodic memories while the cortex stores the semantic, decontextualized information (Nadel and Moscovitch, 1997).

From this observation, it was proposed that semantic cortical representations result from extractions of common elements within hippocampal episodic memories when being replayed. In this line, Trace Transformation Theory (TTT) (Nadel and Moscovitch, 1997; Winocur et al., 2010) suggests that the gist from multiple episodes is extracted during NREM sleep to form a cortical semantic representation. A cognitive model (Lewis and Durrant, 2011) proposed that semantic formation is based on the invariant overlapping and statistical regularities between replayed episodic memories, where areas of overlap are strengthened via Hebbian learning, allowing the abstraction of shared elements among these memories, or the semantic "gist". For example, the reactivation of various memories of "cat experiences" facilitates the extraction and consolidation the "cat" concept from repeating features with episodic memories (four legs, pointed ears, tail, etc.) in cortical representations (Section 1.3.5).

Such semantization effect was supported by experimental studies showing that infants tend to generalize word categories and grammatical structures over sleep (Friedrich et al., 2015; Gómez et al., 2006), that sleep facilitates category formation and learning of linguistic rules in adults (Batterink et al., 2014; Schapiro et al., 2017). Among these experiments, the Deese-Roediger-McDermott (DRM) paradigm, where participants learn a list of semantically related words (e.g., hospital, bandage, operation) with one common theme-word (e.g., doctor) missing. Participants are then asked, after some time spent awake or asleep, to recall the words from that list. In case the participants produced a false memory, i.e., remembering the gist word ("doctor"), this paradigm reveals that a semantic concept has been extracted from the list. Sleep has in fact been shown to enhance false memory formation (Payne et al., 2009; Pardilla-Delgado and Payne, 2017), while some studies report contradictory results (Fenn et al., 2009). Another experimental paradigm is the transitive inference task introduced by Ellenbogen et al. (2007) where participants learned a hierarchy of paired elements (A<B, B<C, C<D, D<E) but were unaware of the overall hierarchy (A<B<C<D<E). Participants were shown to make better inferential judgements (e.g., A<D) if they slept after learning these relations (Ellenbogen et al., 2007; Lau et al., 2010).

Overall, these findings confirm that sleep mediates the abstraction of semantic information from episodic memories and propose a differential memory system between the hippocampus and the neocortex. The TTT thus brings a fundamental view of the role of sleep in learning semantic representations and present similarities with the idea of deep unsupervised learning (discussed in section Section 1.3.5). However, the TTT presents a few shortcomings that remain to be pointed out.

First, the accompanying studies usually assume that semantization occurs over a single night of sleep, while this process should typically occur over a longer time period (Winocur et al., 2010; Sawangjit et al., 2018), potentially explaining the reported contradictory results (Payne et al., 2009; Fenn et al., 2009). Indeed, a study reported that the semantization effects on the DRM paradigm were noticeable only after one year (Lutz et al., 2017).

Second, the proposed models (Lewis and Durrant, 2011; Lewis et al., 2018) lack a mechanistic, circuit-level implementation. For example, it assumes that memories replayed together from the hippocampus share the same semantic content, which suggests that the semantic information is already available before its extraction.

Finally, TTT only considers replay during NREM sleep, but omits the possibility that REM sleep could facilitate this process, such as reported by a few studies (Cai et al., 2009; Djonlagic et al., 2009). For instance, the sequential hypothesis (Giuditta et al., 1995) highlights the importance of the cyclic succession of NREM and REM sleep for memory consolidation, with each sleep stage serving a complementary function, with NREM sleep retaining memories and REM sleep integrating them with preexisting memories. In particular, the role of bizarre dreams, mostly occurring during REM sleep, has not yet been explored extensively and deserves attention.

For the past decades, memory consolidation theories have emphasized the role neuronal reactivations in stabilizing memories, abstracting gist information from life events and constructing semantic cortical representations. However, these theories often omit the role of REM sleep and dreams as they mainly focus on neuronal replay during SWS. To this end, we will next explore the phenomenology of dreams and their suggested roles in healthy brain function, and from there draw a hypothesis on their roles in memory semantization.

1.4.4 Dreaming: virtual generation of sensory information while asleep

Even though disconnecting us from the outside world, sleep still hosts conscious sensory experiences, or dreams, triggered by the generation of an internal, virtual world. Strikingly, these experiences usually give us the feeling of being awake, as similar features to our external sensorium (characters, objects, colors, places, or sounds) are incorporated in a realistic manner. Moreover, similarly as waking experiences, dreams reflect our current concerns, interests and personality, and are highly rich in emotions (Nielsen and Stenstrom, 2005; Nir and Tononi, 2010). However, as soon as we wake up, most of these experiences are forgotten (dream amnesia), and if not, they are easily credited as virtual and distinguished from waking reality. This strange phenomenon has raised many questions about both their origin and their function.

1.4.4.1 Neuronal signatures of dreaming

A first enigma that remains unanswered is how and where dreams are generated in the brain. According to Hobson et al. (2000), dreams are generated through internal signals originating from PGO waves that excite the visual cortex and are later processed by higher-cortical areas. Other studies suggest that dreaming requires similar mechanisms as mental imagery, as lesions in temporo-parieto-occipital junction affect both dreaming and imagination (Kerr and Foulkes, 1981; Solms, 2000; Nir and Tononi, 2010). The hippocampus is also involved in dream generation and influence its episodic content (Spanò et al., 2020). Recently, Siclari et al. (2017) found a parieto-occipital hot zone as the neural correlate of dreaming, where a decrease in low-frequency EEG activity predicted that a participant was dreaming. However, due to the inability to decipher whether a person is currently dreaming or not, and only relying on post-hoc dream reports, it is still unclear what neuronal mechanisms underly the generation of dreams.

1.4.4.2 Dreams are bizarre and do not replay previous experiences

Despite their realism, dreams are often bizarre. Most contains impossible (talking to the deceased, cat being morphed into a car) or improbable (being hit by a tornado) features. This bizarreness is also characterized by uncertainties (not sure if I was talking to my brother or a friend), incongruities (I was looking through the window of my room in Paris, and I could see the Pacific ocean) and scene shifts (I was talking to a friend and I suddenly was playing a football game) (Williams et al., 1992; Mamelak and Hobson, 1989; Zadra and Stickgold, 2021).

These bizarre features are also related to the fact that dreams do not replay previous experiences (Fosse et al., 2003; Schwartz, 2003; Wamsley, 2014). In a study examining dream reports and waking activities from participants over 14 days, Fosse et al. (2003) showed that while 65% of dream reports incorporate aspects of waking life experiences, the exact replay of waking events was found in only 1-2 %. Similarly, playing extensively video games before sleep like Tetris (Stickgold et al., 2000) or ski-simulators (Wamsley et al., 2010a) led to dreams involving the learned task, but not actually playing the game. Instead, dreams are made of various isolated episodic fragments, sometimes non-obviously related (Llewellyn, 2016b; Lewis et al., 2018; Zadra and Stickgold, 2021) which partly explains their "bizarre" aspect.

Moreover, dream nature and bizarreness differ across the night. Indeed, unlike initially presumed (Aserinsky and Kleitman, 1953; Hobson and McCarley, 1977), dreams also occur during NREM sleep where they appear to be less bizarre, vivid, emotional and more episodic, reflecting recent waking experiences (Wamsley et al., 2007; Spanò et al., 2020; Zadra and Stickgold, 2021). Dream-like experiences also occur at sleep onset through hypnagogic hallucinations, often related to thoughts immediately before falling asleep (Waters et al., 2016). There seems to be a continuum of dream bizarreness across the night, which correlates with differential activity patterns observed during NREM and REM sleep (Section 1.4.2).

1.4.4.3 The potential functions of dreams

The bizarreness of dreams raise the question of their potential function - how such a novel, hallucinatory and fantastic experience would benefit our daily lives? Or maybe dreams are just a by-product of the sleeping brain?

Do dreams reveal unconscious wishes? Freud initially proposed that such experiences represent the fulfilment of unconscious wishes from our waking life. Dreaming, in his view, acts as a "day residue", where unacceptable wishes (sexual or agressive) can be expressed without being acted out consciously while awake (Freud, 1900). Even though highly popular in psychoanalysis, this view was later refuted by Hobson and McCarley (1977) who explained that dreams originate from neural signals in the brainstem during REM sleep. According to their activation-synthesis theory, the intermittent bursts of activity during PGO waves trigger a chaotic input that the brain tries to make sense of, arguing against the Freudian idea that dreams have any meaning to be interpreted. Following this turnover, a few theories proposed that dreams serve an actual cognitive function.

Dreams and memory replay. Some theories directly associate dreaming with memory replay. Francis Crick suggested that dreams, by replaying certain spurious memories, promote their selective forgetting through a reversal learning process (Crick and Mitchison, 1983a). In contrast, more contemporary theories propose that dreams, by re-enacting stored memories, favour their consolidation (Wamsley, 2014),

supported by the observation that dreaming of a learning task improved the performance on this task the next day (Wamsley et al., 2010b; Wamsley and Stickgold, 2019). Even though plausible, as discussed above, the creative and bizarre nature of dreams, failing to faithfully reproduce waking episodes, might make them sub-optimal at consolidating specific memories (Hoel, 2021). Dream function might thus additionally benefit from the establishment of new connections between stored memories.

Dreams and emotional processing. In this line, some neurocognitive models support the idea that dreams regulate our emotions, either through fear extinction by allowing fear or traumatic stimuli to be experienced in novel circumstances (Levin and Nielsen, 2009) or as a form of "nighttime therapy" that helps merging emotional concerns and traumatic events with existing memories (Hartmann, 2007). However, emotional regulation could be attributed to sleep itself and not particularly dreams, which are not always emotional (Gruber and Cassoff, 2014; Hoel, 2021).

Dreams and creativity. A more prominent theory of why dreams combine memories in a bizarre manner is that they enhance creativity. Starting with the anecdotal evidence of scientific discoveries from dreams, e.g., benzene structure by Kekule (1865) or the chemical neurotransmission by Loewi (1936) (Mazzarello, 2000), the role of dreams in creativity has then been taken in wider consideration. It has been proposed that the creative associations between unrelated memories during dreaming could lead to the discovery of unexpected solutions at the essence of creativity (Lewis et al., 2018). Through this process, the dreamer would make creative experimentations for potential future situations, as a form of prospective coding (Hobson, 2009; Llewellyn, 2016b), e.g., rehearsing threat perception and avoidance. However, studies report that dreams rarely contain practical solutions to real-life problems, in addition to the fact that most dreams are forgotten (Malcolm-Smith and Solms, 2004; Zadra et al., 2006; Zadra and Stickgold, 2021).

Instead, Zadra and Stickgold (2021) propose that the weak associations created during dreaming are useful to explore potential connections among diverse memories that brain would never normally consider while awake, without referring to a specific situation. Dreams can be seen as a period of "incubation" to discern non-obvious, associative pattern in events or knowledge (Cropley, 2006; Llewellyn, 2016a). This theory partly got experimental support showing where subjects tend to better associate weakly-related words or to solve anagrams when awaken from REM sleep (Stickgold et al., 1999; Walker et al., 2002; Cai et al., 2009).

Dreams for enhancing generalization. Finally, a few theories took a computational account of the role of dreams in enhancing generalization. Based on the predictive coding framework (Rao and Ballard, 1999) described earlier, Hobson et al. (2014) hypothesize that the purpose of dreaming is to optimize the brain's generative model in the absence of sensory input. While during wakefulness this model is trained to minimize sensory prediction errors, dreaming aims to reduce its complexity, or the degree of freedom required to make accurate predictions (Penny et al., 2004), by pruning redundant synapses (Tononi and Cirelli, 2014) and thus preventing overfitting and enhancing generalization. Similarly, Hoel (2021) proposes that dreams evolved to prevent the brain from overtting on its waking experiences by providing out-of-distribution data. However, these computational ideas lack a model with a mechanistic implementation, especially since it is unclear how virtual dreams can improve a model generalization abilities, nor how these dreams are generated in the first place.

Overall, over the past decades, diverse theories proposed potential roles of dreams in cognitive processing, but none of them proposed that they could contribute to constructing semantic representations, as transformation theories (Nadel and Moscovitch, 1997; Lewis and Durrant, 2011) based on NREM hippocampal replay suggest. While theories proposing that dreams enhance creativity (Lewis et al., 2018) or generalization (Hobson et al., 2014) get close to the idea of semantization, it is still unclear how dreams mechanistically serve these purposes. In our work (Chapter 3, Deperrois et al., 2022), we propose that generative models, and in particular GANs, could bring new computational insights on the role of dreams in learning.

Chapter 2

Hypothesis and aim

2.1 Motivation

Observing that sensory representations from high cortical areas contain semantic information, we presented the current feedforward deep learning models explaining how these representations could be acquired. However, these models rely on supervised learning that hardly characterize how animals learn.

In parallel, certain theories suggest that the brain learns without supervision by implementing generative model in feedback pathways that predict sensory inputs from latent representations. However, most of these models (PC, WS, VAE) rely on reconstructing sensory inputs, ignoring that the brain generates more than what it perceives, such as during sleep and dreaming. In contrast, the implicit GANs model turn out to be successful at both generating realistic samples and learning structured representations without directly reconstructing data, potentially revealing how and why the brain generates sensory experiences in absence of sensory inputs.

We then reviewed diverse cognitive theories of sleep suggesting the memory reactivations and dreams might benefit learning on different aspects, sometimes contradicting each other (memory consolidation, semantization, forgetting, emotional processing). We hypothesize that this contradiction partly resides in the lack of a unifying algorithmic description of sleep with a defined brain model. In this thesis, I aim to define this model by taking inspiration from both generative modeling and dreaming phenomenology, thereby suggesting that dream experiences participate in learning semantic representations without supervision using adversarial learning principles.

2.2 Framework

We aim to describe the sensory cortex as deep networks receiving and generating sensory inputs during wakefulness and sleep. In line with the introduced generative models, our cortical architecture implements a feedforward pathway (encoder) inferring high-level (IT) representations from the observed inputs, and a parallel feedback pathway (generator) that generates sensory inputs from latent representations. Following CLS, we also introduce a hippocampal structure that stores and retrieves memories encoded from waking experiences.

Our aim is to explain how this cortical architecture can discover semantic concepts from sensory experiences with minimal supervision. We thereby define a training

paradigm based on learning objectives defined for each physiological phase (wakefulness, NREM and REM sleep). During wakefulness, the system is trained to predict sensory inputs from high-level representations, which are simultaneously stored in the hippocampus. We then simulate NREM sleep with a replay of hippocampal memories leading to the generation of a dream reproducing the previous waking experience. We finally model REM dreams, more bizarre and creative, through a combination of multiple memories from which feedback pathways try to generate a realistic input through an adversarial learning objective (Goodfellow et al., 2014).

2.3   Goals

The combination of these state-specific objectives defines the perturbed and adversarial dreaming (PAD) model for cortical representation learning. We hypothesize that by replaying sensory inputs with sensory perturbations, NREM sleep should improve the robustness of cortical representations, while by inventing new, realistic sensory inputs, REM sleep improves the semantic content of these representations. In order to test this hypothesis, the PAD model is simulated over many wake-sleep cycles illustrating early development in the animal brain. We then highlight the benefits of sleep phases by evaluating latent representations in presence or absence of NREM or REM sleep. We also aim to highlight the meaning of our results in the light of existing sleep theories and discuss the experimental paradigms that could verify our predictions.

Our model not only forms hypotheses NREM and REM sleep functions, but about cortical structure, plasticity mechanisms and neuronal activity patterns. We aim to propose how these features can be implemented with biological substrates and how they could be tested experimentally.

Chapter 3

Learning cortical representations through perturbed and adversarial dreaming

This chapter contains the manuscript *Learning cortical representations through perturbed and adversarial dreaming* published in the eLife Journal.

Authors Nicolas Deperrois[1], Mihai A. Petrovici[1,2] , Walter Senn[1], Jakob Jordan[1]

Author contribution Jakob Jordan and myself wrote the manuscript with critical contributions by Walter Senn and Mihai Petrovici. The project was initially designed by Walter Senn and was closely supervised by Jakob Jordan. Python code writing, model simulations and literature research were were carried out by me. All illustrations were drawn by me and combined with simulation results into the resulting figures.

Outreach Once published, Roberto Inchingolo from the Human Brain Project (HBP) wrote a press release from our publication[3], which was subsequently relayed by several scientific medias. It was also subject to radio and journal interviews reported in Appendix A.2.

Code All code necessary to repeat the experiments is published in the following repository.

---

[1]University of Bern, Switzerland

[2]Kirchhoff-Institute for Physics, Heidelberg University, Germany.

[3]https://www.humanbrainproject.eu/en/follow-hbp/news/2022/05/12/
strange-dreams-might-help-your-brain-learn-better-according-research-hbp-scientists/

Learning cortical representation through perturbed and adversarial dreaming

Nicolas Deperrois[1], Mihai A. Petrovici[1,2], Walter Senn[1], Jakob Jordan[1]

[1] Department of Physiology, University of Bern, 3012 Bern, Switzerland.
[2] Kirchhoff-Institute for Physics, Heidelberg University, 69120 Heidelberg, Germany.

3.1 Abstract

Humans and other animals learn to extract general concepts from sensory experience without extensive teaching. This ability is thought to be facilitated by offline states like sleep where previous experiences are systemically replayed. However, the characteristic creative nature of dreams suggests that learning semantic representations may go beyond merely replaying previous experiences. We support this hypothesis by implementing a cortical architecture inspired by generative adversarial networks (GANs). Learning in our model is organized across three different global brain states mimicking wakefulness, NREM and REM sleep, optimizing different, but complementary objective functions. We train the model on standard datasets of natural images and evaluate the quality of the learned representations. Our results suggest that generating new, virtual sensory inputs via adversarial dreaming during REM sleep is essential for extracting semantic concepts, while replaying episodic memories via perturbed dreaming during NREM sleep improves the robustness of latent representations. The model provides a new computational perspective on sleep states, memory replay and dreams and suggests a cortical implementation of GANs.

3.2 Introduction

After just a single night of bad sleep, we are acutely aware of the importance of sleep for orderly body and brain function. In fact, it has become clear that sleep serves multiple crucial physiological functions (Siegel, 2009; Xie et al., 2013), and growing evidence highlights its impact on cognitive processes (Walker, 2009). Yet, a lot remains unknown about the precise contribution of sleep, and in particular dreams, on normal brain function.

One remarkable cognitive ability of humans and other animals lies in the extraction of general concepts and statistical regularities from sensory experience without extensive teaching (Bergelson and Swingley, 2012). Such regularities in the sensorium are reflected on the neuronal level in invariant object-specific representations in high-level areas of the visual cortex (Grill-Spector et al., 2001; Hung et al., 2005; DiCarlo et al., 2012) on which downstreams areas can operate. These so called semantic representations are progressively constructed and enriched over an organism's lifetime (Tenenbaum et al., 2011; Yee et al., 2013) and their emergence is hypothesized to be facilitated by offline states such as sleep (Dudai et al., 2015).

Previously, several cortical models have been proposed to explain how offline states could contribute to the emergence of high-level, semantic representations. Stochastic hierarchical models which learn to maximize the likelihood of observed data under a generative model such as the Helmholtz machine (Dayan et al., 1995) and the closely related Wake-Sleep algorithm (Hinton et al., 1995; Bornschein and Bengio, 2015) have demonstrated the potential of combining online and offline states to learn semantic

representations. However, these models do not leverage offline states to improve their generative model but are explicitly trained to reproduce sensory inputs during wakefulness. In contrast, most dreams during REM sleep exhibit realistic imagery beyond past sensory experience (Fosse et al., 2003; Nir and Tononi, 2010; Wamsley, 2014) suggesting learning principles which go beyond mere reconstructions.

In parallel, cognitive models inspired by psychological studies of sleep proposed a "trace transformation theory" where semantic knowledge is actively extracted in the cortex from replayed hippocampal episodic memories (Nadel and Moscovitch, 1997; Winocur et al., 2010; Lewis and Durrant, 2011). However, these models lack a mechanistic implementation compatible with cortical structures and only consider the replay of waking activity during sleep.

Recently, implicit generative models which do not explicitly try to reconstruct observed sensory inputs, and in particular generative adversarial networks (GANs; Goodfellow et al., 2014), have been successfully applied in machine learning to generate new but realistic data from random patterns. This ability has been shown to be accompanied by the learning of disentangled and semantically meaningful representations (Radford et al., 2015; Donahue et al., 2016; Liu et al., 2021). They thus may provide computational principles for learning cortical semantic representations during offline states by generating previously unobserved sensory content as reported from dream experiences.

Most dreams experienced during rapid-eye-movement (REM) sleep only incorporate fragments of previous waking experience, often intermingled with past memories (Schwartz, 2003). Suprisingly, such random combinations of memory fragments often results in visual experiences which are perceived as highly structured and realistic by the dreamer. The striking similarity between the inner world of dreams and the external world of wakefulness suggests that the brain actively creates novel experiences by rearranging stored episodic patterns in a meaningful manner (Nir and Tononi, 2010). A few hypothetical functions were attributed to this phenomenon, such as enhancing creative problem solving by building novel associations between unrelated memory elements (Cai et al., 2009; Llewellyn, 2016a; Lewis et al., 2018), forming internal prospective codes oriented toward future waking experiences (Llewellyn, 2016b), or refining a generative model by minimizing its complexity and improving generalization (Hobson et al., 2014; Hoel, 2021). However, these theories do not consider the role of dreams for a more basic function, such as the formation of semantic cortical representations.

Here, we propose that dreams, and in particular their creative combination of episodic memories, play an essential role in forming semantic representations over the course of development. The formation of representations which abstract away redundant information from sensory input and which can thus be easily used by downstream areas is an important basis for memory semantization. To support this hypothesis, we introduce a new, functional model of cortical representation learning. The central ingredient of our model is a creative generative process via feedback from higher to lower cortical areas which mimics dreaming during REM sleep. This generative process is trained to produce more realistic virtual sensory experience in an adversarial fashion by trying to fool an internal mechanism distinguishing low-level activities between wakefulness and REM sleep. Intuitively, generating new but realistic sensory experiences, instead of merely reconstructing previous observations, requires the brain to understand the composition of its sensorium. In line with transformation theories,

this suggests that cortical representations should carry semantic, decontextualized gist information.

We implement this model in a cortical architecture with hierarchically organized forward and backward pathways, loosely inspired by GANs. The connectivity of the model is adapted by gradient-based synaptic plasticity, optimizing different, but complementary objective functions depending on the brain's global state. During wakefulness, the model learns to recognize that low-level activity is externally-driven, stores high-level representations in the hippocampus, and tries to predict low-level from high-level activity (Fig. 3.1a). During NREM sleep, the model learns to reconstruct replayed high-level activity patterns from generated low-level activity, perturbed by virtual occlusions, referred to as perturbed dreaming (Fig. 3.1b). During REM sleep, the model learns to generate realistic low-level activity patterns from random combinations of several hippocampal memories and spontaneous cortical activity, while simultaneously learning to distinguish these virtual experiences from externally-driven waking experiences, referred to as adversarial dreaming (Fig. 3.1c). Together with the wakefulness, the two sleep states, NREM and REM, jointly implement our model of Perturbed and Adversarial Dreaming (PAD).

Over the course of learning, constrained by its architecture and the prior distribution of latent activities, our cortical model trained on natural images develops rich latent representations along with the capacity to generate plausible early sensory activities. We demonstrate that adversarial dreaming during REM sleep is essential for learning representations organized according to object semantics, which are improved and robustified by perturbed dreaming during NREM sleep. Together, our results demonstrate a potential role of dreams and suggest complementary functions of REM and NREM sleep in cortical representation learning.

3.3   Results

3.3.1   Complementary objectives for wakefulness, NREM and REM sleep

We consider an abstract model of the visual ventral pathway consisting of multiple, hierarchically organized cortical areas, with a feedforward pathway, or encoder, transforming neuronal activities from lower to higher areas (Fig. 3.2, E). These high-level activities are compressed representations of low-level activities and are called latent representations, here denoted by z. In addition to this feedforward pathway, we similarly model a feedback pathway, or generator, projecting from higher to lower areas (Fig. 3.2, G). These two pathways are supported by a simple hippocampal module which can store and replay latent representations. Three different global brain states are considered: wakefulness (Wake), non-REM sleep (NREM) and REM sleep (REM). We focus on the functional role of these phases while abstracting away dynamical features such as bursts, spindles or slow waves (Léger et al., 2018), in line with previous approaches based on goal-driven modeling which successfully predict physiological features along the ventral stream (Yamins et al., 2014; Zhuang et al., 2021).

In our model, the three brain states only differ in their objective function and the presence or absence of external input. Synaptic plasticity performs stochastic gradient descent on state-specific objective functions via error backpropagation (LeCun
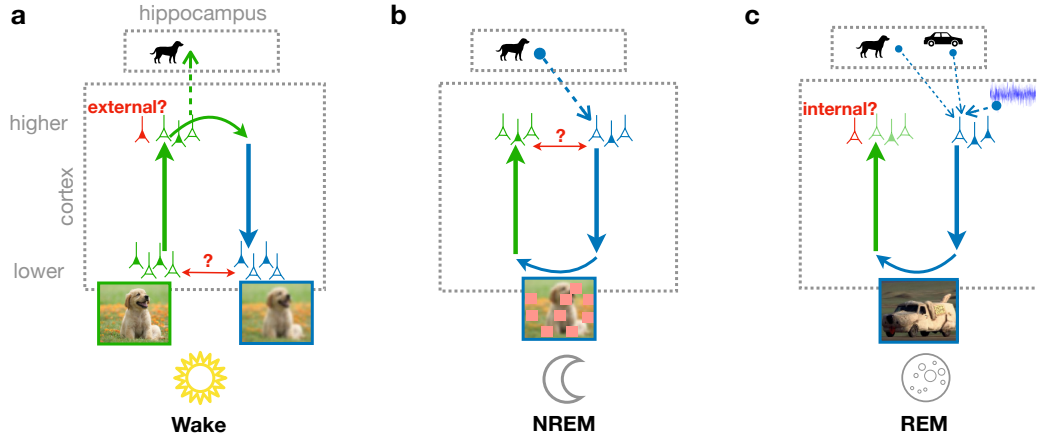
FIGURE 3.1: **Cortical representation learning through perturbed and adversarial dreaming (PAD). (a)** During wakefulness (Wake), cortical feedforward pathways learn to recognize that low-level activity is externally-driven and feedback pathways learn to reconstruct it from high-level neuronal representations. These high-level representations are stored in the hippocampus. **(b)** During NREM sleep (NREM), feedforward pathways learn to reconstruct high-level activity patterns replayed from the hippocampus affected by low-level perturbations, referred to as perturbed dreaming. **(c)** During REM sleep (REM), feedforward and feedback pathways operate in an adversarial fashion, referred to as adversarial dreaming. Feedback pathways generate virtual low-level activity from combinations of multiple hippocampal memories and spontaneous cortical activity. While feedforward pathways learn to recognize low-level activity patterns as internally generated, feedback pathways learn to fool feedforward pathways.

et al., 2015). We assume that efficient credit assignment is realized in the cortex, and focus on the functional consequences of our specific architecture. For potential implementations of biophysically plausible backpropagation in cortical circuits, we refer to previous work (e.g., Whittington and Bogacz, 2019; Lillicrap et al., 2020).

During Wake (Fig. 3.2a), sensory inputs evoke activities x in lower sensory cortex which are transformed via the feedforward pathway E into latent representations z in higher sensory cortex. The hippocampal module stores these latent representations, mimicking the formation of episodic memories. Simultaneously, the feedback pathway G generates low-level activities x' from these representations. Synaptic plasticity adapts the encoding and generative pathways (E and G) to minimize the mismatch between externally-driven and internally-generated activities (Fig. 3.2a). Thus, the network learns to reproduce low-level activity from abstract high-level representations. Simultaneously, E also acts as a 'discriminator' with output d that is trained to become active, reflecting that the low-level activity was driven by an external stimuli. The discriminator learning during Wake is essential to drive adversarial learning during REM. Note that computationally the classification of low-level cortical activities into "externally driven" and "internally generated" is not different from classification into, for example, different object categories, even though conceptually they serve different purposes. The dual use of E reflects a view of cortical information processing in which several network functions are preferentially shared among a single network mimicking the ventral visual stream (DiCarlo et al., 2012). This approach has been previously successfully employed in machine learning models (Huang et al., 2018; Brock et al., 2017; Ulyanov et al., 2017; Munjal et al., 2019; Bang et al., 2020).

For the subsequent sleep phases, the system is disconnected from the external environment, and activity in lower sensory cortex is driven by top-down signals originating
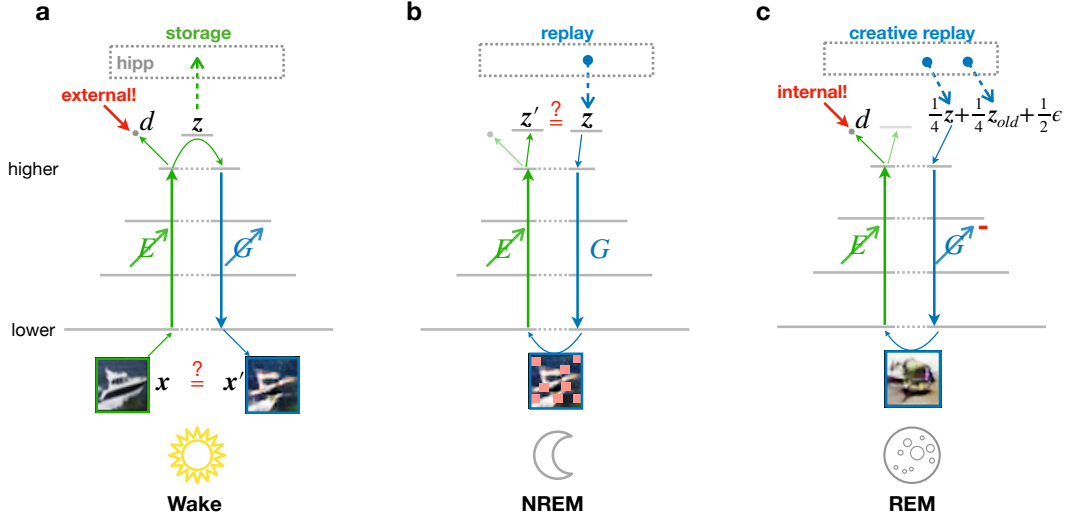
FIGURE 3.2: **Different objectives during wakefulness, NREM, and REM sleep govern the organization of feedforward and feedback pathways in PAD** The variable x corresponds to 32x32 image, z is a 256-dimensional vector representing the latent layer (higher sensory cortex). Encoder (E, green) and generator (G, blue) networks project bottom-up and top-down signals between lower and higher sensory areas. An oblique arrow ($\nearrow$) indicates that learning occurs in a given pathway. **(a)** During Wake, low-level activities x are reconstructed. At the same time, E learns to classify low-level activity as external (red target 'external!') with its output discriminator d. The obtained latent representations z are stored in the hippocampus. **(b)** During NREM, the activity z stored during wakefulness is replayed from the hippocampal memory and regenerates visual input from the previous day perturbed by occlusions, modelled by squares of various sizes applied along the generated low-level activity with a certain probability (see Methods). In this phase, E adapts to reproduce the replayed latent activity. **(c)** During REM, convex combinations of multiple random hippocampal memories (z and z_{old}) and spontaneous cortical activity (ϵ), here with specific prefactors, generate a virtual activity in lower areas. While the encoder learns to classify this activity as internal (red target 'internal!'), the generator adversarially learns to generate visual inputs that would be classified as external. The red minus on G indicates the inverted plasticity implementing this adversarial training.

from higher areas, as previously suggested (Nir and Tononi, 2010; Aru et al., 2020). During NREM (Fig. 3.2b), latent representations z are recalled from the hippocampal module, corresponding to the replay of episodic memories. These representations generate low-level activities which are perturbed by suppressing early sensory neurons, modeling the observed differences between replayed and waking activities (Ji and Wilson, 2007). The encoder reconstructs latent representations from these activity patterns, and synaptic plasticity adjusts the feedforward pathway to make the latent representation of the perturbed generated activity similar to the original episodic memory. This process defines perturbed dreaming.

During REM (Fig. 3.2c), sleep is characterized by creative dreams generating realistic virtual sensory experiences out of the combination of episodic memories (Fosse et al., 2003; Lewis et al., 2018). In PAD, multiple random episodic memories from the hippocampal module are linearly combined and projected to cortex. Reflecting the decreased coupling (Wierzynski et al., 2009; Lewis et al., 2018) between hippocampus and cortex during REM sleep, these mixed representations are diluted with spontaneous cortical activity, here abstracted as Gaussian noise with zero mean and unit variance. From this new high-level cortical representation, activity in lower sensory

cortex is generated and finally passed through the feedforward pathway. Synaptic plasticity adjusts feedforward connections E to silence the activity of the discriminator output as it should learn to distinguish it from externally-evoked sensory activity. Simultaneously, feedback connections are adjusted adversarially to generate activity patterns which appear externally-driven and thereby trick the discriminator into believing that the low-level activity was externally-driven. This is achieved by inverting the sign of the errors that determine synaptic weight changes in the generative network. This process defines adversarial dreaming.

The functional differences between our proposed NREM and REM sleep phases are motivated by experimental data describing a reactivation of hippocampal memories during NREM sleep and the occurrence of creative dreams during REM sleep. In particular, hippocampal replay has been reported during NREM sleep within sharp-wave-ripples (O'Neill et al., 2010), also observed in the visual cortex (Ji and Wilson, 2007), which resembles activity from wakefulness. Our REM sleep phase is built upon cognitive theories of REM dreams (Llewellyn, 2016b; Lewis et al., 2018) postulating that they emerge from random combinations between episodic memory elements, sometimes remote from each other, which appear realistic for the dreamer. This random coactivation could be caused by theta oscillations in the hippocampus during REM sleep (Buzsáki, 2002). The addition of cortical noise is motivated by experimental work showing reduced correlations between hippocampal and cortical activity during REM sleep (Wierzynski et al., 2009), and the occurence of ponto-geniculo-occipital (PGO) waves (Nelson et al., 1983) in the visual cortex often associated with generation of novel visual imagery in dreams (Hobson et al., 2000, 2014). Furthermore, the cortical contribution in REM dreaming is supported by experimental evidence that dreaming still occurs with hippocampal damage, while reported to be less episodic-like in nature (Spanò et al., 2020).

Within our suggested framework, 'dreams' arise as early sensory activity that is internally-generated via feedback pathways during offline states, and subsequently processed by feedforward pathways. In particular, this implies that besides REM dreams, NREM dreams exist. However, in contrast to REM dreams, which are significantly different from waking experiences (Fosse et al., 2003), our model implies that NREM dreams are more similar to waking experiences since they are driven by single episodic memories, in contrast to REM dreams which are generated from a mixture of episodic memories. Furthermore, the implementation of adversarial dreaming requires an internal representation of whether early sensory activity is externally or internally generated, i.e., a distinction whether a sensory experience is real or imagined.

3.3.2 Dreams become more realistic over the course of learning

Dreams in our model arise from both NREM (perturbed dreaming) and REM (adversarial dreaming) phases. In both cases, they are characterized by activity in early sensory areas generated via feedback pathways. To illustrate learning in PAD, we consider these low-level activities during NREM and during REM for a model with little learning experience ("early training") and a model which has experienced many wake-sleep cycles ("late training"; Fig. 3.3). A single wake-sleep cycle consists of Wake, NREM and REM phases. As an example, we train our model on a dataset of natural images (CIFAR-10; Krizhevsky et al., 2013) and a dataset of images of house numbers (SVHN; Netzer et al., 2011). Initially, internally-generated low-level activities during sleep do not share significant similarities with sensory-evoked activities from Wake (Fig. 3.3a); for example, no obvious object shapes are represented (Fig. 3.3b).
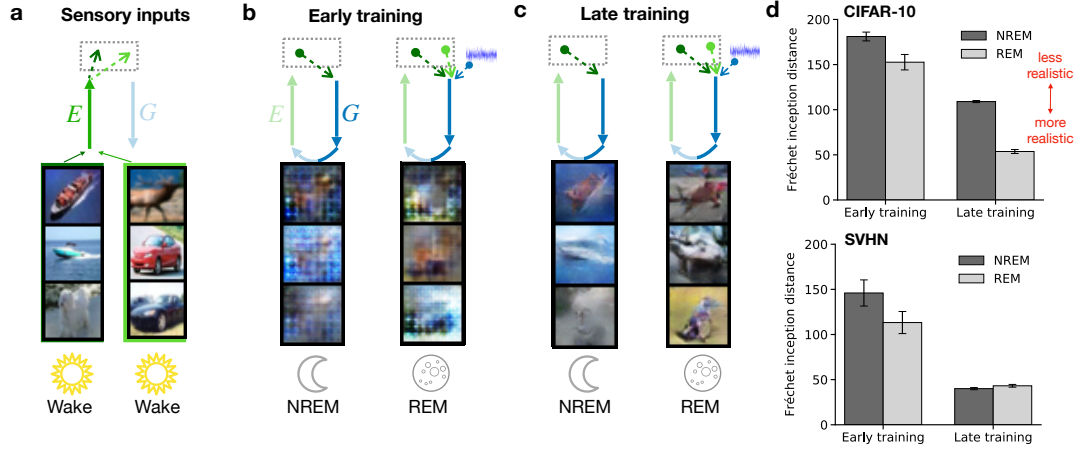
FIGURE 3.3: **Both NREM and REM dreams become more realistic over the course of learning. (a)** Examples of sensory inputs observed during wakefulness. Their corresponding latent representations are stored in the hippocampus. **(b, c)** Single episodic memories (latent representations of stimuli) during NREM from the previous day and combinations of episodic memories from the two previous days during REM are recalled from hippocampus and generate early sensory activity via feedback pathways. This activity is shown for early (epoch 1) and late (epoch 50) training stages of the model. **(d)** Discrepancy between externally-driven and internally-generated early sensory activity as measured by the Fréchet inception distance (FID) (Heusel et al., 2018) during NREM and REM for networks trained on CIFAR-10 (top) and SVHN (bottom). Lower distance reflects higher similarity between sensory-evoked and generated activity. Error bars indicate ±1 SEM over 4 different initial conditions.

After plasticity has organized network connectivity over many wake-sleep cycles (50 training epochs), low-level internally-generated activity patterns resemble sensory-evoked activity (Fig. 3.3c). NREM-generated activities reflect the sensory content of the episodic memory (sensory input from the previous day). REM-generated activities are different from the sensory activities corresponding to the original episodic memories underlying them as they recombine features of sensory activities from the two previous days, but still exhibit a realistic structure. This increase in similarity between externally-driven and internally-generated low-level activity patterns is also reflected in a decreasing Fréchet inception distance (FID, Fig. 3.3d), a metric used to quantify the realism of generated images (Heusel et al., 2018). The increase of dreams realism, here mostly driven by a combination of reconstruction learning (Wake) and adversarial learning (Wake and REM), correlates with the development of dreams in children, that are initially plain and fail to represent objects, people, but become more realistic and structured over time (Foulkes, 1999; Nir and Tononi, 2010).

The PAD training paradigm hence leads to internally-generated low-level activity patterns that become more difficult to discern from externally-driven activities, whether they originate from single episodic memories during NREM or from noisy random combinations thereof during REM. We will next demonstrate that the same learning process leads to the emergence of robust semantic representations.
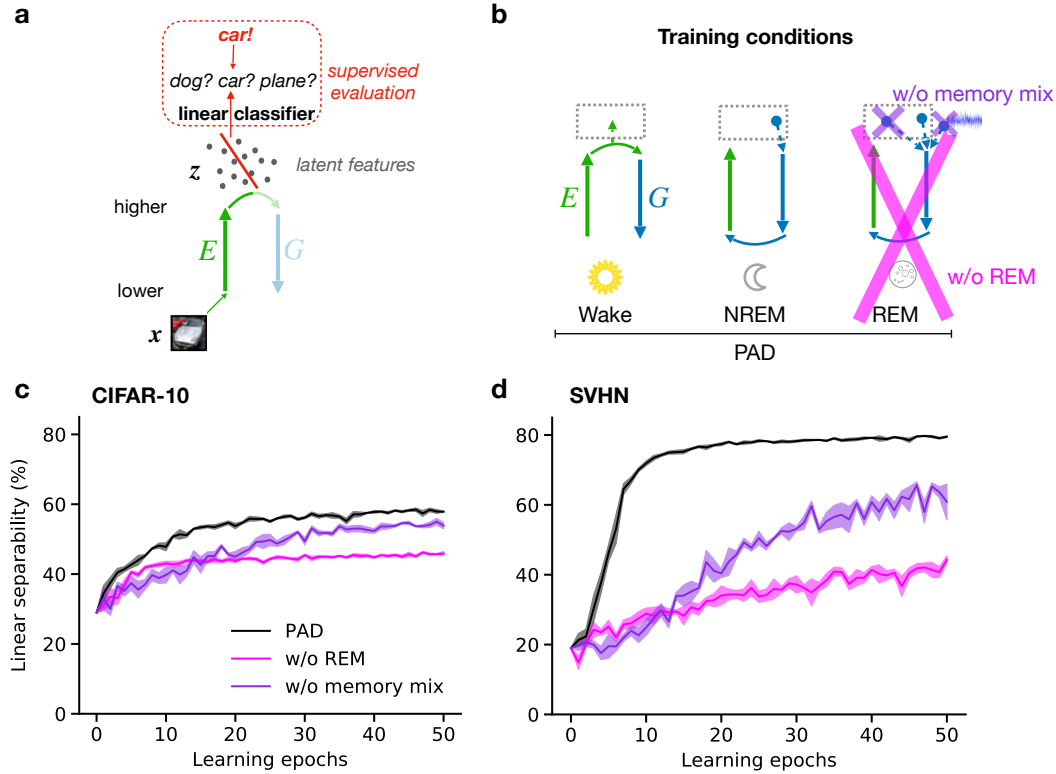
FIGURE 3.4: **Adversarial dreaming during REM improves the linear separability of the latent representation.** **(a)** A linear classifier is trained on the latent representations z inferred from an external input x to predict its associated label (here, the category 'car'). **(b)** Training phases and pathological conditions: full model (PAD, black), no REM phase (pink) and PAD with a REM phase using a single episodic memory only ('w/o memory mix', purple). **(c, d)** Classification accuracy obtained on test datasets (c: CIFAR-10; d: SVHN) after training the linear classifier to convergence on the latent space z for each epoch of the E-G-network learning. Full model (PAD): black line; without REM: pink line; with REM, but without memory mix: purple line. Solid lines represent mean and shaded areas indicate ± 1 SEM over 4 different initial conditions.

3.3.3 Adversarial dreaming during REM facilitates the emergence of semantic representations

Semantic knowledge is fundamental for animals to learn quickly, adapt to new environments and communicate, and is hypothesized to be held by so-called semantic representations in cortex (DiCarlo et al., 2012). An example of such semantic representations are neurons from higher visual areas that contain linearly separable information about object category, invariant to other factors of variation, such as background, orientation or pose (Grill-Spector et al., 2001; Hung et al., 2005; Majaj et al., 2015).

Here we demonstrate that PAD, due to the specific combination of plasticity mechanisms during Wake, NREM and REM, develops such semantic representations in higher visual areas. Similarly as in the previous section, we train our model on the CIFAR-10 and SVHN datasets. To quantify the quality of inferred latent representations, we measure how easily downstream neurons can read out object identity from these. For a simple linear read-out, its classification accuracy reflects the linear separability of different contents represented in a given dataset. Technically, we train a

linear classifier that distinguishes object categories based on their latent representations z after different numbers of wake-sleep cycles ('epochs', Fig. 3.4a) and report its accuracy on data not used during training of the model and classifier ("test data"). While training the classifier, the connectivity of the network (E and G) is fixed.

The latent representation (z) emerging from the trained network (Fig. 3.4b, full model) shows increasing linear separability reaching around 59% test accuracy on CIFAR-10 (Fig. 3.4c, black line, for details see Table 3.1) and 79% on SVHN (Fig. 3.4d, black line), comparable to less biologically plausible machine-learning models (Berthelot et al., 2018). These results show the ability of PAD to discover semantic concepts across wake-sleep cycles in an unsupervised fashion.

Within our computational framework, we can easily consider sleep pathologies by directly interfering with the sleep phases. To highlight the importance of REM in learning semantic representations, we consider a reduced model in which the REM phase with adversarial dreaming is suppressed and only perturbed dreaming during NREM remains (Fig. 3.4b, pink cross). Without REM sleep, linear separability increases much slower and even after a large number of epochs remains significantly below the PAD (see also Fig. 3.12c,d). This suggests that adversarial dreaming during REM, here modeled by an adversarial game between feedforward and feedback pathways, is essential for the emergence of easily readable, semantic representations in the cortex. From a computational point of view, this result is in line with previous work showing that learning to generate virtual inputs via adversarial learning (GANs variants) forms better representations than simply learning to reproduce external inputs (Radford et al., 2015; Donahue et al., 2016; Berthelot et al., 2018).

Finally, we consider a different pathology in which REM is not driven by randomly combined episodic memories and noise, but by single episodic memories without noise, as during NREM (Fig. 3.4b, purple cross). Similarly to removing REM, linear separability increases much slower across epochs, leading to worse performance of the readout (Fig. 3.4c,d, purple lines). For the SVHN dataset, the performance does not reach the level of the PAD even after many wake-sleep cycles (see also Fig. 3.12d). This suggests that combining different, possibly non-related episodic memories, together with spontaneous cortical activity, as reported during REM dreaming (Fosse et al., 2003), leads to significantly faster representation learning.

Our results suggest that generating virtual sensory inputs during REM dreaming, via a high-level combination of hippocampal memories and spontaneous cortical activity and subsequent adversarial learning, allow animals to extract semantic concepts from their sensorium. Our model provides hypotheses about the effects of REM deprivation, complementing pharmacological and optogenetic studies reporting impairments in the learning of complex rules and spatial object recognition (Boyce et al., 2016). For example, our model predicts that object identity would be less easily decodable from recordings of neuronal activity in the Inferior-Temporal (IT) cortex in animal models with chronically impaired REM sleep.

3.3.4 Perturbed dreaming during NREM improves robustness of semantic representations.

Generalizing beyond previously experienced stimuli is essential for an animal's survival. This generalization is required due to natural perturbations of sensory inputs,
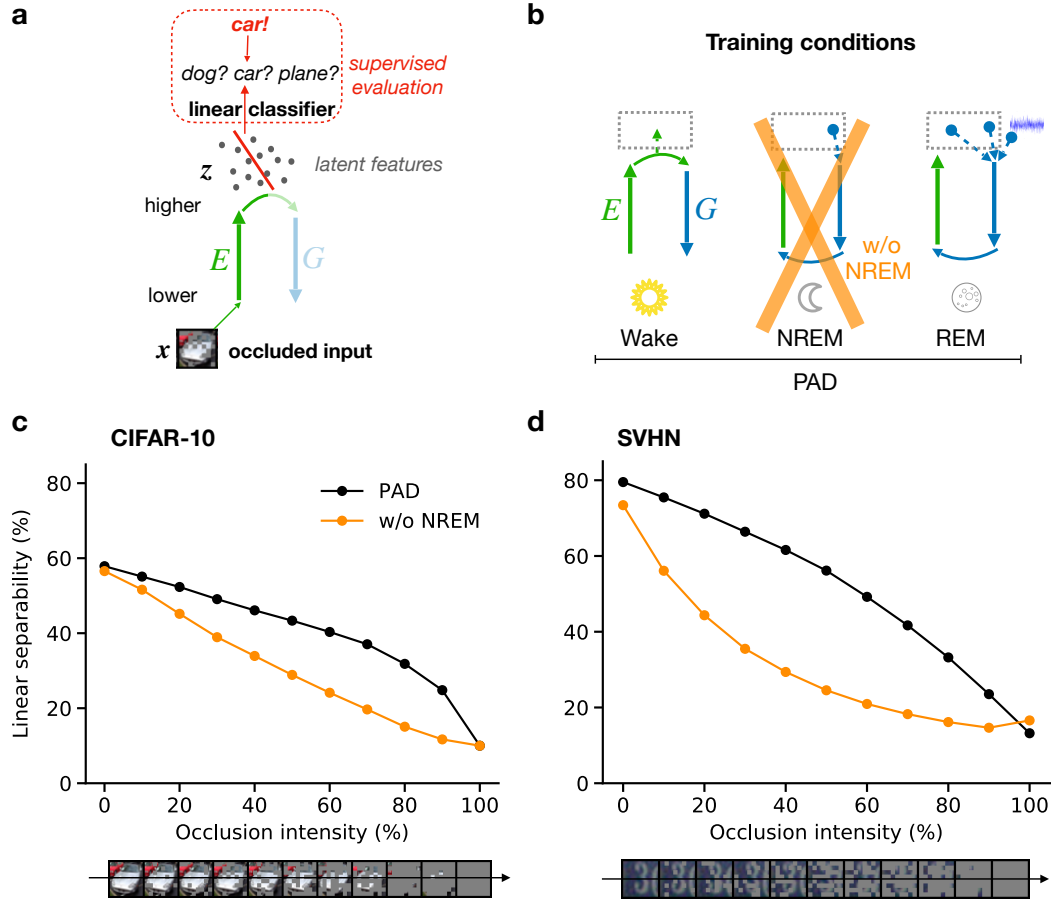
FIGURE 3.5: **Perturbed dreaming during NREM improves robustness of latent representations.** **(a)** A trained linear classifier (cf. Fig. 3.4) infers class labels from latent representations. The classifier was trained on latent representations of original images, but evaluated on representations of images with varying levels of occlusion. **(b)** Training phases and pathological conditions: full model (PAD, black), without NREM phase (orange). **(c, d)** Classification accuracy obtained on test dataset (C: CIFAR-10; D: SVHN) after 50 epochs for different levels of occlusion (0 to 100%). Full model (PAD): black line; w/o NREM: orange line. SEM over 4 different initial conditions overlap with data points. Note that due to an unbalanced distribution of samples the highest performance of a naive classifier is 18.9% for the SVHN dataset.

for example partial occlusions, noise, or varying viewing angles. These alter the stimulation pattern, but in general should not change its latent representation subsequently used to make decisions.

Here, we model such sensory perturbations by silencing patches of neurons in early sensory areas during the stimulus presentation (Fig. 3.5a). As before, linear separability is measured via a linear classifier that has been trained on latent representations of un-occluded images and we use stimuli which were not used during training. Adding occlusions hence directly tests the out-of-distribution generalization capabilities of the learned representations. For the model trained with all phases (Fig. 3.5b, full model), the linear separability of latent representations decreases as occlusion intensity increases, until reaching chance level for fully occluded images (Fig. 3.5c,d; black line).

We next consider a sleep pathology in which we suppress perturbed dreaming during the NREM phase while keeping adversarial dreaming during REM (Fig. 3.5b, orange cross). Without NREM, linear separability of partially occluded images is significantly decreased for identical occlusion levels (Fig. 3.5c,d; compare black and orange lines). In particular, performance degrades much faster with increasing occlusion levels. Note that despite the additional training objective, the full PAD develops equally good or even better latent representations of unoccluded images (0% occlusion intensity) compared to this pathological condition without perturbed dreams.

Crucially, the perturbed dreams in NREM are generated by replaying single episodic memories. If the latent activity fed to the generator during NREM was of similar origin as during REM, i.e. obtained from a convex combination of multiple episodic memories coupled with cortical spontaneous activity, the quality of the latent representations significantly decreases (see also Fig. 3.15). This suggests that only replaying single memories, as hypothesized to occur during NREM sleep (O'Neill et al., 2010), rather than their noisy combination, is beneficial to robustify latent representations against input perturbations.

This robustification originates from the training objective defined in the NREM phase, forcing feedforward pathways to map perturbed inputs to the latent representation corresponding to their clean, non-occluded version. This procedure is reminiscent of a regularization technique from machine learning called 'data augmentation' (Shorten and Khoshgoftaar, 2019), which increases the amount of training data by adding stochastic perturbations to each input sample. However, in contrast to data augmentation methods which directly operate on samples, here the system autonomously generates augmented data in offline states, preventing interference with online cognition and avoiding storage of the original samples. Our 'dream augmentation' suggests that NREM hippocampal replay not only maintains or strengthens cortical memories, as traditionally suggested (Klinzing et al., 2019), but also improves latent representations when only partial information is available. For example, our model predicts that animals lacking such dream augmentation, potentially due to impaired NREM sleep, fail to react reliably to partially occluded stimuli even though their responses to clean stimuli are accurate.

3.3.5 Latent organization in healthy and pathological models

The results so far demonstrate that perturbed and adversarial dreaming (PAD), during REM and NREM sleep states, contribute to cortical representation learning by increasing the linear separability of latent representations into object classes. We next investigate how the learned latent space is organized, i.e., whether representations of sensory inputs with similar semantic content are grouped together even if their low-level structure may be quite different, for example due to different viewing angles, variations among an object category, or (partial) occlusions.

We illustrate the latent organization by projecting the latent variable z using Principal Component Analysis (PCA, Fig. 3.6a, Jolliffe and Cadima, 2016). This method is well-suited for visualizing high-dimensional data in a low-dimensional space while preserving as much of the data's variation as possible.

For PAD, the obtained PCA projection shows relatively distinct clusters of latent representations according to the semantic category ("class identity") of their corresponding images (Fig. 3.6b). The model thus tends to organize latent representations
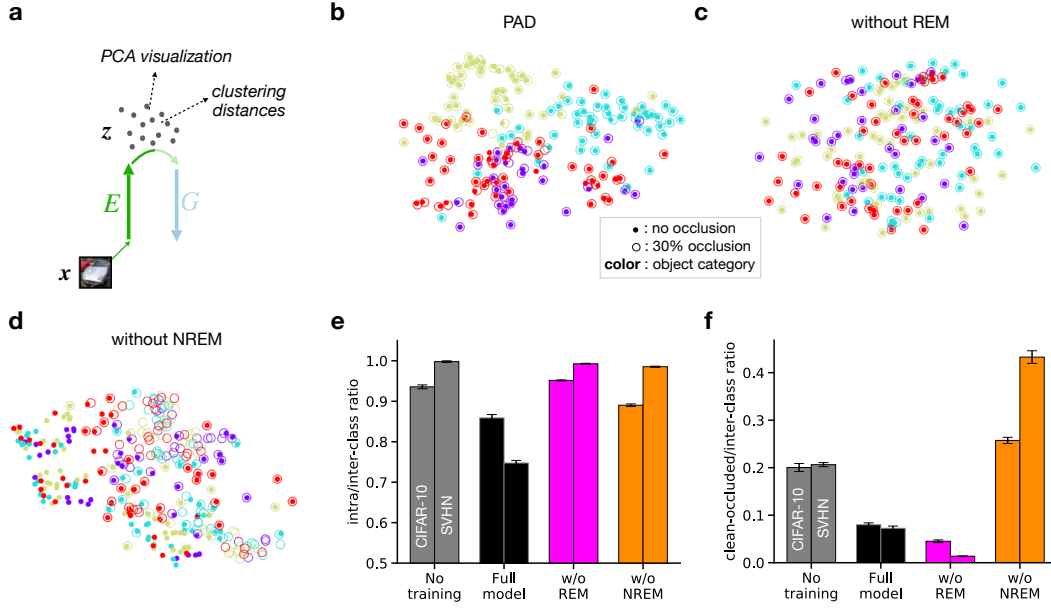
FIGURE 3.6: **Effects of NREM and REM sleep on latent representations.** **(a)** Inputs x are mapped to their corresponding latent representations z via the encoder E. Principal Component Analysis (Jolliffe and Cadima, 2016) is performed on the latent space to visualize its structure (b-d). Clustering distances (e,f) are computed directly on latent features z. **(b, c, d)** PCA visualization of latent representations projected on the first two principal components. Full circles represent clean images, open circles represent images with 30% occlusion. Each color represents an object category from the SVHN dataset (purple:'0', cyan:'1', yellow:'2', red:'3'). **(e)** Ratio between average intra-class and average inter-class distances in latent space for randomly initialized networks (no training, grey), full model (black), model trained without REM sleep (w/o REM, pink) and model trained without NREM sleep (w/o NREM, orange) for un-occluded inputs. **(f)** Ratio between average clean-occluded (30% occlusion) and average inter-class distances in latent space for full model (black), w/o REM (pink) and w/o NREM (orange). Error bars represent SEM over 4 different initial conditions.

such that high-level, semantic clusters are discernable. Furthermore, partially occluded objects (Fig. 3.6b, empty circles) are represented closeby their corresponding un-occluded version (Fig. 3.6b, full circles).

As shown in the previous sections, removing either REM or NREM has a negative impact on the linear separability of sensory inputs. However, the reasons for these effects are different between REM and NREM. If REM sleep is removed from training, representations of unoccluded images are less organized according their semantic category, but still match their corresponding occluded versions (Fig. 3.6c). REM is thus necessary to organize latent representations into semantic clusters, providing an easily readable representation for downstream neurons. In contrast, removing NREM causes representations of occluded inputs to be remote from their un-occluded representations (Fig. 3.6d).

We quantify these observations by computing the average distances between latent representations from the same object category (intra-class distance) and between representations of different object category (inter-class distance). Since the absolute distances are difficult to interpret, we focus on their ratio (Fig. 3.6e). On both datasets, this ratio increases if the REM phase is removed from training (Fig. 3.6e,

compare black and pink bars), reaching levels comparable to the one with the untrained network. Moreover, removing NREM from training also increases this ratio. These observations suggest that both perturbed and adversarial dreaming jointly reorganize the latent space such that stimuli with similar semantic structure are mapped to similar latent representations. In addition, we compute the distance between the latent representations inferred from clean images and their corresponding occluded versions, also divided by the inter-class distance (Fig. 3.6f). By removing NREM from training, this ratio increases significantly, highlighting the importance of NREM in making latent representations invariant to input perturbations.

3.3.6   Cortical implementation of PAD

We have shown that perturbed and adversarial dreaming (PAD) can learn semantic cortical representations useful for downstream tasks. Here we hypothesize how the associated mechanisms may be implemented in cortex.

First, PAD implies the existence of discriminator neurons that would learn to be differentially active during wakefulness and REM sleep. It also postulates a conductor that orchestrates learning by providing a teaching ('nudging') signal to the discriminator neurons during Wake and REM. Experimental evidence suggests that discriminator neurons, differentiating between internally generated end externally driven sensory activity, may reside in the anterior cingulate cortex (ACC) or the medial prefrontal cortex (mPFC), but functionally similar neurons may be located across cortex to deliver local learning signals (Subramaniam et al., 2012; Simons et al., 2017; Gershman, 2019; Benjamin and Kording, 2021b).

Second, learning in PAD is orchestrated across three different phases: (i) learning stimulus reconstruction during Wake, (ii) learning latent variable reconstruction during NREM sleep (perturbed dreaming), and (iii) learning to generate realistic sensory activity during REM sleep (adversarial dreaming). Our model suggests that objective functions and synaptic plasticity are affected by these phases (Fig. 3.7). Wakefulness is associated with increased activity of modulatory brainstem neurons releasing neuromodulators such as acetylcholine (ACh) and noradrenaline (NA), hypothesized to prioritize the amplification of information from external stimuli (Adamantidis et al., 2019; Aru et al., 2020). In contrast, neuromodulator concentrations during NREM are reduced compared to Wake, while REM is characterized by high ACh and low NA levels (Hobson, 2009). We postulate that the state-specific modulation provides a high activity target for the discriminator during Wake which is decreased during REM and entirely gated off during NREM. Furthermore, we suggest that adversarial learning is implemented by a sign-switched plasticity in the generative network during REM sleep, with respect to Wake. During wakefulness, plasticity in these apical synapses may be enhanced by noradrenaline (NA) as opposed to NREM (Adamantidis et al., 2019; Aru et al., 2020). The presence of acetylcholine (ACh) alone during REM (Hobson et al., 2000) may switch the sign of plasticity in apical synapses of (hippocampal) pyramidal neurons (McKay et al., 2007). Furthermore, it is known that somato-dendritic synchrony is reduced in REM versus NREM sleep (Seibt et al., 2017); this suggests a reduced somato-dendritic backpropagation of action potentials, which, in turn, is known to switch the sign of apical plasticity (Sjöström and Häusser, 2006).

Third, learning in our model requires the computation of reconstruction errors, i.e., mismatches between top-down and bottom-up activity. So far, two non-exclusive

	Wake	NREM	REM
ACh levels	High	Low	High
NA levels	High	Low	Low
Sensory activity	Externally driven	Internally generated	Internally generated
Discriminator output d	High activity	Gated off	Low activity
Plasticity in generator G	On	Off	Sign switch
Network meta-state	Wake	Perturbed dream	Adversarial dream
Phenomenology	☀️ **Wake**	🌙 **NREM**	⊙ **REM**

FIGURE 3.7: **Model features and physiological counterparts during Wake, NREM and REM phases.** ACh: acetylcholine; NA: noradrenaline. "Sign switch" indicates that identical local errors lead to opposing weight changes between Wake and REM sleep.

candidates for computing mismatch signals have been proposed. One suggests a dendritic error representation in layer 5 pyramidal neurons that compare bottom-up with top-down inputs from our encoding (E) and generative (G) pathways (Guerguiev et al., 2017; Sacramento et al., 2018). The other suggests an explicit mismatch representation by subclasses of layer 2/3 pyramidal neurons (Keller and Mrsic-Flogel, 2018).

Fourth, our computational framework assumes effectively separate feedforward and feedback streams. A functional separation of these streams does not necessarily imply a structural separation at the network level. Indeed, such cross-projections are observed in experimental data (Gilbert and Li, 2013) and also used in, e.g., the predictive processing framework (Rao and Ballard, 1999). In our model, an effective separation of the information flows is required to prevent "information shortcuts" across early sensory cortices which would prevent learning of good representations in higher sensory areas. This suggests that for significant periods of time, intra-areal lateral interactions between cortical feedforward and feedback pathways are effectively gated off in most of the areas.

Fifth, similar to previous work (Káli and Dayan, 2004), the hippocampus is not explicitly modeled but rather mimicked by a buffer allowing simple store and retrieve operations. An extension of our model could replace this simple mechanism with attractor networks which have been previously employed to model hippocampal function (Tang et al., 2010). The combination of episodic memories underlying REM dreams in our model could either occur in hippocampus or in cortex. In either case, we would predict a nearly simultaneous activation of different episodic memories in hippocampus that results in the generation of creative virtual early cortical activity.

Finally, beyond the mechanisms discussed above, our model assumes that cortical circuits can efficiently perform credit assignment, similar to the classical error backpropagation algorithm. Most biologically plausible implementations for error-backpropagation involve feedback connections to deliver error signals (Whittington and Bogacz, 2019; Richards et al., 2019; Lillicrap et al., 2020), for example to the apical dendrites of pyramidal neurons (Sacramento et al., 2018; Guerguiev et al., 2017;

Haider et al., 2021). An implementation of our model in such a framework would hence require additional feedforward and feedback connections for each neuron. For example, neurons in the feedforward pathway would not only project to higher cortical areas to transmit signals, but additionally project back to earlier areas to allow these to compute the local errors required for effective learning. Overall, our proposed model could be mechanistically implemented in cortical networks through different classes of pyramidal neurons with a biological version of supervised learning based on a dendritic prediction of somatic activity (Urbanczik and Senn, 2014), and a corresponding global modulation of synaptic plasticity by state-specific neuromodulators.

3.4 Discussion

Semantic representations in cortical networks emerge in early life despite most observations lacking an explicit class label, and sleep has been hypothesized to facilitate this process (Klinzing et al., 2019). However, the role of dreams in cortical representation learning remains unclear. Here we proposed that creating virtual sensory experiences by randomly combining episodic memories during REM sleep lies at the heart of cortical representation learning. Based on a functional cortical architecture, we introduced the perturbed and adversarial dreaming model (PAD) and demonstrated that REM sleep can implement an adversarial learning process which, constrained by the network architecture and the choice of latent prior distributions, builds semantically organized latent representations. Additionally, perturbed dreaming based on the episodic memory replay during NREM stabilizes the cortical representations against sensory perturbations. Our computational framework allowed us to investigate the effects of specific sleep-related pathologies on cortical representations. Together, our results demonstrate complementary effects of perturbed dreaming from individual episodes during NREM and adversarial dreaming from mixed episodes during REM. PAD suggests that the generalization abilities exhibited by humans and other animals arise from distinct processes during the two sleep phases: REM dreams organize representations semantically and NREM dreams stabilize these representations against perturbations. Finally, the model suggests how adversarial learning inspired by GANs can potentially be implemented by cortical circuits and associated plasticity mechanisms.

3.4.1 Relation to cognitive theories of sleep

PAD focuses on the functional role of sleep, and in particular dreams. Many dynamical features of brain states during NREM and REM sleep, such as cortical oscillations (Léger et al., 2018) are hence ignored here but will potentially become relevant when constructing detailed circuit models of the suggested architectures, for example for switching between memories (Korcsak-Gorzo et al., 2021). Our proposed model of sleep is complementary to theories suggesting that sleep is important for physiological and cognitive maintenance (McClelland et al., 1995; Káli and Dayan, 2004; Rennó-Costa et al., 2019; van de Ven et al., 2020). In particular, Norman et al. (2005) proposed a model where autonomous reactivation of memories (from cortex and hippocampus) coupled with oscillating inhibition during REM sleep helps detect weak parts of memories and selectively strengthen them, to overcome catastrophic forgetting. While our REM phase serves different purposes, an interesting commonality is the view of REM as a period where the cortex "thinks about what it already knows" from past and recent memories and reorganizes its representations by replaying them

together, as opposed to NREM where only recent memories are replayed and consolidated. Recent work has also suggested that the brain learns using adversarial principles, either as a reality monitoring mechanism potentially explaining delusions in some mental disorders (Gershman, 2019), in the context of dreams to overcome overfitting and promote generalization (Hoel, 2021), and for learning inference in recurrent biological networks (Benjamin and Kording, 2021b).

Cognitive theories propose that sleep promotes the abstraction of semantic concepts from episodic memories through a hippocampo-cortical replay of waking experiences, referred to as "memory semantization" (Nadel and Moscovitch, 1997; Lewis and Durrant, 2011). The learning of organized representations is an important basis for semantization. An extension of our model would consider the influence of different sensory modalities on representation learning (Guo et al., 2019), which is known to significantly influence cortical schemas (Lewis et al., 2018) and can encourage the formation of computationally powerful representations (Radford et al., 2021).

Finally, sleep has previously been considered as a state where 'noisy' connections acquired during wakefulness are selectively forgotten (Crick and Mitchison, 1983b; Poe, 2017), or similarly, as a homeostatic process to desaturate learning and renormalize synaptic strength (synaptic homeostasis hypothesis; Tononi and Cirelli, 2014, 2020). In contrast, our model offers an additional interpretation of plasticity during sleep, where synapses are globally readapted to satisfy different but complementary learning objectives than Wake, either by improving feedforward recognition of perturbed inputs (NREM) or by adversarially tuning top-down generation (REM).

3.4.2 Relation to representation learning models

Recent advances in machine learning, such as self-supervised learning approaches, have provided powerful techniques to extract semantic information from complex datasets (Liu et al., 2021). Here, we mainly took inspiration from self-supervised generative models combining autoencoder and adversarial learning approaches (Radford et al., 2015; Donahue et al., 2016; Dumoulin et al., 2017; Berthelot et al., 2018; Liu et al., 2021). It is theoretically not yet fully understood how linearly separable representations are learned from objectives which do not explicitly encourage them, i.e., reconstruction and adversarial losses. We hypothesize that the presence of architectural constraints and latent priors, in combination with our objectives, enable their emergence (see also Alemi et al., 2018; Tschannen et al., 2020). Note that similar generative machine learning models often report a higher linear separability of network representations, but use all convolutional layers as a basis for the readout (Radford et al., 2015; Dumoulin et al., 2017), while we only used low-dimensional features z. Approaches similar to ours, i.e., those which perform classification only on the latent features, report comparable performance to ours (Berthelot et al., 2018; Hjelm et al., 2019; Beckham et al., 2019a).

Furthermore, in contrast to previous GAN variants, our model removes many optimization tricks such as batch-normalization layers (Ioffe and Szegedy, 2015), spectral normalization layers (Miyato et al., 2018) or optimizing the min-max GAN objective in three steps with different objectives, which are challenging to implement in biological substrates. Despite their absence, our model maintains a high quality of latent representations. As our model is relatively simple, it is amenable to implementations within frameworks approximating backpropagation in the brain (Whittington and

Bogacz, 2019; Richards et al., 2019; Lillicrap et al., 2020). However, some components remain challenging for implementations in biological substrates, for example convolutional layers (but see Pogodin et al., 2021) and batched training (but see Marblestone et al., 2016).

3.4.3 Dream augmentations, mixing strategies and fine-tuning

To make representations robust, a computational strategy consists of learning to map different sensory inputs containing the same object to the same latent representation, a procedure reminiscent of data augmentation (Shorten and Khoshgoftaar, 2019). As mentioned above, unlike standard data augmentation methods, our NREM phase does not require the storage of raw sensory inputs to create altered inputs necessary for such data augmentation and instead relies on (hippocampal) replay being able to regenerate similar inputs from high-level representations stored during wakefulness. Our results obtained through perturbed dreaming during NREM provide initial evidence that this dream augmentation may robustify cortical representations.

Furthermore, as discussed above, introducing more specific modifications of the replayed activity, for example mimicking translations or rotations of objects, coupled with a negative phase where latent representations from different images are pushed apart, may further contribute to the formation of invariant representations. Along this line, recent self-supervised contrastive learning methods (Gidaris et al., 2018; Chen et al., 2020; Zbontar et al., 2021) have been shown to enhance the semantic structure of latent representations by using a similarity objective where representations of stimuli under different views are pulled together in a first phase, while, crucially, embedding distances between unrelated images are increased in a second phase.

In our REM phase, different mixing strategies in the latent layer could be considered. For instance, latent activities could be mixed up by retaining some vector components of a representation and using the rest from a second one (Beckham et al., 2019a). Moreover, more than two memory representations could have been used. Alternatively, our model could be trained with spontaneous cortical activity only. In our experimental setting we do not observe significant differences between using a combination of episodic memories with spontaneous activity or only using spontaneous activity (Fig. 3.13). However, we hypothesize that for models which learn continuously, a preferential replay of combinations of recent episodic memories encourages the formation of cortical representations that are useful in the present.

Here, we used a simple linear classifier to measure the quality of latent representations, which is an obvious simplification with regard to cortical processing. Note however that also for more complex 'readouts', organized latent representations enable more efficient and faster learning (Silver et al., 2017; Ha and Schmidhuber, 2018b; Schrittwieser et al., 2020). In its current form, PAD assumes that training the linear readout does not lead to weight changes in the encoder network. However, in cortical networks, cognitive or motor tasks leveraging latent representations likely shape the encoder network, which could in our model be reflected in 'fine-tuning' the encoder for specific tasks (compare Liu et al., 2021).

Finally, our model does not show significant differences in performance when the order of sleep phases is switched (Fig. 3.14). However, NREM and REM are observed to occur in a specific order throughout the night (Diekelmann and Born, 2010) and this order has been hypothesized to be important for memory consolidation ("sequential

hypothesis", Giuditta et al., 1995). The independence of phases in our model may be due to the relatively small synaptic changes occurring in each phase. We expect the order of sleep phases to influence model performance if these changes become larger, either due to longer phases or increased learning rates. The latter may become particularly relevant in continual learning settings where it becomes important to control the emphasis put on recent observations.

3.4.4 Signatures of generative learning

PAD makes several experimentally testable predictions at the neuronal and systems level. We first address generally whether the brain learns via generative models during sleep before discussing specific signatures of adversarial learning.

First, our NREM phase assumes that hippocampal replay generates perturbed wake-like early sensory activity (see also Ji and Wilson, 2007) which is subsequently processed by feedforward pathways. Moreover, our model predicts that over the course of learning, sensory-evoked neuronal activity and internally-generated activity during sleep become more similar. In particular, we predict that (spatial) activity in both NREM and REM become more similar to Wake, however, patterns observed during REM remain distinctly different due to the creative combination of episodic memories. Future experimental studies could confirm these hypotheses by recording early sensory activity during wakefulness, NREM and REM sleep at different developmental stages and evaluating commonalities and differences between activity patterns. Previous work has already demonstrated increasing similarity between stimulus-evoked and spontaneous (generated) activity patterns during wakefulness in ferret visual cortex (Berkes et al., 2011; but see Avitan et al., 2021).

On a behavioral level, the improvement of internally-generated activity patterns correlates with the development of dreams in children, that are initially unstructured, simple and plain, and gradually become full-fledged, meaningful, narrative, implicating known characters and reflecting life episodes (Nir and Tononi, 2010). In spite of their increase in realism, REM dreams in adulthood are still reported as bizarre (Williams et al., 1992). Bizarre dreams, such as a "flying dogs", are typically defined as discontinuities or incongruities of the sensory experience (Mamelak and Hobson, 1989) rather than completely structureless experiences. This definition hence focuses on high-level logical structure, not on the low-level sensory content. In contrast, the low FID score, i.e., high realism, of REM dreams in our experiments reflects that the low-level structure on which this evaluation metric mainly focuses (e.g., Brendel and Bethge, 2019) is similar to actual sensory input. Capturing the "logical realism" of our generated neuronal activities most likely requires a more sophisticated evaluation metric and an extension of the model capable of generating temporal sequences of sensory stimulation. We note, however, that even such surreal dreams as "flying dogs" can be interpreted as altered combinations of episodic memories and thus, in principle, can arise from our model.

Second, our model suggests that the development of semantic representations is mainly driven by REM sleep. This allows us to make predictions which connect the network with the systems level, in the specific case of acquiring skills from complex and unfamiliar sensory input. For humans, this could be learning a foreign language with unfamiliar phonetics. Initially, cortical representations cannot reflect relevant nuances in these sounds. Phonetic representations develop gradually over experience and are reflected in changes of the sensory evoked latent activity, specifically in the

reallocation of neuronal resources to represent the relevant latent dimensions. We hypothesize that in case of impaired REM sleep, this change of latent representations is significantly reduced, which goes hand in hand with decreased learning speed. Future experimental studies could investigate these effects for instance by trying to decode sound identity from high-level cortical areas in patients where REM sleep is impaired over long periods through pharmacological agents such as anti-depressants (Boyce et al., 2017). An equivalent task in the non-human animal domain would be song acquisition in songbirds (Fiete et al., 2007). On a neuronal level, one could selectively silence feedback pathways during REM sleep in animal models over many nights, for example via optogenetic tools. Our model predicts that this silencing would significantly impact the animal's learning speed, as reported from animals with reduced theta rhythm during REM sleep (Boyce et al., 2017).

3.4.5 Signatures of adversarial learning

The experimental predictions discussed above mainly address whether the brain learns via generative models during sleep. Here we make experimental predictions which would support our hypotheses and contrast it to alternative theories of learning during offline states.

Existence of an external/internal discriminator. The discriminator provides our model with the ability to distinguish externally driven from internally generated low-level cortical activity. Due to this unique property, the discriminator may be leveraged to distinguish actual from imagined sensations. According to our model, reduced REM sleep would lead to an impaired discriminator, and could thus result in an inability of subjects to realize that self-generated imagery is not part of the external sensorium. This may result in the formation of delusions, as previously suggested (Gershman, 2019). For instance, hallucinations in schizophrenic patients, often mistaken for veridical perceptions (Waters et al., 2016), could be partially caused by abnormal REM sleep patterns, related to observed reduced REM latency and density (Cohrs, 2008). Based on these observations, we predict in the context of our model a negative correlation between REM sleep quality and delusional perceptions of hallucinations. Systematic differences in REM sleep quality may hence explain why some patients are able to recognize that their hallucinations are self-generated while some others mistake them to be real. Moreoever, although locating discriminator neurons may prove non-trivial (but see "Cortical implementation of PAD" for specific suggestions), we predict that once the relevant cells have been identified, perturbing them may lead to detrimental effects on differentiating between external sensory inputs and internally generated percepts.

The state-specific activity of the discriminator population makes predictions about plasticity on synapses in the feedforward stream during wakefulness and sleep. In our model, the discriminator is trained to distinguish externally from internally generated patterns by opposed targets imposed during Wake and REM. After many wake-sleep cycles, the KL loss as well as the reconstruction loss (see Methods) in our model become small compared to the adversarial loss (Fig. 3.10, Fig. 3.11), which remains non-zero due to a balance between discriminator and generator. The same low-level activity pattern would hence cause opposite weight changes during wakefulness and sleep on feedforward synapses. This could be tested experimentally by actively instantiating similar spatial activity patterns in low-level sensory cortex during wakefulness and REM and compare the statistics of observed changes in (feedforward) downstream synapses.

Adversarial training of a generator during sleep. To drive adversarial learning and maintain a balance between the generator and discriminator, the generative network must be trained in parallel to the discriminative (encoder) network during REM. In contrast, in alternative representation learning models which involve offline states such as the Wake-Sleep algorithm (Hinton et al., 1995), generative pathways are not trained to produce realistic dreams during the sleep phase. Rather, they are trained by reconstruction on real input data during the wake phase. This allows an experimental distinction between our model and Wake-Sleep-like models: while our model predicts plasticity in both bottom-up and top-down pathways both during wake and during REM sleep, Wake-Sleep models alternate between training feedback and feedforward connections during online and offline states, respectively.

Previous work has developed methods to infer plasticity rules from neuronal activity (Lim et al., 2015; Senn and Sacramento, 2015) or weight changes (Nayebi et al., 2020). In the spirit of existing *in vivo* experiments, we suggest to optogenetically monitor and potentially modulate apical dendritic activities in cortical pyramidal neurons of mice during wakefulness and REM sleep (Li et al., 2017; Voigts and Harnett, 2020; Schoenfeld et al., 2022). From the statistics of the recorded dendritic and neuronal activity, the plasticity rules could be inferred and compared to the state-dependent rules suggested by our model, in particular to the predicted sign-switch of plasticity between wakefulness and REM sleep.

Adversarial learning and creativity. Adversarial learning, for example in GANs, enables a form of creativity, reflected in their ability to generate realistic new data or to create semantically meaningful interpolations (Radford et al., 2015; Berthelot et al., 2018; Karras et al., 2018). This creativity might be partly caused by the freedom in generating sensory activity that is not restricted by requiring good reconstructions, but is only guided by the internal/external judgment (Goodfellow, 2016). This is less constraining on the generator than direct reconstruction losses used in alternative models such as variational auto-encoders (Kingma and Welling, 2013) or the Wake-Sleep algorithm (Hinton et al., 1995). We thus predict that REM sleep, here implementing adversarial learning, should boost creativity, as previously reported (Cai et al., 2009; Llewellyn, 2016a; Lewis et al., 2018). Furthermore, we predict that REM sleep influences a subject's ability to visualize creative mental images, for instance associating non-obvious visual patterns from distinct memories. For example, we predict that participants chronically lacking REM sleep would perform worse than control participants at a creative synthesis task (Palmiero et al., 2015), consisting of combining different visual components into a new, potentially useful object.

Adversarial learning and lucid dreaming. Finally, adversarial dreaming offers a theoretical framework to investigate neuronal correlates of normal versus lucid dreaming (Dresler et al., 2012; Baird et al., 2019). While in normal dreaming the internally generated activity is perceived as externally caused, in lucid dreaming it is perceived as what it is, i.e., internally generated. We hypothesize that the "neuronal conductor" that orchestrates adversarial dreaming is also involved in lucid dreaming, by providing to the dreamer conscious access to the target "internal" that the conductor imposes during REM sleep. Our cortical implementation suggests that the neuronal conductor could gate the discriminator teaching via apical activity of cortical pyramidal neurons. The same apical dendrites were also speculated to be involved in conscious perception (Takahashi et al., 2020), dreaming (Aru et al., 2020), and in representing the state and content of consciousness (Aru et al., 2019).

Our model demonstrates that adversarial learning during wakefulness and sleep can provide significant benefits to extract semantic concepts from sensory experience. By bringing insights from modern artificial intelligence to cognitive theories of sleep function, we suggest that cortical representation learning during dreaming is a creative process, orchestrated by brain-state-regulated adversarial games between separated feedforward and feedback streams. Adversarial dreaming may further be helpful to understand learning beyond the standard student-teacher paradigm. By 'seeing' the world from new perspectives every night, dreaming represents an active learning phenomenon, constantly improving our understanding, our creativity and our awareness.

3.5   Methods

3.5.1   Network architecture

The network consists of two separate pathways, mapping from the pixel to the latent space ('encoder'/'discriminator') and from the latent to pixel space ('generator'). Encoder/Discriminator and Generator architectures follow a similar structure as the DCGANs model (Radford et al., 2015). The encoder E_z has four convolutional layers (LeCun et al., 2015) containing $64, 128, 256$ and 256 channels respectively (Fig. 3.8). Each layer uses a 4×4 kernel, a padding of 1 (0 for last layer), and a stride of 2, i.e.,
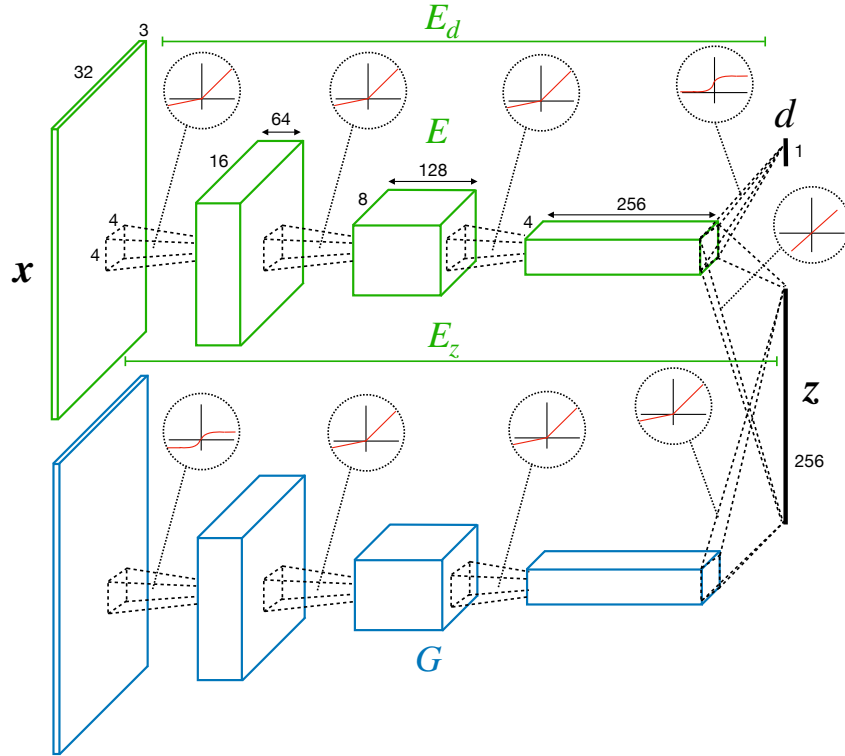
FIGURE 3.8: **Convolutional neural network (CNN) architecture of encoder/discriminator and generator used in PAD.**

feature size is halved in each layer. All convolutional layers except the last one are followed by a LeakyReLU non-linearity (Maas et al., 2013). We denote the activity in the last convolutional layer as z. An additional convolutional layer followed by a sigmoid non-linearity is added on top of the second-to-last layer of the encoder and

maps to a single scalar value d, the internal/external discrimination (with putative teaching signal 0 or 1). We denote the mapping from $\boldsymbol{x}$ to d by E_d. E_z and E_d thus share the first three convolutional layers. We jointly denote them by E, where $E(\boldsymbol{x}) = (E_z(\boldsymbol{x}), E_d(\boldsymbol{x})) = (\boldsymbol{z}, d)$ (Fig. 3.8).

Mirroring the structure of E_z, the generator G has four deconvolutional layers containing $256, 128, 64$, and 3 channels. They all use a 4×4 kernel, a padding of 1 (0 for first deconvolutional layer) and a stride of 2, i.e, the feature-size is doubled in each layer. The first three deconvolutional layers are followed by a LeakyReLU non-linearity, and the last one by a tanh non-linearity.

As a detailed hippocampus model is outside the scope of this study, we mimic hippocampal storage and retrieval by storing and reading latent representations to and from memory.

3.5.2 Datasets

We use the CIFAR-10 (Krizhevsky et al., 2013) and SVHN (Netzer et al., 2011) datasets to evaluate our model. They consist of 32×32 pixel images with three color channels. We consider their usual split into a training set and a smaller test set.

3.5.3 Training procedure

We train our model by performing stochastic gradient-descent with mini-batches on condition-specific objective functions, in the following also referred to as loss functions, using the ADAM-optimizer ($\beta_1 = 0.5$, $\beta_2 = 0.999$; Kingma and Ba, 2017) with learning rate of 0.0002 and mini-batch size of 64. We rely on our model being fully differentiable. The following section describes the loss functions for the respective conditions.

3.5.3.1 Loss functions

Wake In the Wake condition, we minimize the following objective function, composed of a loss for image encoding, a regularization, and a real/fake (external/internal) discriminator,

$$\mathcal{L}_{\text{Wake}} = \mathcal{L}_{\text{img}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{real}} \,. \tag{3.1}$$

E_z and G learn to reconstruct the mini-batch of images $\boldsymbol{X} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, ..., \boldsymbol{x}^{(b)}\}$ similarly to autoencoders (Bengio et al., 2013) by minimizing the image reconstruction loss $\mathcal{L}_{\text{img}}$ defined by

$$\mathcal{L}_{\text{img}} = \frac{1}{b} \sum_{i=1}^{b} \|\boldsymbol{x}^{(i)} - G(E_z(\boldsymbol{x}^{(i)}))\|^2 \,, \tag{3.2}$$

where b denotes the size of the mini-batch. We store the latent vectors $\boldsymbol{Z} = E_z(\boldsymbol{X})$ corresponding to the current mini-batch for usage during the NREM and REM phases.

We additionally impose a Kullback-Leibler (KL) divergence loss on the encoder E_z. This acts as a regularizer and encourages latent activities to be Gaussian with zero

---

Algorithm 1: Training procedure

θ_E, θ_G ;   // initialize network parameters
for *number of training iterations* **do**

    Wake
    $\boldsymbol{X} \leftarrow \{\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(b)}\}$;   // random mini-batch from dataset
    $\boldsymbol{Z}, \boldsymbol{D} \leftarrow E(\boldsymbol{X})$;   // infer latent and discriminative outputs
    $\boldsymbol{X}' \leftarrow G(\boldsymbol{Z})$;   // reconstruct input via generator
    $\mathcal{L}_{\text{img}} \leftarrow \frac{1}{b} \sum_{i=1}^{b} \|\boldsymbol{x}^{(i)} - \boldsymbol{x}'^{(i)}\|^2$;   // compute reconstruction loss
    $\mathcal{L}_{\text{KL}} \leftarrow \text{D}_{\text{KL}}(q(\boldsymbol{Z})\|p(\boldsymbol{Z}))$;   // compute KL-loss
    $\mathcal{L}_{\text{real}} \leftarrow -\frac{1}{b} \sum_{i=1}^{b} \log(\boldsymbol{d}^{(i)})$;   // compute discriminator loss on real
    samples
    $\theta_E \leftarrow \theta_E - \nabla_{\theta_E}(\mathcal{L}_{\text{img}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{real}})$;   // update encoder/discriminator
    parameters
    $\theta_G \leftarrow \theta_G - \nabla_{\theta_G}\mathcal{L}_{\text{img}}$;   // update generator parameters

    NREM sleep
    $\boldsymbol{Z} \leftarrow \{\boldsymbol{z}^{(1)}, ..., \boldsymbol{z}^{(b)}\}$;   // mini-batch of latent vectors from Wake
    $\boldsymbol{X}' \leftarrow G(\boldsymbol{Z})$;   // reconstruct input via generator
    $\boldsymbol{Z}' \leftarrow E_z(\boldsymbol{X}' \odot \boldsymbol{\Omega})$;   // infer perturbed input
    $\mathcal{L}_{\text{NREM}} \leftarrow \frac{1}{b} \sum_{i=1}^{b} \|\boldsymbol{z}^{(i)} - \boldsymbol{z}'^{(i)}\|^2$;   // compute reconstruction loss
    $\theta_E \leftarrow \theta_E - \nabla_{\theta_E}\mathcal{L}_{\text{NREM}}$

    REM sleep
    if *first iteration* **then**
      | $\boldsymbol{Z}_{\text{mix}} \leftarrow \boldsymbol{Z}$
    else
      | $\boldsymbol{Z}_{\text{mix}} \leftarrow \lambda'(\lambda\boldsymbol{Z} + (1-\lambda)\boldsymbol{Z}_{\text{old}}) + (1-\lambda')\boldsymbol{\epsilon}$;   // convex combination of
      current and old latent vectors with noise
    end
    $\boldsymbol{D} \leftarrow E_d(G(\boldsymbol{Z}_{\text{mix}}))$
    $\mathcal{L}_{\text{REM}} \leftarrow -\frac{1}{b} \sum_{i=1}^{b} \log(1 - \boldsymbol{d}^{(i)}))$;   // compute adversarial loss
    $\theta_E \leftarrow \theta_E - \nabla_{\theta_E}\mathcal{L}_{\text{REM}}$
    $\theta_G \leftarrow \theta_G + \nabla_{\theta_G}\mathcal{L}_{\text{REM}}$;   // gradient ascent on discriminator loss
    $\boldsymbol{Z}_{\text{old}} \leftarrow \boldsymbol{Z}$;   // keep current vectors for next iteration
end

---

mean and unit variance:

$$\mathcal{L}_{\text{KL}} = \text{D}_{\text{KL}}(q(\boldsymbol{Z}|\boldsymbol{X})\|p(\boldsymbol{Z})) , \tag{3.3}$$

where $q(\boldsymbol{Z}|\boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ is a distribution over the latent variables $\boldsymbol{Z}$, parametrized by mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$, and $p(\boldsymbol{Z}) \sim \mathcal{N}(0, 1)$ is the prior distribution over latent variables. E_z is trained to minimize the following loss:

$$\mathcal{L}_{\text{KL}} = \frac{1}{2n_z} \sum_{j=1}^{n_z} \left(\mu_j^{(\boldsymbol{Z})^2} + \sigma_j^{(\boldsymbol{Z})^2} - 1 - \log(\sigma_j^{(\boldsymbol{Z})^2}) \right) , \tag{3.4}$$

where n_z denotes the dimension of the latent space and where $\mu_j^{(\boldsymbol{Z})}$ and $\sigma_j^{(\boldsymbol{Z})}$ represent

the jth elements of respectively the empirical mean $\boldsymbol{\mu}^{(\boldsymbol{Z})}$ and empirical standard deviation $\boldsymbol{\sigma}^{(\boldsymbol{Z})}$ of the set of latent vectors $E_z(\boldsymbol{X}) = \boldsymbol{Z}$.

As part of the adversarial game, E_d is trained to classify the mini-batch of images as real. This corresponds to minimizing the loss defined as sum across the mini-batch size b,

$$\mathcal{L}_{\text{real}} = \mathcal{L}_{\text{GAN}}(E_d(\boldsymbol{X}), 1) = -\frac{1}{b} \sum_{i=1}^{b} \log(E_d(\boldsymbol{x}^{(i)})) \,. \tag{3.5}$$

Note that, in principle, $\mathcal{L}_{\text{GAN}}$ can be any GAN-specific loss function (Gui et al., 2020). Here we choose the binary cross-entropy loss.

NREM sleep Each Wake phase is followed by a NREM phase. During this phase we make use of the mini-batch of latent vectors $\boldsymbol{z}$ stored during the Wake phase. Starting from a mini-batch of latent vectors, we generate images $G(z)$. Each obtained image of $G(\boldsymbol{z})$ is multiplied by a binary occlusion mask $\boldsymbol{\omega}$ of the same dimension. This mask is generated by randomly picking two occlusion parameters, occlusion intensity and square size (for details see Section 3.5.3.2). The encoder E_z learns to reconstruct the latent vectors $\boldsymbol{z}$ by minimizing the following reconstruction loss:

$$\mathcal{L}_{\text{NREM}} = \frac{1}{b} \sum_{i=1}^{b} \| \boldsymbol{z}^{(i)} - E_z \left(G(\boldsymbol{z}^{(i)}) \odot \boldsymbol{\omega} \right) \|^2 \,, \tag{3.6}$$

where $\odot$ denotes the element-wise product.

REM sleep In REM, each latent vector from the mini-batch considered during Wake is combined with the latent vector from the previous mini-batch, the whole being convex combined with a mini-batch of noise vectors $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$, where I is the identity matrix, leading to a mini-batch of latent vectors $\boldsymbol{Z}_{\text{mix}} = \lambda'(\lambda \boldsymbol{Z} + (1 - \lambda)\boldsymbol{Z}_{\text{old}}) + (1 - \lambda')\boldsymbol{\epsilon}$. Here, $\lambda = 0.5$ and $\lambda' = 0.5$, where $\boldsymbol{Z}_{\text{old}}$ is the previous mini-batch of latent activities. This batch of latent vectors is passed through G to generate the associated images $G(\boldsymbol{Z}_{\text{mix}})$. In this phase, the loss function encourages E_d to classify $G(\boldsymbol{Z}_{\text{mix}})$ as fake, while adversarially pushing G to generate images which are less likely to be classified as fake by the minimax objective

$$\min_{E_d} \max_{G} \mathcal{L}_{\text{REM}} \,, \tag{3.7}$$

where

$$\mathcal{L}_{\text{REM}} = \mathcal{L}_{\text{GAN}}(E_d(G(\boldsymbol{Z}_\lambda)), 0) = -\frac{1}{b} \sum_{i=1}^{b} \log(1 - E_d(G(\mathbf{z}_\lambda^{(i)}))) \,. \tag{3.8}$$

In our model, the adversarial process is simply described by a full backpropagation of error through E_d and G with a sign switch of weight changes in G.

In summary, each Wake-NREM-REM cycle consists of: 1) reconstructing a mini-batch $\boldsymbol{x}$ of images during Wake, 2) reconstructing a mini-batch of latent activities $\boldsymbol{Z} = E_z(\boldsymbol{X})$ during NREM with perturbation of $G(z)$, and 3) replaying $\boldsymbol{Z}$ convex combined with $\boldsymbol{Z}_{\text{old}}$ and noise from the $(n-1)$-th cycle. In PAD training, all losses are weighted equally and we did not use a schedule for $\mathcal{L}_{\text{KL}}$, as opposed to standard Variational Autoencoder (VAE) training (Kingma and Welling, 2013). One training epoch is defined by the number of mini-batches necessary to cover the whole dataset. The

evolution of losses with training epochs is shown in Fig. 3.10 and Fig. 3.11. The whole training procedure is summarized in the pseudo-code implemented in Algorithm 1.

3.5.3.2 Image occlusion

FIGURE 3.9: **Varying size and intensity of occlusions on example images from CIFAR-10.** Image occlusions vary along 2 parameters: occlusion intensity, defined by the probability to apply a grey square at a given position, and square size (s).

Following previous work (Zeiler and Fergus, 2013), grey squares of various sizes are applied along the image with a certain probability (Fig. 3.9). For each mini-batch, a probability and square size were randomly picked between 0 and 1, and $1-8$ respectively. We divide the image into patches of the given size and we replace each patch with a constant value (here, 0) according to the defined probability.

3.5.4 Evaluation

3.5.4.1 Training of linear read-out

A linear classifier is trained on top of latent features $\boldsymbol{Z} = E_z(\boldsymbol{X})$, with $\boldsymbol{Z} \in \mathbb{R}^{N \times 256}$, where N is the number of training dataset images. A latent feature $\boldsymbol{z} \in \mathbb{R}^{256}$ is projected via a weight matrix $W \in \mathbb{R}^{10 \times 256}$ to the label neurons to obtain the vector $\boldsymbol{y} = W\boldsymbol{z}$.

This weight matrix is trained in a supervised fashion by using a multi-class cross-entropy loss. For a feature $\boldsymbol{z}$ labelled with a target class $t \in \{0, 1, .., 9\}$, the per-sample classification loss is given by

$$\mathcal{L}^C(\boldsymbol{z}, t; W) = -\log p_W(Y = t | \boldsymbol{z}) \ . \tag{3.9}$$

Here, $p_{\mathbf{W}}$ is the conditional probability of the classifier defined by the linear projection and the softmax function

$$p_W(Y = t | \boldsymbol{z}) = \frac{e^{y_t}}{\sum_{i=0}^{9} e^{y_i}} \ . \tag{3.10}$$

The classifier is trained by mini-batch ($b = 64$) stochastic gradient descent on the loss $\mathcal{L}^C$ with a learning rate $\eta = 0.2$ for 20 epochs, using the whole training dataset.

3.5.4.2 Linear separability

Following previous work (Hjelm et al., 2019), we define linear separability as the classification accuracy of the trained classifier on inferred latent activities $E_z(\boldsymbol{X}_{\text{test}})$ from a separate test dataset $\boldsymbol{X}_{\text{test}}$. Given a latent feature $\boldsymbol{z}$, class prediction is made by picking the index of the maximal activity in the vector $\boldsymbol{y}$. We ran several simulations for 4 different initial parameters of E and G and report the average test accuracy and standard error of the mean over trials. To evaluate performance on occluded data, we applied random square occlusion masks on each sample from

$\boldsymbol{X}_{\text{test}}$ for a fixed probability of occlusion and square size. We report only results for occulusions of size 4, after observing similar results with other square sizes.

3.5.4.3 PCA visualization

To visualize the 256-dimensional latent representation $E_z(\boldsymbol{x})$ of the trained model we used the Principal Component Analysis reduction algorithm (Jolliffe and Cadima, 2016). We project the latent representations to the first two principle components.

3.5.4.4 Latent-space organization metrics

Intra-class distance is computed by randomly picking $1,000$ pairs of images of the same class, projecting them to the encoder latent space $\boldsymbol{z}$ and computing their Euclidian distance. This process is repeated over the 10 classes in order to obtain the average over 10 classes. Similarly, inter-class distance is computed by randomly picking $10,000$ pairs of images of different classes, projecting them to the encoder latent space $\boldsymbol{z}$ and computing their Euclidian distance. The ratio of intra- and inter-class distance is obtained by dividing the mean intra-class distance by the mean inter-class distance. Clean-occluded distance is computed by randomly picking $10,000$ pairs of non-occluded/occluded images, projecting them to the encoder latent space and computing their Euclidian distance. The ratio of clean-occluded and inter-class distance is obtained by dividing the clean-occluded distance by the mean inter-class distance. We performed this analysis for several different trained networks with different initial conditions and report the mean ratios and standard error of the mean over trials.

3.5.4.5 Fréchet inception distance

Following Heusel et al. (2018), Fréchet inception distance (FID) is computed by comparing the statistics of generated (NREM or REM) samples to real images from the training dataset projected through an Inception-v3 network pre-trained on ImageNet

$$\text{FID} = \|\mu_{\text{real}} - \mu_{\text{gen}}\|^2 + \text{Tr}(\Sigma_{\text{real}} + \Sigma_{\text{gen}} - 2(\Sigma_{\text{real}}\Sigma_{\text{gen}})^{1/2}) \qquad (3.11)$$

where μ and Σ represent the empirical mean and covariance of the 2048-dimensional activations of the Inception v3 pool3 layer for $10,000$ pairs of data samples and generated images. Results represent mean FID and standard error of the mean FID over 4 different trained networks with different initializations.

3.5.4.6 Modifications specific to pathological models

To evaluate the differential effects of each phase, we removed NREM and/or REM phases from training (Fig. 3.4, Fig. 3.5, Fig. 3.6). For instance, for the condition w/o NREM, the network is never trained with NREM.

A few adjustments were empirically observed to be necessary in order to obtain a fair comparison between each condition. When removing the REM phase during training, we observed a decrease of linear separaribility after some (> 25) epochs. We suspect that this decrease is a result of overfitting due to unconstrained autoencoding objective of E and G. Models trained without REM hence would not provide a good baseline to reveal the effect of adversarial dreaming on linear separability. For models without the REM phase, we hence added a vector of Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 0.5 \cdot I)$ to the encoded activities $E_z(\boldsymbol{X})$ of dimension n_z before feeding them to the generator.

Thus, Eq. Eq. 3.2 becomes:

$$\mathcal{L}_{\text{img}} = \frac{1}{b} \sum_{i=1}^{b} \| \boldsymbol{x}^{(i)} - G\left(E_z(\boldsymbol{x}^{(i)}) + \boldsymbol{\epsilon} \right) \|^2 \, , \tag{3.12}$$

which stabilizes linear separability of latent activities around its maximal value for both CIFAR-10 and SVHN datasets until the end of training.

Furthermore, we observed that the NREM phase alters linear performance in the absence of REM (w/o REM condition). To overcome this issue, we reduced the effect of NREM by scaling down its loss with a factor of 0.5. This enabled to benefit from NREM (recognition under image occlusion) without altering linear separability on full images.

3.6 Acknowledgements

3.7 Supplementary information

3.7.1 Training losses for full and pathological models

In the following, we report the measured losses over training for the various different pathological conditions. $\mathcal{L}_{\text{img}}$ and $\mathcal{L}_{\text{KL}}$ are optimized for each condition and systematically decrease with learning, while $\mathcal{L}_{\text{NREM}}$ is significantly reduced in models with NREM (Fig. 3.10, Fig. 3.11). Its initial increase in the models with REM is explained to its competitive optimization with the GAN losses. Generator loss $\mathcal{L}_{\text{fake}} = \mathcal{L}_{\text{REM}}$ and discriminator loss $\mathcal{L}_{\text{real}} + \mathcal{L}_{\text{fake}}$ are only optimized in models with REM, showing a progressive decrease of the discriminator loss in parallel with an increase of the generator loss, reflecting adversarial learning between the two streams.

3.7.2 Linear classification performance

We report the mean and standard error of the mean (SEM) of the final linear classification performance (epoch 50) on latent representations of from the PAD and pathological models in Table 3.1.

We also report the linear classification performance for the full and pathological models over 100 epochs. Linear separability for the "w/o REM" (Fig. 3.12c,d, pink curves) and "w/o memory mix" (Fig. 3.12d, purple curve) conditions do not reach levels of the full model (Fig. 3.12c,d, black curves) even after many training epochs.
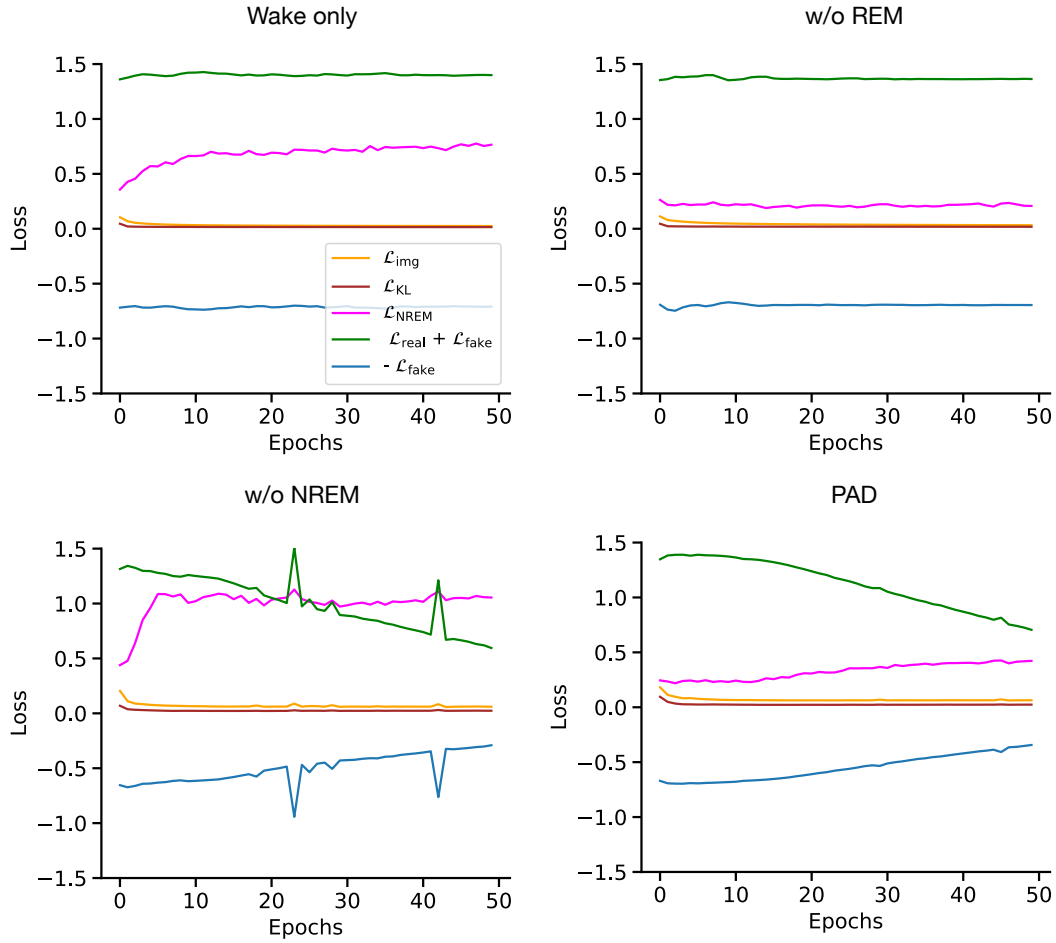
FIGURE 3.10: **Training losses for full and pathological models with CIFAR-10 dataset.** Evolution of training losses used to optimize E and G networks (see Methods) over training epochs for full and pathological models.

Furthermore, without NREM (Fig. 3.12c,d, "w/o NREM" and "Wake only", orange and gray curves), linear separability tends to decrease after many training epochs, suggesting that NREM helps to stabilize performance with training by preventing overfitting.

3.7.3   Comparison of performance with REM driven by convex combination or noise

We report the linear classifier performance for PAD using different latent inputs to the generator. In the main text, we use a convex combination of mixed memories (being a

Dataset	PAD	w/o memory mix	w/o REM	w/o NREM	Wake only
CIFAR-10	58.25 ± 0.70	53.87 ± 0.85	46.00 ± 0.43	58.00 ± 0.34	42.25 ± 0.54
SVHN	78.92 ± 0.40	60.87 ± 5.07	42.30 ± 1.51	73.25 ± 0.22	41.93 ± 0.65

TABLE 3.1:   **Final classification performance for full model and all pathological conditions for un-occluded images .** Mean and SEM over 4 different initial condition of linear separability of latent representations at the end of training (epoch 50) for PAD and its pathological variants.
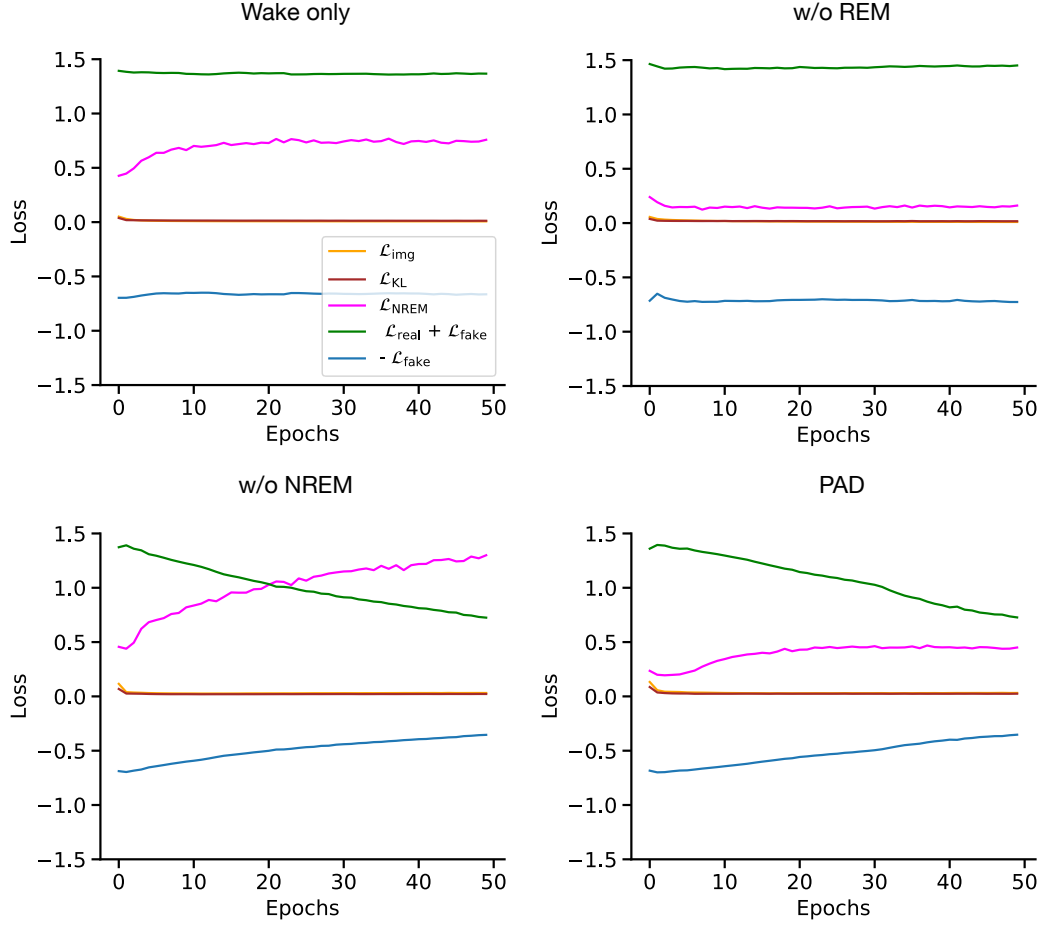
FIGURE 3.11: **Training losses for full and pathological models with SVHN dataset.**

convex combination of two different replayed latent vectors) and noise sampled from a Gaussian unit distribution (Fig. 3.13, black). We here show the results when only random Gaussian noise is used (Fig. 3.13, green) and when only a convex combination of memories is used (Fig. 3.13, red). These different mixing strategies do not show a big difference in linear separability over training epochs.

3.7.4 The order of sleep phases has no influence on the performance of the linear classifier

To investigate the role of the order of NREM and REM sleep phases, we consider a variation in which their order is reversed with respect to the model described in the main manuscript. The performance of the linear classifier is not influenced by this change (Fig. 3.14).

3.7.5 Replaying multiple episodic memories during NREM sleep

While in the main text we considered NREM to use only a single episodic memory, here we report results for a model in which also NREM uses multiple (here: two) episodic memories. In the full model (Fig. 3.15, black curves, same data as in Fig. 3.5c,d), NREM uses a single stored latent representation. Here we additionally
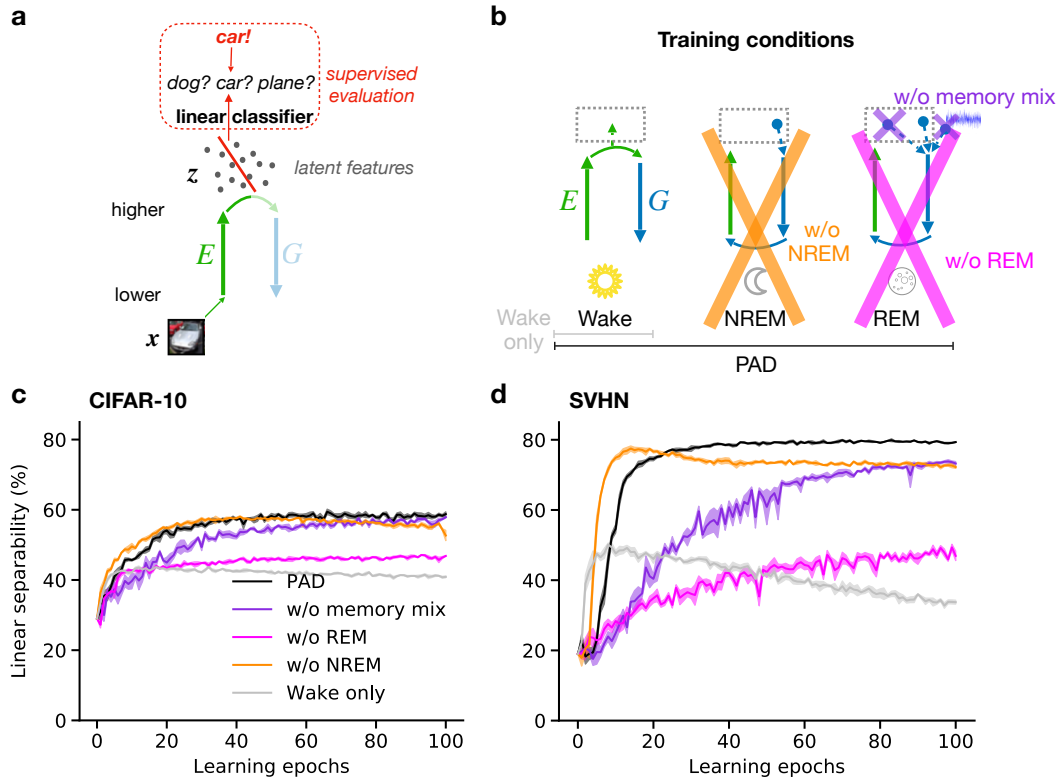
FIGURE 3.12: **Linear classification performance for full model and all pathological conditions.** For details see Fig. 3.4.

FIGURE 3.13: **Linear classification performance for different mixing strategies during REM.** Linear separability of latent representations with training epochs for PAD trained with different REM phases: one driven by a convex combination of mixed memories and noise (black), one by pure noise (green), and one by mixed memories only (red). For details see Fig. 3.4.

consider an additional model in which these representations are obtained from a convex combination of mixed memories and spontaneous cortical activity. The better performance of a single replay suggests that replay from single episodic memories as postulated to occur during NREM sleep is more efficient to robustify latent representations against input perturbations.
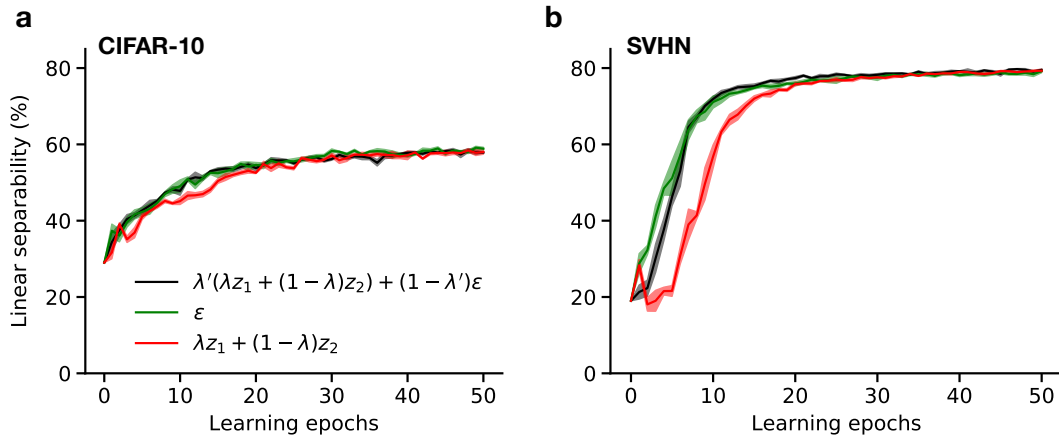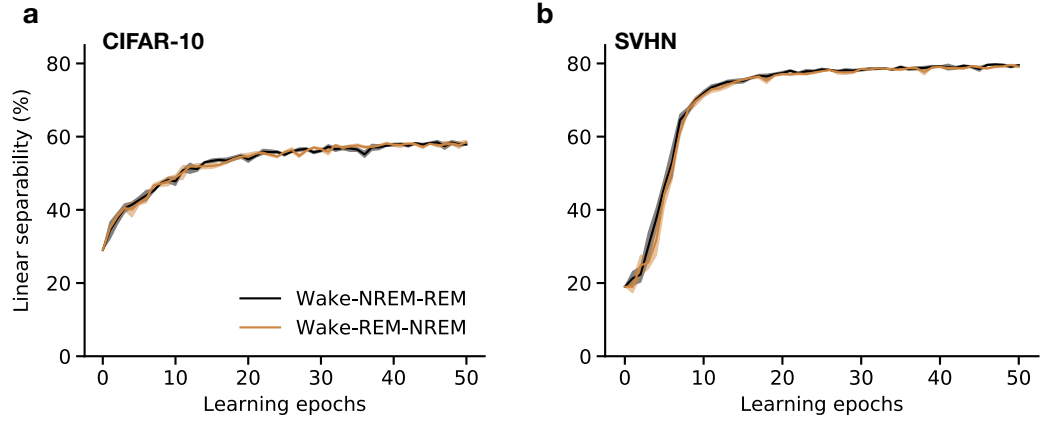
FIGURE 3.14: **Linear classification performance for different order of sleep phases.** Linear separability of latent representations with training epochs for PAD trained when NREM precedes REM phase (Wake-NREM-REM, black) or when REM precedes NREM (Wake-REM-NREM, brown).
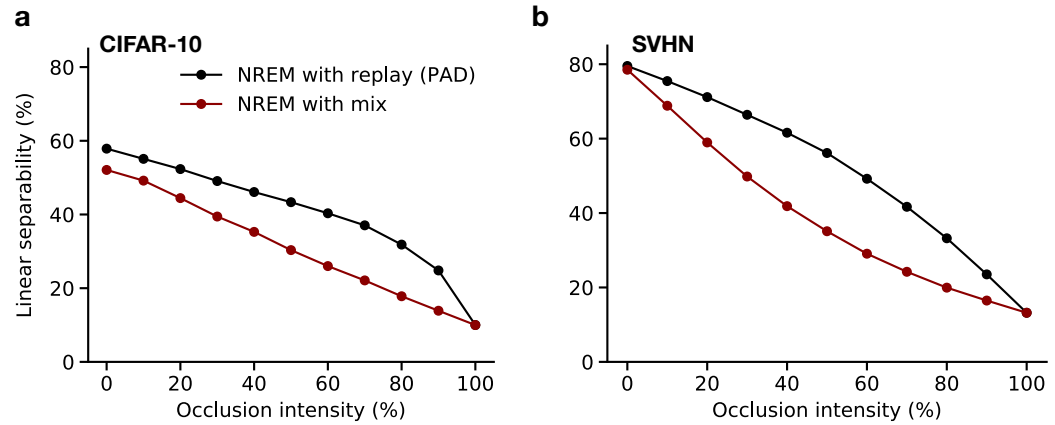
FIGURE 3.15: **Importance of replaying single hippocampal memories during NREM.** Linear separability of latent representations at the end of learning with occlusion intensity for a model trained with all phases.

Chapter 4

A role of dreaming in a semi-supervised regime

This chapter contains an extension of the PAD model from the manuscript *Learning cortical representations through perturbed and adversarial dreaming* to a semi-supervised learning regime.

Model simulations and figure plots were performed by me. Research design and results were discussed with Jakob Jordan.

4.1 Introduction

In the previous chapter, we introduced a computational model inspired from generative modeling that characterized the potential role of NREM and REM dreams in learning semantic representations. We demonstrated that a combination of state-dependent unsupervised objectives could by itself facilitate the emergence of robust and semantically organized representations. In particular, we showed that REM creative dreams together with wakefulness could host an adversarial game between feedforward and feedback pathways that significantly improves the performance of the model over solely reconstructing external inputs during Wake. This observation is in line with machine learning studies reporting a better performance at representation learning for GANs and their variants over explicit models such as AEs and VAEs (Donahue et al., 2016; Berthelot et al., 2018; Liu et al., 2021; Bond-Taylor et al., 2021).

These results triggered new questions about the role of perturbed and adversarial dreaming in cortical representation learning. What if, instead of being fully deprived from external supervision, the model agent could partially access the object category of sensory inputs during Wake? Indeed, human infants also receive sparse teaching signals throughout their development, such as when parents indicate the name of an object, or forbid them from eating certain food. We hypothesize that these signals supposedly tune cortical representations toward a better semantic separation. Such a learning regime is reminiscent of semi-supervised learning (van Engelen and Hoos, 2020), where an artificial network is trained in an unsupervised way on the whole dataset while trained in a supervised way on a fixed subset of labeled examples. In this setting, the unsupervised objective aims to improve the performance of the model over solely training it with sparse supervised data.

4.2 Methods

We here evaluate the benefits of our unsupervised learning phases (Wake, NREM and REM) in a semi-supervised learning regime. To this aim, we add to the architecture of the PAD model a linear projection of the z layer to a classifier output of 10 units c, where the class category can be learned (Fig. 4.1a). Note that we previously used such a linear read-out for evaluating z representations (Fig. 3.4), except that here we use it to train the entire E network. To train the encoder on supervised data, we create a labeled dataset D_y of N images as a subset of the entire training set D. D is used to train the classical PAD model in an unsupervised manner (see Chapter 3). In addition, at each Wake phase, we randomly pick a mini-batch from D_y to train the encoder on the classification task by providing class information to the output c. The number of labeled images N in the dataset D_y can vary from the full dataset size (50000 for CIFAR-10 and 73732 for SVHN), to 10000, 5000, 1000, 100 or 0 which brings us back to the unsupervised case (Deperrois et al., 2022). We thus add a loss term $\mathcal{L}_{\text{class}}$ to the loss function in Eq. 3.1:

$$\mathcal{L}_{\text{Wake}} = \mathcal{L}_{\text{img}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{real}} + \mathcal{L}_{\text{class}} \qquad (4.1)$$

$$\mathcal{L}_{\text{class}} = \frac{1}{b} \sum_{i=1}^{b} \sum_{k=0}^{9} y_k^{(i)} \log(c_k^{(i)}) \,, \qquad (4.2)$$
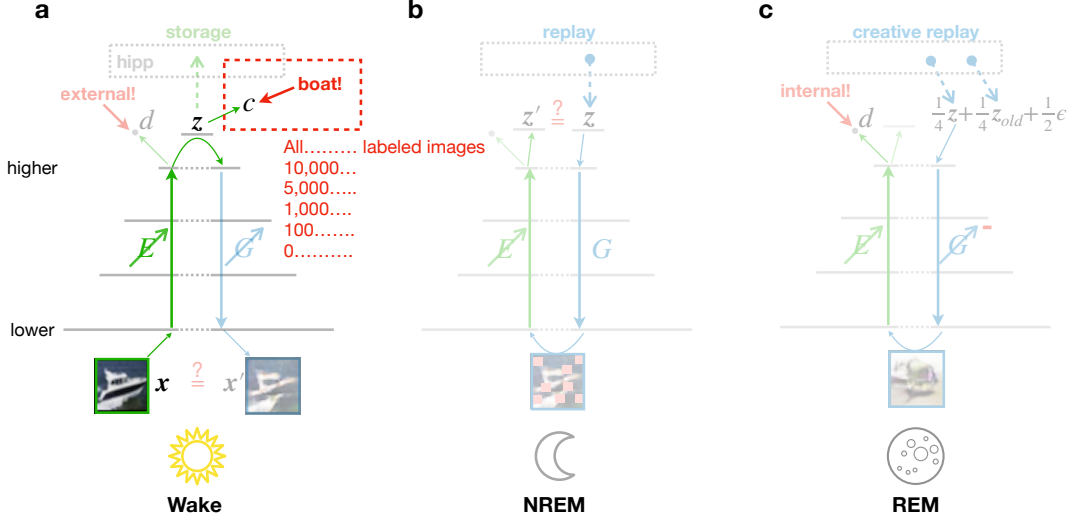
FIGURE 4.1: **Learning in PAD with different levels of supervision (a)** During Wake, in addition to the unsupervised objectives, a classifier output c on top of z learns to predict the correct class of the observed input x using the explicit label information. The number of labeled images provided to this output can vary from 0 (Deperrois et al., 2022), 100, 1,000, 5,000, 10,000 to the full dataset. Remaining features of the model are further detailed in Fig. 4.1.

where $y_k^{(i)} = 1$ if the ith image of the minibatch from D_y is of label k and $y_k^{(i)} = 0$ otherwise, $c_k^{(i)}$ is the Softmax probability for the kth class of the ith image provided by the output layer c. The remaining objectives of each phase remain the same as in Algorithm 1.

Considering this extension, we would like to evaluate the quality of learned representations for full and pathological models in a semi-supervised regime. The evaluation procedure is identical to Fig. 3.4 and Fig. 3.5, where we freeze E and G and train a linear classifier on-top of learned representations z with the full training dataset (Fig. 4.2a) and report the linear classification accuracy on the test dataset (linear separability).

The model can be trained under different conditions (Fig. 4.2b): full model (PAD), if REM is removed from training (w/o REM), if NREM is removed (w/o NREM), if both NREM and REM are removed (Wake only), or without any unsupervised learning objectives (None). All these models are trained with different amount of labeled images during Wake, going from zero to all labeled images.

4.3 Results

We then report the linear separability of learned representations at the end of training (epoch 50) for each model and different number of labeled examples for CIFAR-10 (Fig. 4.2c) and SVHN (Fig. 4.2d) datasets. We additionally provide a supervised upper bound (red cross) where the encoder network is trained on a full labeled dataset with additional drop-out regularization (Srivastava et al., 2014) to prevent overfitting on the training set, reaching test accuracies of approximately 75 % for CIFAR-10 and 93 % for SVHN.

For the model only trained on classification (Fig. 4.2c-d, red lines, "None"), the performance on the full dataset is slightly smaller than the supervised upper bound (due to the absence of drop-out regularization) where all labeled images are provided. However, by decreasing the number of labeled images, linear separability decreases rapidly.

Adding the Wake phase increases the performance for low levels of supervision ($N < 5000$), which shows that reconstruction objectives from explicit models such as our derived AE objective are beneficial for semi-supervised learning with little data. Adding NREM leads to little or no improvement of model performance (Fig. 4.2c-d, pink lines), probably because it only improves its performance on perturbed data but does not further extract semantic information from raw data.

However, adding REM adversarial dreaming (Fig. 4.2c-d, orange and black lines) significantly improves linear separability of learned representations for almost all levels of supervision. Indeed, the performance under the PAD is still high in a low supervision regime. The relative increase from higher supervision levels is then smaller than for other ablated cases.

4.4   Discussion

By extending our PAD model to learning from supervised signals, our results show that learning only from the available teaching signals or with the reconstruction of external inputs during Wake fails to produce semantically organized representations in a weakly supervised regime. However, in presence of REM adversarial dreaming, latent representations become better organized especially when only sparse teaching signals are provided throughout development. As the number of teaching signals increase, the importance of REM dreaming in learning semantic representations diminishes, as the brain would fully benefit from teaching signals and does not need to discover underlying structure through unsupervised mechanisms. However, as stated above, such a high amount of supervision (e.g., from 1000 labeled data in Fig. 4.2) is unrealistic in biological development (Bergelson and Swingley, 2012; Slone and Johnson, 2015; Lindsay, 2021). Animals essentially learn statistic regularities of sensory inputs without being systematically taught to do so and might generalize over a few teaching signals that the environment can provide. Creative dreams such as during REM sleep would then help them to enrich their representations over other object configurations that were not observed or explicitly taught while they were awake.

Our results are in line with the reported benefits of GAN learning in a semi-supervised learning regime (Salimans et al., 2016; Odena, 2016; Dai et al., 2017). In these models, providing a classifier network the ability to discriminate between real and generated data significantly improves the classifier performance in a weakly supervised regime. We recognize that the reported performance are usually higher in these settings, however the model architectures implemented are much deeper and additional ingredients than purely GAN learning were used to improve model performance, such as feature matching (Salimans et al., 2016) or the design of a "bad" generator (Dai et al., 2017). As in Deperrois et al. (2022), our purpose mainly consists of highlighting potential benefits of offline states over purely wake-driven learning using biologically consistent architectures and learning objectives, and a competitive improvement of performance against state-of-the-art models is beyond the scope of this project.
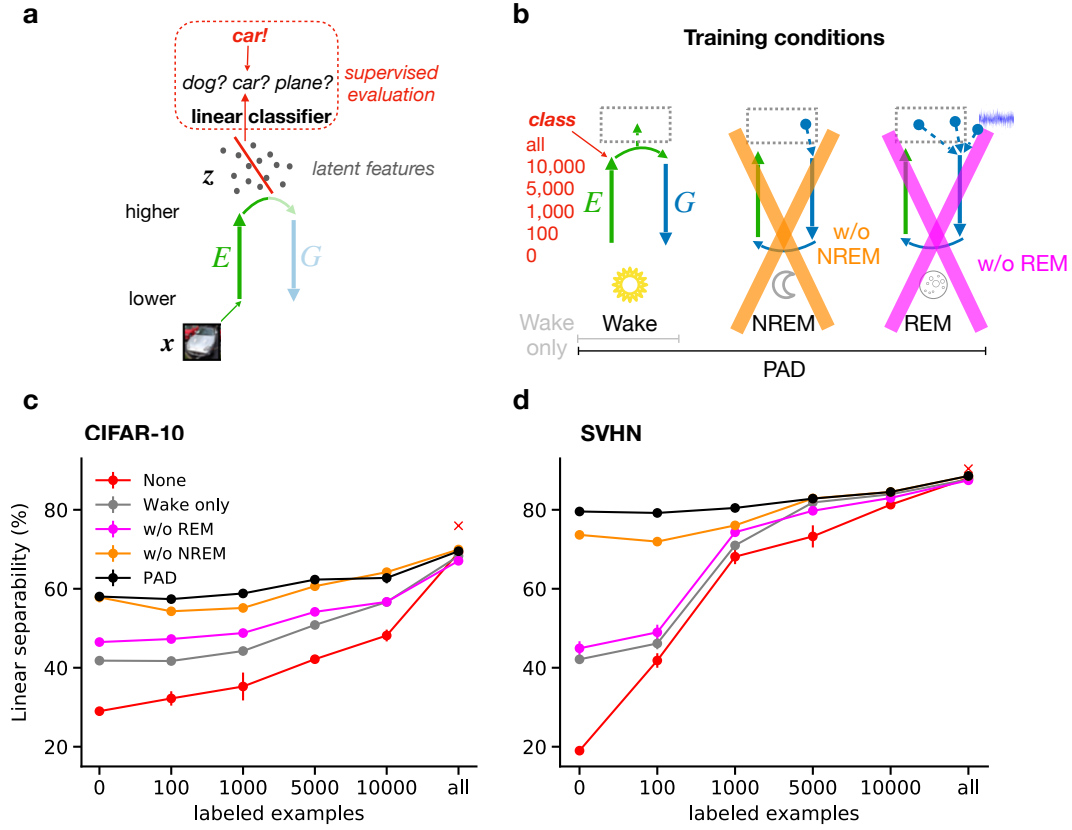
FIGURE 4.2: **Effects of NREM and REM on latent representations in a semi-supervised regime** **(a)** A linear classifier is trained on the latent representations z inferred from an external input x to predict its associated label (here, the category 'car'). **(b)** Training phases and pathological conditions: full model (PAD, black), no REM phase (pink), no NREM phase (orange), Wake only (grey), None (red). **(c, d)** Classification accuracy obtained on test datasets (c: CIFAR-10; d: SVHN) after training the linear classifier to convergence on the latent space z at the final epoch of E-G-network learning for different amount of labeled data. Full model (PAD): black line; without REM: pink line; without NREM: orange line; Wake only: grey line; None: red line. Red cross indicates a supervised upper bound with (None + drop-out regularization). Error bars indicate ± 1 SEM over 2 different initial conditions.

Chapter 5

Discussion

Why does the eye see a thing more clearly in dreams than the imagination when awake?

— Leonardo da Vinci

In this work, we proposed through an AI-inspired model that dreams facilitate the extraction of semantic information from sensory inputs with little or no supervision. Here, we extend the discussion from Chapter 3 to provide further analysis of the model and discuss potential extensions. First, we briefly summarize the main results from our findings.

5.1 Main results

The idea that sleep and dreams contribute to learning has been around for decades. However, previously it was unclear what exactly the brain learns during these offline states, and which mechanisms contribute to these processes. Here, we hypothesized that dreams participate in learning cortical representations with little or no supervision.

While cognitive theories characterize memory semantization, a form of unsupervised learning process, by the simple replay of waking memories during NREM sleep (Nadel and Moscovitch, 1997; Lewis and Durrant, 2011; Winocur et al., 2010; Dudai et al., 2015), our study emphasizes the creative aspect of dreams. Indeed, a central result from our work is that REM dreams, by mixing several hippocampal memories and, together with wakefulness, hosting an adversarial game between cortical forward and backward pathways, facilitate the extraction of semantic information from sensory inputs. Notably, we show that both combination of memories and adversarial learning are required to separate object categories within high-level network features.

We still however consider the importance of memory replay as we suggest from our results that it could improve the robustness of cortical representations. This effect is obtained if the generated NREM dreams are augmented with sensory perturbations, and if only a single memory triggers these dreams.

Finally, by providing our learning system increasingly sparse supervised signals, our results demonstrate that REM adversarial dreaming is mainly beneficial in a weakly supervised regime, within which humans and animals usually learn.

Together, our results suggest a computational role for NREM and REM dreams in mammalian learning and draw new lines of experimental investigations for future cognitive and neurobiological studies.

5.2 Representation learning and the brain

In this work, we emphasize that animal learning is mostly unsupervised, and that the brain might be endowed with unsupervised learning objectives to construct semantic representations. Here, we further analyze the different learning objectives that we used to propose an unsupervised learning paradigm during Wake, NREM and REM sleep, and reflect them in light of neuroscientific and machine learning literature.

5.2.1 Explicit generative learning by predicting sensory inputs

In Section 1.3.2, we introduced the view of the brain as a generative model to perform unsupervised learning from sensory data. In particular, we explained that the most popular generative models, such as hierarchical predictive coding (Rao and Ballard, 1999) or Wake-Sleep (Hinton et al., 1995), are explicit, as they aim to maximize the likelihood of sensory data by learning to predict sensory inputs via element-wise reconstruction errors (e.g., Eq. 1.18, Eq. 1.26).

In our model, this approach was confined to the Wake phase. Indeed, the Eq. 3.2 corresponds to a reconstruction error between the input x and a top-down prediction $G(E_z(x))$ typically used by autoencoder models to learn a compressed representation $E_z(x) = z$ of input data. However, by adding the KL-loss to a Gaussian unit distribution (Eq. 3.3), and some additional Gaussian noise ϵ (Eq. 3.12), our Wake objective resembles the VAE objective (Eq. 1.26). The choice for these additional regularizations are justified by the observation that the autoencoder objective alone is not sufficient to extract relevant high-level information from data, as it might end up learning the identity function when given enough capacity in the generator network, even if the latent space is low-dimensional (Goodfellow et al., 2016). Moreover, the KL-divergence loss forces latent representations to be confined to a fixed (prior) distribution, which in turn force convex combinations during REM not to be too sparse, necessary to train GANs (Goodfellow et al., 2014).

However, the Wake phase remains different from VAEs as the encoder only predicts the mean of the posterior distribution (and not the variance) to which a constant noise is applied. Thus, this phase can be seen as a VAE with a fixed variance as posterior distribution.

Besides its computational advantages, this objective is in line with previous theories of Bayesian computations (Rao and Ballard, 1999; Clark, 2013) as it characterizes the brain updating its expectations about a hidden state of the world by predicting the external input through a hierarchical generative network with an element-wise objective.

Furthermore, it allows the hippocampus to store a reliable compressed, low-dimensional representation z without having to store the high-dimensional image input x. Indeed, without this reconstruction objective, the generated dream $G(z)$ during the subsequent NREM phase would not correspond to the previously observed input, and thus the latent reconstruction objective during NREM (Eq. 3.6) would not improve the robustness of latent representations, and would even impair learning. Furthermore,

this objective allows to keep the cycle-consistency between the data and latent spaces, especially since the generator is systematically updated via adversarial learning during REM, and thus allowing the encoder network to invert the generator (see below Section 5.2.2).

5.2.2 Adversarial generative learning by inventing sensory inputs

Even though our Wake phase was inspired from explicit generative models, in order to generate new, realistic inputs despite the absence of external inputs, we took inspiration from the implicit generative adversarial objective that only relies on the discriminator teaching signal, and not on element-wise reconstruction objectives (Goodfellow, 2016). This is further illustrated by the observation that REM adversarial learning make the generator synthesize creative samples from random latent activity, while a model without REM only produces blurry samples reflecting images from the dataset, learned from the element-wise objective used during Wake (Appendix A.1.1). We argued, throughout our study, that the adversarial objective could constitute a potential explanation for bizarre, creative but realistic dreams occurring during the REM state.

The use of an adversarial objective in our model was also motivated by previous work reporting that GANs and their variants performed better at representation learning than other generative models (Radford et al., 2015; Brock et al., 2017; Dumoulin et al., 2017; Beckham et al., 2019b; Donahue and Simonyan, 2019). Our results confirm that the GANs objective, given our architecture, performs better than the pure autoencoding objective learned during Wake (Fig. 3.4, Fig. 4.2).

Initially, the representation learning performance of GANs was directly evaluated on the discriminator features, which achieved relatively good performance on CIFAR-10 and SVHN datasets (Radford et al., 2015). These results, as discussed in Section 4.4, inspired subsequent models to combine classifier and discriminator objectives in a semi-supervised regime (Odena, 2016; Dai et al., 2017; Salimans et al., 2016). While the success of the discriminator as a feature extractor is still difficult to explain, recent theoretical work argues that this effect cannot be attributed to the discriminator's objective itself, but to the need to prevent the entire GANs from mode collapse during training (Mao et al., 2020).

Other works revealed that the features learned from an encoder trained to invert the GANs generator also achieved equal or better performance (Donahue et al., 2016; Dumoulin et al., 2017). This consists of designing an encoder that predicts the latent activity z that would generate a given image x through G. Different methods were explored to perform this task, for instance by learning a couple encoder-generator adversarially against a discriminator (Dumoulin et al., 2017; Donahue et al., 2016; Donahue and Simonyan, 2019), by augmenting autoencoders with an adversarial objective on combinations of representations (Berthelot et al., 2018; Beckham et al., 2019b; Ulyanov et al., 2017), or by combining VAEs and GANs objectives (Rosca et al., 2017). Learning this encoder corresponds to learning the approximate posterior $q_\phi(z|x)$ in a similar manner as VAEs and other generative models, that would in principle infer high-level properties of the observed data, considering that the GAN latent space $p(z)$ is well structured and interpolations within it produce smooth and realistic variations in the data space. However, this configuration requires to train three networks: a generator and a discriminator to perform GAN-learning, and an additional encoder, which is overall costly in terms of computational ressources and less likely matches with cortical structure (Gilbert and Li, 2013; DiCarlo et al., 2012).

In our model, similarly as previous work (Brock et al., 2017), the feedforward pathway acts both as an encoder and a discriminator. The adversarial objective run through the Wake (Eq. 3.5) and REM (Eq. 3.8) phases implements the discriminator function while the reconstruction objective (Eq. 3.2) combined with the KL-divergence loss to a Gaussian distribution (Eq. 3.3) during Wake favours the learning of an encoder inverting G.

In an additional experiment (Appendix A.1.2), we test our model with separate networks for discriminator and encoder functions. In this case, the encoder only implements Eq. 3.2 while the discriminator network only implements Eq. 3.5 and Eq. 3.8. We show that separately, learned features from encoder and discriminator perform worse than if both functions are combined, as in our PAD model. Thus, these results show that given our architecture, not only combining both encoding and discriminator functions save computational ressource, but provide an overall better network performance.

These results also provide a biological prediction about the organization of cortical feedforward pathways. Fundamentally, one cannot easily distinguish these different functional components (encoding and discriminating) along the ventral stream, as they are possibly shared within a single feedforward network (DiCarlo et al., 2012). Here, given our network architecture, we demonstrate that this biological constraint is advantageous.

5.2.3 Contrastive learning by comparing sensory inputs

Following the classical formalism of the brain as a generative model (Rao and Ballard, 1999; Friston, 2010; Marino, 2022) and their ability to learn good representations, we previously emphasized the importance of generative models as a way to alleviate the need for supervised labels (Section 1.3.2, Section 5.2.1, Section 5.2.2). However, the recent success of contrastive learning objectives in representation learning (Jaiswal et al., 2020) could provide further insight on biological learning. Although mentioned in our paper discussion Section 3.4.3, we here further explore this concept and the possibility of extending our model with such objectives.

Generative approaches with latent variable models rely on the idea that the training data originates from an underlying, physical generative process. By capturing this generative process through a latent prior distribution $p(z)$ and a likelihood $p_\theta(x|z)$, latent variables z might reveal an effective description about the type of sensory input $G(z)$ they generate, and thus store useful, semantic information (i.e., object shape, animal or object, gender, etc.). However, modeling the training data distribution can be computationally expensive and may not be always necessary for representation learning (Le-Khac et al., 2020).

In comparison, contrastive learning algorithms use one or several feedforward encoders to learn representations of data and do not learn the data distribution with a generator. In a nutshell, an encoder is trained to compare samples between each other, either by pulling together similar inputs or pushing apart dissimilar inputs (Jaiswal et al., 2020; Le-Khac et al., 2020). Similar (positive) examples are usually obtained by applying a series of data augmentations such as cropping, resizing, blur, color distorsion to a given sample (Chen et al., 2020), and negative examples are simply different samples from the dataset. This comparison can be learned with a loss function $l_{i,j}$ defined on a positive pair (i,j) and a large number of negative pairs
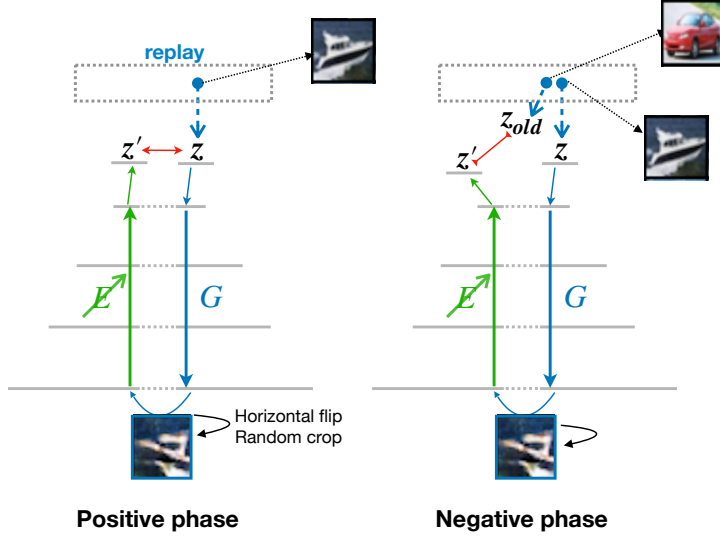
FIGURE 5.1: **Potential extensions of NREM to contrastive learning.** Future work could further investigate if our NREM phase can improve latent representations of the PAD. One direction would be to identify which augmentations (e.g., horizontal flip, random crop) leads to this improvement (Chen et al., 2020). Another one would consist of storing "negative" examples in the hippocampus and replaying them along with the current memories to learn the denominator of the contrastive learning objective (Eq. 5.1).

$(i, k)_{k \neq i}$ such as:

$$l_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}, \tag{5.1}$$

where $\text{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^T \boldsymbol{v}/\|\boldsymbol{u}\| \|\boldsymbol{v}\|$ denotes the dot product between l_2 normalized $\boldsymbol{u}$ and $\boldsymbol{v}$ and τ denotes the temperature parameter (Chen et al., 2020). This measures the cosine of the angle between two vectors, which is a better effective measure than the euclidian distance as it is scale invariant. Through this learning objective, the network aims to reduce the distance between the representations of positive pairs $(\boldsymbol{z}_i, \boldsymbol{z}_j)$ and increase the distance between the representations of negative pairs $(\boldsymbol{z}_i, \boldsymbol{z}_k)_{k \neq i}$.

Recent work has shown that neural networks learned with contrastive learning methods predict recordings of visual cortex activity as well or better than networks learned with full supervision (Zhuang et al., 2021; Konkle and Alvarez, 2022). From these results, authors proposed that such networks could constitute strong candidates for primate sensory learning. We here investigate, within our sleep framework, the biological mechanisms that could implement such objectives.

As mentioned in Section 3.4.3, our NREM phase is somewhat reminiscent of contrastive learning algorithms as it tries to pull the representation $\boldsymbol{z}'$ of an occluded, replayed input $\boldsymbol{x}'$ towards the representation $\boldsymbol{z}$ of the original input $\boldsymbol{x}$ observed during the previous Wake phase. However, this phase presents some major differences that need to be addressed to turn it into a contrastive learning objective.

First, in our framework, the "positive" example $\boldsymbol{x}'$ is not an augmentation of the input $\boldsymbol{x}$ itself, but of a reconstructed, blurry version from the stored $\boldsymbol{z}$ in the hippocampus ("dream augmentation"). If this reconstruction is good enough, the semantic content of this example should be unaltered, and our model would not need to store the entire data image $\boldsymbol{x}$ to obtain its augmented version.

Second, the applied occlusions might not be optimal for contrastive learning objectives, as previous work showed that a specific type and series of augmentations are required (e.g., random cropping and color distorsion, Chen et al., 2020) that could also be biologically plausible (e.g., activation of a subset of V1 neurons for random cropping). Future work could explore these other types of augmentations would turn this phase beneficial for learning semantic representations.

Third, our NREM phase does not provide the negative examples necessary to compute Eq. 5.1's denominator. We however argue that by storing representations from past days (mini-batches) in the hippocampus, the negative examples can be provided along the positive (current day) example (Fig. 5.1, right). This idea was proven successful in recent contastive learning works that use a memory bank to store and retrieve representations of negative samples (Misra and van der Maaten, 2020; He et al., 2020). Even though negative examples are an efficient way to prevent the encoder from learning to map any input to a constant representation (collapsing representations), recent algorithms avoid requiring them. For instance, one can use an additional *target* encoder, updated through a running average of the first *online* encoder's weights, to provide the positive examples (Grill et al., 2020). Another possibility is to maintain the variance of each embedding dimension above a certain threshold (Bardes et al., 2021).

5.3   Dreaming and the brain

5.3.1   A hypothesis for dream generation

The origin of dreams is still quite unclear (see Section 1.4.4.1). A main message from our model is that the generation of dreams could be explained the activation of feedback pathways implementing a generative model of the sensorium. We proposed that this generative model is learned via reconstruction of external inputs while awake and via adversarial learning against a feedforward discriminator during Wake and REM sleep. In parallel, the feedforward encoder (and discriminator) learns to invert the feedback generator to ensure cycle consistency between low and high-level activities (discussed in Section 5.2.1).

If our hypothesis is correct, early sensory activity could provide a window into the content of dreams. For example, if an external input (e.g., a cat image) triggers a certain pattern of early sensory activity during wakefulness, and that a similar activity is observed during a subsequent dream, it is likely that the content from this dream reflects this input (i.e., a cat is present in the dream). Interestingly, this is in line with a functional magnetic resonance imaging (fMRI) study from Horikawa et al. (2013). In this study, diverse images were initially presented to participants and the elicited cortical activity was simultaneously recorded. Then, linear classifiers were trained on the recorded recorded activities based on the image categories (faces, car, doors, etc.). When these classifiers were matched to the participants' cortical activity recorded during a dreaming state (here, hypnagogic state), researchers found a striking agreement between the classifier predictions and the content from the corresponding dream reports. In other words, the cortical activity patterns representing images in our dreams are created by reactivating patterns elicited when similar images were observed while awake.

Our model reflects this observation as reactivating latent activities stored from Wake triggers NREM dreams that contain the elements from the most recent waking sensory experience. By mixing several hippocampal memories with additional noise, REM dreams contain elements from the nearest neighbouring latent activities associated with particular waking sensory inputs. Through this phase-specific description, our approach reconciles dreaming theories that either claim that dreams originate from the replay of stored memories (Wamsley, 2014; Wamsley and Stickgold, 2019) or instead emphasize that dreams are not replaying a specific memory, but rather combine different elements from non-related memories (Fosse et al., 2003; Schwartz, 2003).

Finally, our proposal for REM dream generation also reflects the initial activation-synthesis theory from Hobson and McCarley (1977) that claims that REM dreams result from the brain "making the best of a bad job in producing even partially coherent dream imagery from the relatively noisy signals sent up to it from the brain stem". In that sense, our model characterizes the noisy signal during REM sleep through the random combinations in the latent space, and the brain trying to make sense of it through the adversarial learning process, increasing the realism of this randomly initiated sensory experience.

5.3.2 A hypothesis for dream function

The existing theories for dream function (e.g., creativity, emotional processing, generalization, etc., see Section 1.4.4.3 for details) are quite diverse and seem to differ from our hypothesised function, that is the improvement of cortical representations. This difference arguably resides in the computational nature of our work that tends to shape our hypothesis towards the modeling choices, such as the datasets (here, images), architecture (here, convolutional networks) and evaluation metrics (e.g., linear separability).

Although different, some aspects of the existing theories for sleep and dreams functions are still supported by our work. We argued in Section 3.4.1 that our ideas could be associated with the concept of memory semantization from transformation theories (Nadel and Moscovitch, 1997; Winocur et al., 2010; Lewis and Durrant, 2011) (Section 1.4.3.2). However, in cognitive science, memory semantization mainly refers to a transformation of highly detailed, contextualized hippocampal episodic memories into abstracted, decontextualized cortical memories, while in our model, our hippocampal "episodic" memories are simple copies of cortical representations. The semantic "transformation" in fact occurs within cortical networks through the gradual adaptation of cortical synapses with learning.

As discussed in Section 3.4.5, our hypothesis also reflects the possibility that dreams improve creativity (Hobson, 2009; Lewis et al., 2018) considering that random memory combinations are explored, however, there is, to our knowledge, no metric that could evaluate creativity in neural networks. It also relates to the proposed role of dreams in enhancing generalization (Hobson et al., 2014; Hoel, 2021), as in presence of REM dreams, our model still maintains a good linear separability of latent representations for test (never encountered) data even when little or no supervised labels are provided as compared to a model without REM (Chapter 4).

5.3.3   Feedback pathways beyond dreaming

In this work, we mainly emphasize on the role of cortical feedback pathways in generating virtual sensory experiences during dreaming or reproducing sensory inputs during wakefulness. We however suppose that mental imagery would employ the same mechanisms as dreams (Pearson, 2019) while awake, i.e., via cortical feedback pathways initiated by random latent activities. However, due to the occurrence of mental imagery in the awake state, the perception of external inputs might interfere with these imagined activities. We hence argue that REM sleep constitutes the ideal stage to generate virtual experiences as the brain is fully disconnected from external inputs (Hobson et al., 2014; Llewellyn, 2016a).

Finally, we assume that the previously proposed roles for feedback pathways (spatial attention, object expectation, McManus et al., 2011; Gilbert and Li, 2013; Moore and Zirnsak, 2017) could also be accounted in our model if feedforward signals carried by the encoder are merged with feedback signals from the generator, for example by convex combining them or multiplying them (gain modulation, Ferguson and Cardin, 2020). For instance, one could consider a certain context z_{context} elicited by previous inputs, that modulates an externally-driven low-level activity x could be modulated through an attention signal coming from higher areas $G(z_{\mathrm{context}})$ through multiplication of both activities, i.e.,

$$x_{\mathrm{modulated}} = x * G(z_{\mathrm{context}}) \tag{5.2}$$

This low-level activity would conceptually integrate top-down information obtained through G, and mimick the effects of top-down attention on visual perception by emphasizing which part of the input are relevant for a given context. Future work could investigate whether such mixing of bottom-up with top-down information could lead to a better representation $E(x_{\mathrm{modulated}})$, and then provide a link between dreaming and top-down attention. Previous computational studies already revealed that endowing a network with re-entrant feedback connections, resulting in recurrent neuronal dynamics, improved network performance over a fully feedforward network on occluded images (Spoerer et al., 2017; Tang et al., 2018). However, in these studies feedback connections do not learn a generative model of sensory inputs but only effectively increase the capacity of the feedforward network.

5.4   Outlook

5.4.1   Suggested experiments in humans

Our model aims to describe how semantic cortical representations are acquired throughout development. In order to test our predictions, experimental studies could record the stimulus-evoked activity in high-cortical areas (Grill-Spector et al., 2001; Hung et al., 2005) over a long period and evaluate their linear separability according to object categories, in presence of absence of NREM and REM sleep. We previously mentioned possible lines of experimental investigations to test our hypothesis (see Section 3.4.4). We here suggest an additional, simple and short experimental investigation that could be performed on human subjects and highlight the results from our model.

First, to keep the study on a relatively short time scale, subjects could be presented with novel objects, never encountered before, such as abstract 3D objects composed of multiple cubes (Tarr, 1995). The time required to acquire representations about these objects would be supposedly shorter than the time required to construct general object representations (see Section 1.2.1).

Second, recording single-unit neuronal activity as done in Hung et al. (2005) is highly invasive and cannot be performed in humans. On the other hand, only assessing classification accuracy from human participants is a fairly coarse-grained method that only allows to assess if, but not how object representations change over the course of learning (Geirhos et al., 2020). In psychophysics, object representations are assumed to be embedded in an internal multidimensional space (Ashby, 2014; Richler and Palmeri, 2014) where distances in this space correspond to perceived similarities. To compute this embedding space, participants are first asked to assess similarity between objects (from the same or different categories). This data is then analyzed with a recently developed machine learning embedding algorithm that estimates the dimensions of the embedding space, as well as the coordinates of these object representations in that space (Terada and Luxburg, 2014; Haghiri et al., 2020). We suggest that by tracking similarity judgements from human participants over the training period, one could investigate the representational changes within this space, which would serve as a cognitive approximation of high-level cortical representations.

Third, evaluating the effects of dreaming on object representation learning requires to deprive subjects from dreaming when asleep. This procedure is experimentally challenging as one cannot attest whether a subject is dreaming or not, neither what is dreaming about (discussed in Section 1.4.4.1), even when completely ignoring all ethical issues with trying to prevent humans from dreaming. While some anti-depressant drugs affect REM sleep (Boyce et al., 2017), dreaming still occurs in these subjects (Oudiette et al., 2012). Mental imagery, in contrast, implies that the generation of internal percepts is voluntarily triggered and its content is relatively controlled (Pearson, 2019), making it more exploitable for testing the effects of internally generated experiences on learning. As discussed above (Section 5.3.3), considering that mental imagery shares the same neuronal substrates as dreaming, we suggest that within our proposed novel object recognition task (previous paragraph), human subjects could be asked to perform mental imagery training sessions following the presentation of novel objects. In parallel, another control group of subjects would not perform the mental imagery task after being presented with novel objects. From our theory, we expect that the representation from subjects performing these imagery sessions more linearly separable according to the novel object categories.

5.4.2 Dreaming for the future?

By inventing new sensory inputs from new latent activities, our model follows to some extent Llewellyn (2016b) ideas that distinguish "predictive coding", attributed to waking perception where the brain anticipates upcoming inputs by identifying their latent causes, from "prospective coding" where the brain creates prospective codes oriented toward future situations. Note that however, our model does not capture the possibility that these prospective codes serve future situations, but only suggest that they reorganize representations such that causes underlying sensory experiences are better inferred in the future.
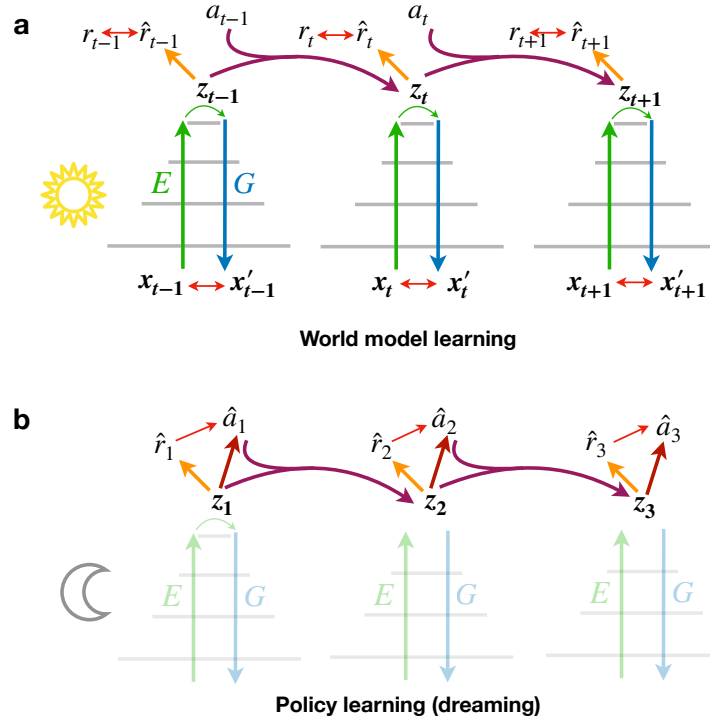
FIGURE 5.2: **Learning behaviors by dreaming of a world model.** **(a)** From a dataset of observations x_t associated with a rewards r_t, the world model learns a transition model $q(z_t|z_{t-1}, a_{t-1})$ and a reward model $q(r_t|z_t)$. **(b)** During "dreaming", the agent optimizes its policy by learning which actions maximize the imagined rewards $\hat{r}$ based on the learned reward model $q(r_t|z_t)$ and transition model $q(z_t|z_{t-1}, a_{t-1})$. This learning can directly be performed on the latent space dynamics. Figure adapted from Hafner et al. (2019).

As previously discussed (Section 1.4.4.3), the idea that dreams prepare us for future situations is more anecdotic (Mazzarello, 2000) than factual, but we expect that future AI-inspired models of the brain could capture this property, for instance considering model-based reinforcement learning (Ha and Schmidhuber, 2018a; Moerland et al., 2020; Hafner et al., 2019, 2020). In this setting, an agent learns a model of the environment and optimizes its policy, i.e., which actions to take in which situation in order to maximize the cumulative rewards, based on the simulations of this internal model. In addition to a feedfoward encoder inferring the latent state z_t from the observation x_t, these models predict next state of the environment z_t from the previous state z_{t-1} and action a_{t-1} (transition model $q(z_t|z_{t-1}, a_{t-1})$, purple, Fig. A.1a). The behavior can then be directly learned from the imagined latent dynamics and rewards derived from this world model in absence of further input (Fig. A.1b). Back to the actual environment, the agent transfers the policy learned from the imagined states (Ha and Schmidhuber, 2018a; Hafner et al., 2019, 2020).

These models can thus provide insights on how an agent can prepare itself for future situations by simulating such experiences while dreaming. However, latent dynamics are usually learned by recurrent architectures that require backpropagation-through-time (Werbos, 1990) which biological plausibility remains to be elucidated (but see Bellec et al., 2020).

Note also that in these models, "dreaming" is viewed as a state where behavior is optimized, but not the world model itself, in contrast to the PAD model. Future work in model-based RL could consider updating the world model in addition to the

behavior model during these offline phases, for example through adversarial learning. We expect that the resulting disentangled state representations could facilitate the learning of latent dynamics and therefore the behavior of the agent. Through this lens, it would highlight the importance of learning structured cortical representations during sleep, beyond the separation between object categories.

5.5 Conclusion

In this thesis, we explored the hypothesis that sleep promotes learning of semantic representations by proposing that cortical networks implement a model of the sensorium that gets differently reactivated and optimized during NREM and REM sleep. Our results suggest that dreaming, by extending the domain of possible observations, can improve the quality of cortical representations and thus participate to sensory development. While the proposed underlying mechanisms are inspired from deep generative modeling, the originality of this work resides in their application to a widely debated biological phenomenon, taking a substantial consideration of cortical structure, existing cognitive theories and experimental data. In this final chapter, we reinterpreted our model in light of state-of-the art machine learning algorithms, described the new hypothesis for dreaming in the brain, and suggested new lines of modeling and experimental studies. We hope that this work, beyond its results, will serve as a source of inspiration for future research on sleep and biological representation learning.

Appendix A

Supplementary information

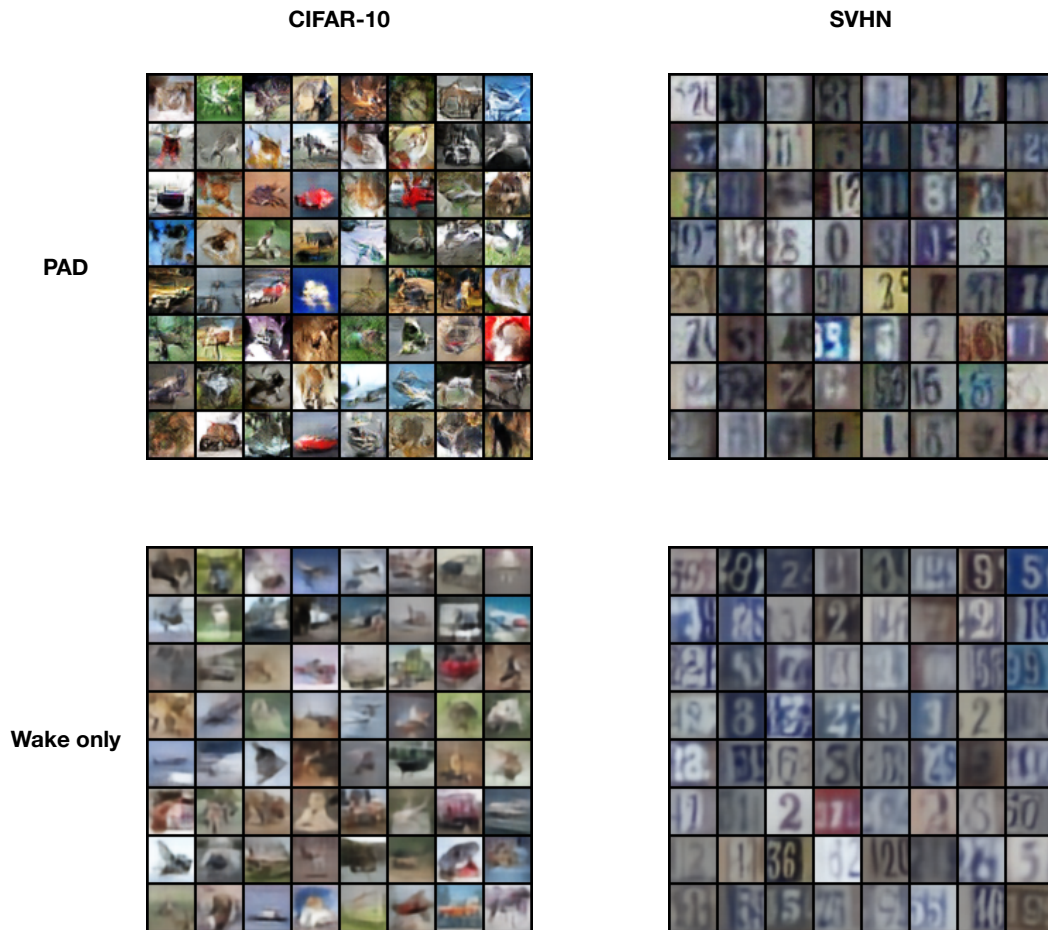## A.1 Additional results

A.1.1 Generated samples from PAD

FIGURE A.1: **Samples generated from PAD in presence or absence of REM adversarial learning.**

We here display additional samples from Fig. 3.3 obtained from the REM phase, i.e., by feeding random memory combinations with additional noise into the generator network of the PAD model at the end of learning. As a baseline, we show the samples
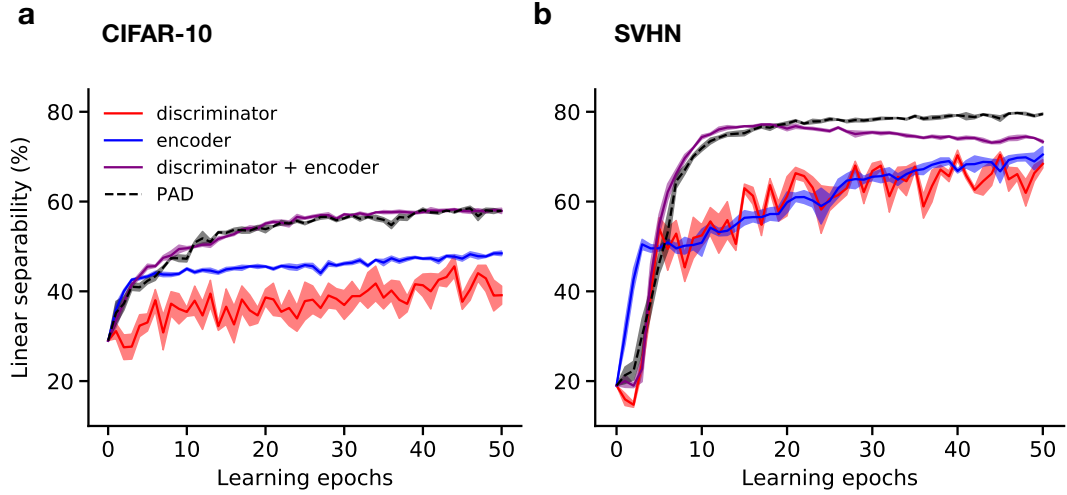
FIGURE A.2: **Effect of combining encoder and discriminator functions into one single network on latent representations.** Latent representation linear separability obtained if encoder and discriminator functions are implemented in separate networks (blue and red) or if both functions are performed by the same network (purple). Note that in all these conditions, the NREM phase was not included in training for better analysis of individual effects. Solid lines represent mean and shaded areas indicate ± 1 SEM over 4 different initial conditions.

obtained through this process in the "Wake-only" condition, where REM adversarial learning is not implemented.

We observe that in the Wake-only condition (VAE objective), generated samples are blurry and pale and hardly differ from images of the dataset. The visible "objects" observed in this condition are due to the reconstruction objective of the Wake phase that trains the generative network to reproduce images element-by-element. The resulting image seems to either reproduce images from the dataset or average two images in the pixel space.

In the PAD model, generated samples are much sharper and crispier than in the Wake-only condition. They however relate much less with specific images from the dataset, but rather combine features from different objects (e.g., shape of car with texture of a dog, unobserved digits), which partly explain their bizarre aspects, also characteristic of REM dreams. This shows that adversarial learning helped the generator to synthesize realistic images from latent representations that were not associated with a particular image during the Wake phase.

A.1.2 Effect of discriminator and encoder learning on latent features

In Fig. A.2, we show the latent representations linear separability if the encoder and discriminator functions are implemented in separated networks (blue and red) or if both functions are performed by the same network (purple). As a baseline, we also show the results obtained by adding the NREM phase (PAD, black), that tends to help the encoder to further invert the generator network due to its latent reconstruction objective.

A.2 Questions & Answers

A.2.1 VeryWell Health Q&A

The website VeryWell Health[1] was interested in communicating the findings from our publication. Jocelyn Solis-Moreira provided us a series of questions and wrote an article[2] based on our answers that I report here.

What made you interested in exploring the role of dreams in learning?

Dreams are an interesting phenomenon. If you're one of the lucky people who remembers their dreams, try to recall your most recent one. What about it stood out to you? Were you reliving a specific experience from your previous day, or rather taking part in a crazy movie? There is a high chance that you experienced the latter. So I would wonder: What can this be possibly good for? It's unlikely that this imagined experience prepares me for what exactly is going to happen tomorrow. Still, dreams are a consistent phenomenon of human sleep. Unfortunately, so far it was unclear what purpose dreams could serve. To tackle this mystery, we took inspiration from theoretical principles of artificial intelligence stating that generating virtual experiences is a way to learn about the structure of the world.

Were there any results you found surprising or did not expect to find?

After hypothesizing that generating virtual experiences, in the form of dreams, is a way for our brain to learn about the structure of the world, we needed to decide what serves as a basis for a single dream. In the simplest case, one could use a single memory to generate a dream. Alternatively, one could combine multiple, possibly unrelated, memories. What we found in our model is that it is indeed the combination of several memories that leads to learn better. This was not obvious from the start, but it provides a hint of why our dreams often creatively combine elements from different episodes of our lives.

People try to find a deeper meaning to their dreams, but according to your study, dreams may not "mean" anything but rather be used to organize the brain. Does this suggest dreaming has an evolutionary purpose?

I think it's important here to distinguish between two different interpretations of "meaning". In the common understanding, dreams are interpreted to have personal, often emotional meaning, such as telling you something about your future or your relationships. Our model does not capture this dimension of human experience and we can hence not draw any conclusions here. However, in an alternative interpretation, "meaning" could refer to the importance of dreams for brain function. In this sense, our model suggests that dreams do have "meaning", in line with previous dream theories such as the activation-synthesis hypothesis (Hobson). Intuitively, generating new but realistic virtual experiences requires the brain to learn a lot about the structure of our world, knowledge that is precious to navigate our environment while being awake. In other words, our model suggests that it is not important *what* you are dreaming, it's important *that* you are dreaming. Consequently we would also strongly agree that dreaming has an evolutionary purpose.

---

[1] https://www.verywellhealth.com
[2] https://www.verywellhealth.com/weird-dreams-process-experience-5324057

Does this mean people should prioritize sleep/try to sleep deeper to help their brain learn and get organized?

I think it is important to mention that we are not pursuing clinical, but basic research, hence I cannot give medical advice on sleep patterns. That being said, I would not be the first to point out the importance of good sleep for healthy brain function. It is well known that sleep serves many physiological needs, such as recovering our motor functions, removing waste products from our brain, and strengthening our memories. Our study extends this list by suggesting that dreams are important for learning about the structure of our world.

Sleep benefits everyone, but given that dreaming can serve as an active learning process, would you say it has major benefits for children and teens to retain what they learned in school?

I would like to repeat the same warning as before: since we are pursuing basic research, we can not give medical advice. Previous work has already shown that for the retention of specific memories, sleep is important, supporting your suggestion that healthy sleep is crucial during times where we accumulate a lot of new knowledge. Given the often bizarre nature of dreams, in combination with our model, I think dreams are not the main driver of retaining specific memories, but they are rather relevant for organizing our memories according to specific concepts.

During childhood development, we do not only retain specific episodes but learn general concepts about the world, which build the major structure of our brain: learning how to recognize objects, walk or speak. According to our model, dreaming would be a key contributor to learning these fundamentals. Furthermore, I would like to point out at any stage of adult life we keep learning new concepts and skills, and thus keep restructuring our brain. Our model suggests that dreams are involved in this restructuring. Indeed, it is not uncommon for our dreams to be occupied with a specific new skill which we are currently learning, for example when starting to practice playing an instrument. This is in line with research showing that REM sleep (in which creative dreams occur) remains present throughout adult life and occupies a significant fraction of the time spent asleep.

What is the most important thing our readers should know about your study? For example, should people be tracking their sleep health?

The main take-away from our study is that dreaming at night may be just as important for your brain as gathering new experiences during the day. Remember that for every two hours you spend awake, perceiving new information, you sleep one hour, with no information coming in. Dreams, due to their sensory isolation and their hallucinatory nature, might be an ideal stage to creatively re-process waking experiences and extract concepts and meaning from them

What are you working on next? For example, are you building on this study to look at dreaming in humans themselves instead of a computational model?

Indeed, together with our colleagues in the psychology department, we are currently designing an experimental paradigm to test the hypotheses from our model in human participants on a behavioral level. More specifically, we will test how the hypothesized dream-processes affect visual object representation. For example we will test

if participants better recognize unfamiliar objects after being exposed to related objects on the previous day. Our model suggests that the dreams occuring during REM sleep should lead to increased performance. Furthermore, we would like to investigate whether one can use "artificial dreams" delivered via virtual reality to replace or augment REM dreams. Such a replacement could be important since in certain pathologies, or due to medication, REM sleep may be reduced in patients.

Is there anything I haven't asked that you think is essential for readers to know about your study?

I think it is important for readers to understand that we are approaching the role of sleep from a computational rather than a clinical perspective. Nevertheless, our model makes specific predictions on the behavioral as well as neuronal level which can be investigated both with human participants and animal models. We thereby hope to contribute to our understanding of brain function and in particular to deciphering the mysterious role that dreams play in our lives.

A.2.2 Radio Q&A

Similarly, our study also raised interest of the Irish radio FM104[3] that came across the HBP press release. I here report some of the main questions that they posed us and the answers we provided.

How was the study conducted ?

We present a computational model of how the brain learns to represent visual inputs (in our case images) in the cortex. These representations can be used by other parts of the brain to perform tasks, for example to reach for an object with your hands. The study is based on the numerical simulation of the cortical neuronal networks during wakefulness and sleep Each of these phases present different objectives to train the synapses that connect the neurons of the network. During wakefulness, the visual inputs (images) are presented and stored in the artificial brain. During sleep, memories are reactivated and the network generates dreams from them. This generation of dream is inspired by artificial intelligence algorithms that were shown to be successful at generating new data (images, music, etc). We alternate wake and sleep phases to train our model, simulating learning over years of development, and evaluate how well the network organizes the visual representations over time.

The study offers a new theory on the significance of dreams using machine learning inspired methodology and brain simulation. Can you explain these different techniques and how they worked simultaneously?

These techniques were inspired from artificial intelligence, in particular "deep learning". They consist of learning complex tasks, such as recognizing an object in an image, convert spoken words into text, or in our interest here, generate data. These algorithms rely on "artificial neural networks" made of millions of units, often referred to as "artificial neurons", all connected in a specific architecture. The goal is to learn a specific task (recognizing a cat in an image of cat, or generating the image of the cat), by adapting all these connections. Due to the abundance of units and connections in these networks, and the computational resources that allow us to

---

[3]https://www.fm104.ie

train them efficiently, these algorithms showed impressive results, sometimes close to human-level performance.

In our case, we mainly took inspiration from generative algorithms. They work by the following: we feed them with thousands, millions of pictures of cats, and at the end of training, they should be able to generate pictures of cats that are different from the ones we have provided before. In particular, so called "generative adversarial networks" were found to be very good at this task. They consist of two networks that are trained in parallel. A generator network that tries to generate realistic images of cats, and a discriminator network that tries to decide whether a specific image is real or generated by the other network. At the end of learning, the generator should be able to generate images that look so realistic that the discriminator is confused and cannot distinguish them from real images.

These generation principles present some intriguing commonalities with the dreaming phenomenon, and we wondered whether they could help us to understand how the brain constructs these dreams we used these models to model dreaming in the cortex and evaluated how such dreams influence the learning of cortical representations.

We know sleep is important and restless nights can be tiresome. What are some of the effects a single restless night can have on our cognition?

The effects of a restless night have already been reported to have direct impact on healthy brain function the next day: memory retention, motor execution, attention and awareness are reduced. Driving after a restless night can be as dangerous as after having consumed alcohol! That being said, these are short-term effects of a lack of sleep. They could be recovered (at least partially) by catching up on sleep the next nights. In our model, we study the long-term effects of sleep: discovering general concepts about the world, such as recognizing different objects, understanding the meaning of words, which occur over weeks, months, or even years. From our model, we suggest that with patients with irregular REM sleep, such as in the case of certain pathologies or the administration of antidepressants, would significantly impact the learning of these general concepts.

During sleep there are different phases. Can you explain these phases and what happens to our brain during each phase?

There are basically two states: Rapid Eye Movement (REM) sleep, often referred to as paradoxical sleep, and non-REM sleep, also characterized as slow wave sleep, or deep sleep. During REM sleep, the brain is highly active, almost as much as during wakefulness, However the only muscles that are activated are the ones of the eyes and for breathing, All other skeletal muscles are inhibited, which confers this "paradoxical" aspect. Their inhibition concur with the abundance of vivid and emotional dreams that should not be acted out. During non-REM sleep, the brain generates slow waves of activity, roughly four waves per second. It is widely believed that in each wave, the memory from an episode of the last day stored in the hippocampus is replayed and consolidated in the neocortex. NREM and REM follow each other during the night and constitute cycles of about 1h30. However, we tend to have more REM sleep in the second half of the night.

The researchers simulated the cortex during three distinct states REM, Non-REM and Wakefulness, what was the result of this simulation in each state?

Let's consider that today, you're driving on the high-way and mostly observe cars passing by. Now let's picture what would happen in the subsequent night in our model. In our model, during NREM sleep, specific memories from the previous day are replayed and our artificial brain tries to reproduce the corresponding visual inputs, here, images of cars. We additionally add some perturbations on these inputs so that our brain is also used to perceiving these images when only partial information is available (e.g., hidden by a tree).

Now, let's talk about REM sleep. In your brain, you have other memories, from previous days, like when you were in the countryside and you saw many cows and sheeps. During REM sleep, our model assumes that we combine the memories from the previous day (cars), with the past memories (cows), and try to produce something realistic out of it. Even though it would lead to something bizarre (a car with features from a cow, like black and white patches and ears), it is not obvious for our brain to generate such a new, creative image. This is where the GANs enter: it teaches our brain to generate creative dreams. We postulate that during REM dreams, our brain is actively learning to generate something realistic, and by doing so, this improves our knowledge about the structure of the world.

Bibliography

Adamantidis, A. R., Gutierrez Herrera, C., and Gent, T. C. (2019). Oscillating circuitries in the sleeping brain. *Nature Reviews Neuroscience*, 20(12):746–762.

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., and Murphy, K. (2018). Fixing a broken elbo.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

Aru, J., Siclari, F., Phillips, W. A., and Storm, J. F. (2020). Apical drive—A cellular mechanism of dreaming? *Neuroscience & Biobehavioral Reviews*, 119:440–455.

Aru, J., Suzuki, M., Rutiku, R., Larkum, M. E., and Bachmann, T. (2019). Coupling the State and Contents of Consciousness. *Frontiers in Systems Neuroscience*, 13(August):1–9.

Aserinsky, E. and Kleitman, N. (1953). Regularly occurring periods of eye motility, and concomitant phenomena, during sleep. *Science*, 118(3062):273–274.

Ashby, F. G. (2014). *Multidimensional models of perception and cognition*. Psychology Press.

Avitan, L., Pujic, Z., Mölter, J., Zhu, S., Sun, B., and Goodhill, G. J. (2021). Spontaneous and evoked activity patterns diverge over development. *Elife*, 10:e61942.

Baird, B., Mota-Rolim, S. A., and Dresler, M. (2019). The cognitive neuroscience of lucid dreaming. *Neuroscience & Biobehavioral Reviews*, 100(May 2018):305–323.

Bang, D., Kang, S., and Shim, H. (2020). Discriminator feature-based inference by recycling the discriminator of gans. *International Journal of Computer Vision*, 128(10-11):2436–2458.

Bardes, A., Ponce, J., and LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*.

Barlow, H. (1995). The neuron doctrine in perception. In *The cognitive neurosciences.*, pages 415–435. The MIT Press, Cambridge, MA, US.

Barlow, H. B. et al. (1961). Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01).

Batterink, L. J., Oudiette, D., Reber, P. J., and Paller, K. A. (2014). Sleep facilitates learning a new linguistic rule. *Neuropsychologia*, 65:169–179.

Beckham, C., Honari, S., Verma, V., Lamb, A. M., Ghadiri, F., Hjelm, R. D., Bengio, Y., and Pal, C. (2019a). On Adversarial Mixup Resynthesis. page 12.

Beckham, C., Honari, S., Verma, V., Lamb, A. M., Ghadiri, F., Hjelm, R. D., Bengio, Y., and Pal, C. (2019b). On adversarial mixup resynthesis. *Advances in neural information processing systems*, 32.

Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., and Maass, W. (2020). A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature communications*, 11(1):1–15.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Benington, J. H. and Craig Heller, H. (1995). Restoration of brain energy metabolism as the function of sleep. *Progress in Neurobiology*, 45(4):347–360.

Benjamin, A. S. and Kording, K. P. (2021a). Learning to infer in recurrent biological networks.

Benjamin, A. S. and Kording, K. P. (2021b). Learning to infer in recurrent biological networks. *arXiv:2006.10811 [cs, q-bio, stat]*. arXiv: 2006.10811.

Bergelson, E. and Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, 114(49):12916–12921.

Bergelson, E. and Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9):3253–3258.

Berkes, P., Orbán, G., Lengyel, M., and Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331(6013):83–87.

Berthelot, D., Raffel, C., Roy, A., and Goodfellow, I. (2018). Understanding and Improving Interpolation in Autoencoders via an Adversarial Regularizer. *arXiv:1807.07543 [cs, stat]*. arXiv: 1807.07543.

Bishop, C. M. (1998). Latent variable models. In *Learning in graphical models*, pages 371–403. Springer.

Bond-Taylor, S., Leach, A., Long, Y., and Willcocks, C. G. (2021). Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1. arXiv:2103.04922 [cs, stat].

Booth, M. and Rolls, E. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral cortex (New York, N.Y. : 1991)*, 8:510–23.

Bornschein, J. and Bengio, Y. (2015). Reweighted wake-sleep.

Bova, S., Fazzi, E., Giovenzana, A., Montomoli, C., Signorini, S., Zoppello, M., and Lanzi, G. (2007). The Development of Visual Object Recognition in School-Age Children. *Developmental neuropsychology*, 31:79–102.

Boyce, R., Glasgow, S., Williams, S., and Adamantidis, A. (2016). Causal evidence for the role of rem sleep theta rhythm in contextual memory consolidation. *Science*, 352:812 – 816.

Boyce, R., Williams, S., and Adamantidis, A. (2017). REM sleep and memory. *Current Opinion in Neurobiology*, 44:167–177.

Brendel, W. and Bethge, M. (2019). Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*.

Brock, A., Donahue, J., and Simonyan, K. (2019). Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv:1809.11096 [cs, stat]*. arXiv: 1809.11096.

Brock, A., Lim, T., Ritchie, J. M., and Weston, N. (2017). Neural Photo Editing with Introspective Adversarial Networks. *arXiv:1609.07093 [cs, stat]*. arXiv: 1609.07093.

Burgess, N., Maguire, E. A., and O'Keefe, J. (2002). The Human Hippocampus and Spatial and Episodic Memory. *Neuron*, 35(4):625–641.

Buzsáki, G. (2002). Theta Oscillations in the Hippocampus. *Neuron*, 33(3):325–340.

Buzsáki, G. and Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science*, 304(5679):1926–1929.

Cai, D. J., Mednick, S. A., Harrison, E. M., Kanady, J. C., and Mednick, S. C. (2009). REM, not incubation, improves creativity by priming associative networks. *Proceedings of the National Academy of Sciences*, 106(25):10130–10134.

Carandini, M. (2005). Do We Know What the Early Visual System Does? *Journal of Neuroscience*, 25(46):10577–10597.

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. *arXiv:2002.05709 [cs, stat]*. arXiv: 2002.05709.

Chen, Z. and Wilson, M. A. (2017). Deciphering Neural Codes of Memory during Sleep. *Trends in Neurosciences*, 40(5):260–275.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204.

Cohrs, S. (2008). Sleep Disturbances in Patients with Schizophrenia. *CNS Drugs*, 22(11):939–962.

Connor, C. E., Brincat, S. L., and Pasupathy, A. (2007). Transformation of shape information in the ventral pathway. *Current Opinion in Neurobiology*, page 8.

Crick, F. and Mitchison, G. (1983a). The function of dream sleep. *Nature*, 304(5922):111–114.

Crick, F. and Mitchison, G. (1983b). The function of dream sleep. *Nature*, 304(5922):111–114.

Cropley, A. (2006). In praise of convergent thinking. *Creativity research journal*, 18(3):391–404.

Dai, Z., Yang, Z., Yang, F., Cohen, W. W., and Salakhutdinov, R. R. (2017). Good Semi-supervised Learning That Requires a Bad GAN. page 11.

Datta, S. (1999). Pgo wave generation: mechanism and functional significance. *Rapid eye movement sleep*, pages 91–106.

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The Helmholtz Machine. *Neural Computation*, 7(5):889–904.

Deperrois, N., Petrovici, M. A., Senn, W., and Jordan, J. (2022). Learning cortical representations through perturbed and adversarial dreaming. *eLife*, 11:e76384.

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, 73(3):415–434.

Diekelmann, S. and Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience*, 11(2):114–126.

Djonlagic, I., Rosenfeld, A., Shohamy, D., Myers, C., Gluck, M., and Stickgold, R. (2009). Sleep enhances category learning. *Learning & Memory*, 16(12):751–755.

Donahue, J., Krähenbühl, P., and Darrell, T. (2016). Adversarial Feature Learning. *arXiv:1605.09782 [cs, stat]*. arXiv: 1605.09782.

Donahue, J. and Simonyan, K. (2019). Large scale adversarial representation learning. *Advances in neural information processing systems*, 32.

Dosovitskiy, A. and Brox, T. (2016). Generating Images with Perceptual Similarity Metrics based on Deep Networks. *arXiv:1602.02644 [cs]*. arXiv: 1602.02644.

Dresler, M., Wehrle, R., Spoormaker, V. I., Koch, S. P., Holsboer, F., Steiger, A., Obrig, H., Sämann, P. G., and Czisch, M. (2012). Neural correlates of dream lucidity obtained from contrasting lucid versus non-lucid REM sleep: A combined EEG/fMRI case study. *Sleep*, 35(7):1017–1020.

Dudai, Y., Karni, A., and Born, J. (2015). The Consolidation and Transformation of Memory. *Neuron*, 88(1):20–32.

Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., and Courville, A. (2017). Adversarially Learned Inference. *arXiv:1606.00704 [cs, stat]*. arXiv: 1606.00704.

Dumoulin, V. and Visin, F. (2016). A guide to convolution arithmetic for deep learning.

Ellemberg, D., Lewis, T. L., Hong Liu, C., and Maurer, D. (1999). Development of spatial and temporal vision during childhood. *Vision Research*, 39(14):2325–2333.

Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D., and Walker, M. P. (2007). Human relational memory requires time and sleep. *Proceedings of the National Academy of Sciences*, 104(18):7723–7728.

Farroni Teresa, Johnson Mark H., Menon Enrica, Zulian Luisa, Faraguna Dino, and Csibra Gergely (2005). Newborns' preference for face-relevant stimuli: Effects of contrast polarity. *Proceedings of the National Academy of Sciences*, 102(47):17245–17250. Publisher: Proceedings of the National Academy of Sciences.

Fenn, K. M., Gallo, D. A., Margoliash, D., Roediger, H. L., and Nusbaum, H. C. (2009). Reduced false memory after sleep. *Learning & memory*, 16(9):509–513.

Ferguson, K. A. and Cardin, J. A. (2020). Mechanisms underlying gain modulation in the cortex. *Nature Reviews Neuroscience*.

Fiete, I. R., Fee, M. S., and Seung, H. S. (2007). Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances. *Journal of Neurophysiology*, 98(4):2038–2057. PMID: 17652414.

Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). "Who" Is Saying "What"? Brain-Based Decoding of Human Voice and Speech. *Science*, 322(5903):970–973.

Fosse, M. J., Fosse, R., Hobson, J. A., and Stickgold, R. J. (2003). Dreaming and episodic memory: A functional dissociation? *Journal of Cognitive Neuroscience*, 15(1):1–9.

Foulkes, D. (1999). *Children's dreaming and the development of consciousness*. Harvard University Press.

Freud, S. (1900). *The Interpretation of Dreams*. The Modern Library.

Friedrich, M., Wilhelm, I., Born, J., and Friederici, A. D. (2015). Generalization of word meanings during infant sleep. *Nature Communications*, 6(1):6004.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815–836.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202.

Geirhos, R., Meding, K., and Wichmann, F. A. (2020). Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, 33:13890–13902.

Gennari, G., Marti, S., Palu, M., Fló, A., and Dehaene-Lambertz, G. (2021). Orthogonal neural codes for speech in the infant brain. *Proceedings of the National Academy of Sciences*, 118(31):e2020410118.

Gershman, S. J. (2019). The Generative Adversarial Brain. *Frontiers in Artificial Intelligence*, 2.

Gidaris, S., Singh, P., and Komodakis, N. (2018). Unsupervised Representation Learning by Predicting Image Rotations. *arXiv:1803.07728 [cs]*. arXiv: 1803.07728.

Gilbert, C. D. and Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14(5):350–363.

Giuditta, A., Ambrosini, M. V., Montagnese, P., Mandile, P., Cotugno, M., Zucconi, G. G., and Vescia, S. (1995). The sequential hypothesis of the function of sleep. *Behavioural Brain Research*, 69(1):157 – 166. The Function of Sleep.

Goldstein, A. N. and Walker, M. P. (2014). The Role of Sleep in Emotional Brain Function. *Annual Review of Clinical Psychology*, 10(1):679–708.

Gómez, R. L., Bootzin, R. R., and Nadel, L. (2006). Naps promote abstraction in language-learning infants. *Psychological science*, 17(8):670–674.

Goodfellow, I. (2016). NIPS 2016 Tutorial: Generative Adversarial Networks. *arXiv:1701.00160 [cs]*. arXiv: 1701.00160.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks.

Gott, J. A., Liley, D. T. J., and Hobson, J. A. (2017). Towards a Functional Understanding of PGO Waves. *Frontiers in Human Neuroscience*, 11.

Greenberg, D. L. and Verfaellie, M. (2010). Interdependence of episodic and semantic memory: Evidence from neuropsychology. *Journal of the International Neuropsychological Society*, 16(5):748–753.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 290(1038):181–197.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., and Valko, M. (2020). Bootstrap your own latent - a new approach to self-supervised learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc.

Grill-Spector, K., Kourtzi, Z., and Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, 41(10):1409 – 1422.

Gross, C. G. (2002). Genealogy of the "grandmother cell". *The Neuroscientist*, 8(5):512–518.

Gruber, R. and Cassoff, J. (2014). The interplay between sleep and emotion regulation: conceptual framework empirical evidence and future directions. *Current psychiatry reports*, 16(11):1–9.

Guerguiev, J., Lillicrap, T. P., and Richards, B. A. (2017). Towards deep learning with segregated dendrites. *eLife*, 6:e22901.

Gui, J., Sun, Z., Wen, Y., Tao, D., and Ye, J. (2020). A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications. *arXiv:2001.06937 [cs, stat]*. arXiv: 2001.06937.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. *Advances in neural information processing systems*, 30.

Guo, W., Wang, J., and Wang, S. (2019). Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394.

Ha, D. and Schmidhuber, J. (2018a). Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31.

Ha, D. and Schmidhuber, J. (2018b). World models. *arXiv preprint arXiv:1803.10122*.

Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. (2019). Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.

Hafner, D., Lillicrap, T., Norouzi, M., and Ba, J. (2020). Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*.

Haghiri, S., Wichmann, F. A., and von Luxburg, U. (2020). Estimation of perceptual scales using ordinal embedding. *Journal of vision*, 20(9):14–14.

Haider, P., Ellenberger, B., Kriener, L., Jordan, J., Senn, W., and Petrovici, M. (2021). Latent equilibrium: Arbitrarily fast computation with arbitrarily slow neurons. *Advances in Neural Information Processing Systems*, 34.

Hartmann, E. (2007). The nature and functions of dreaming. *The new science of dreaming*, 3:171–192.

Haxby James V., Gobbini M. Ida, Furey Maura L., Ishai Alumit, Schouten Jennifer L., and Pietrini Pietro (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, 293(5539):2425–2430. Publisher: American Association for the Advancement of Science.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, Seattle, WA, USA. IEEE.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2018). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *arXiv:1706.08500 [cs, stat]*. arXiv: 1706.08500.

Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D., and Botvinick, M. (2021). Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature Communications*, 12(1):6456.

Hinton, G. (1984). *Boltzmann Machines: Constraint Satisfaction Networks that Learn*. Carnegie-Mellon University, Department of Computer Science.

Hinton, G., Dayan, P., Frey, B., and Neal, R. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. (2019). Learning deep representations by mutual information estimation and maximization. *arXiv:1808.06670 [cs, stat]*. arXiv: 1808.06670.

Hobson, J. A. (2009). REM sleep and dreaming: towards a theory of protoconsciousness. *Nature Reviews Neuroscience*, 10(11):803–813.

Hobson, J. A., Hong, C. C.-H., and Friston, K. J. (2014). Virtual reality and consciousness inference in dreaming. *Frontiers in Psychology*, 5.

Hobson, J. A. and McCarley, R. W. (1977). The brain as a dream state generator: an activation-synthesis hypothesis of the dream process. *The American journal of psychiatry*.

Hobson, J. A., Pace-Schott, E. F., and Stickgold, R. (2000). Dreaming and the brain: Toward a cognitive neuroscience of conscious states. *Behavioral and Brain Sciences*, 23(6):793–842.

Hoel, E. (2021). The overfitted brain: Dreams evolved to assist generalization. *Patterns*, 2(5):100244.

Horikawa, T., Tamaki, M., Miyawaki, Y., and Kamitani, Y. (2013). Neural Decoding of Visual Imagery During Sleep. *Science*, 340(6132):639–642.

Huang, H., Li, Z., He, R., Sun, Z., and Tan, T. (2018). Introvae: Introspective variational autoencoders for photographic image synthesis.

Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. *Science*, 310(5749):863–866.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167.

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., and Haxby, J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences*, 96(16):9379–9384.

Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology*, 73(1):218–226.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. (2020). A Survey on Contrastive Self-Supervised Learning. *Technologies*, 9(1):2.

Jenkins, J. G. and Dallenbach, K. M. (1924). Obliviscence during Sleep and Waking. *The American Journal of Psychology*, 35(4):605–612. Publisher: University of Illinois Press.

Ji, D. and Wilson, M. A. (2007). Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature Neuroscience*, 10(1):100–107.

Johnson, M. H., Dziurawiec, S., Ellis, H., and Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, 40(1):1–19.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

Joyce, J. M. (2011). Kullback-leibler divergence. In *International encyclopedia of statistical science*, pages 720–722. Springer.

Káli, S. and Dayan, P. (2004). Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nature Neuroscience*, 7(3):286–294.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. (2020). Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114.

Karras, T., Laine, S., and Aila, T. (2018). A Style-Based Generator Architecture for Generative Adversarial Networks. *arXiv:1812.04948 [cs, stat]*. arXiv: 1812.04948.

Keller, G. B. and Mrsic-Flogel, T. D. (2018). Predictive Processing: A Canonical Cortical Computation. *Neuron*, 100(2):424–435.

Kerr, N. H. and Foulkes, D. (1981). Right hemispheric mediation of dream visualization: A case study. *Cortex*, 17(4):603–609.

Kingma, D. P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. arXiv: 1412.6980.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes.

Kingma, D. P. and Welling, M. (2019). An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

Klinzing, J. G., Niethard, N., and Born, J. (2019). Mechanisms of systems memory consolidation during sleep. *Nature Neuroscience*, 22(10):1598–1610.

Knutson, K. L., Spiegel, K., Penev, P., and Van Cauter, E. (2007). The metabolic consequences of sleep deprivation. *Sleep Medicine Reviews*, 11(3):163–178.

Koch, C. and Poggio, T. (1999). Predicting the visual world: silence is golden. *nature neuroscience*, 2(1):9–10.

Konkle, T. and Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, 13(1):491.

Korcsak-Gorzo, A., Müller, M. G., Baumbach, A., Leng, L., Breitwieser, O. J., van Albada, S. J., Senn, W., Meier, K., Legenstein, R., and Petrovici, M. A. (2021). Cortical oscillations implement a backbone for sampling-based computation in spiking neural networks.

Kreiman, G., Koch, C., and Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience*, 3(9):946–953.

Krizhevsky, A., Nair, V., and Hinton, G. (2013). Cifar-10 (canadian institute for advanced research).

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Lange, T., Dimitrov, S., and Born, J. (2010). Effects of sleep and circadian rhythm on the human immune system: Sleep, rhythms, and immune functions. *Annals of the New York Academy of Sciences*, 1193(1):48–59.

Lau, H., Tucker, M., and Fishbein, W. (2010). Daytime napping: Effects on human direct associative and relational memory. *Neurobiology of learning and memory*, 93(4):554–560.

Le-Khac, P. H., Healy, G., and Smeaton, A. F. (2020). Contrastive Representation Learning: A Framework and Review. *IEEE Access*, 8:193907–193934.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, 1(4):541–551.

Léger, D., Debellemaniere, E., Rabat, A., Bayon, V., Benchenane, K., and Chennaoui, M. (2018). Slow-wave sleep: From the cell to the clinic. *Sleep Medicine Reviews*, 41:113–132.

Levin, R. and Nielsen, T. (2009). Nightmares, bad dreams, and emotion dysregulation: A review and new neurocognitive model of dreaming. *Current Directions in psychological science*, 18(2):84–88.

Lewis, P. A. and Durrant, S. J. (2011). Overlapping memory replay during sleep builds cognitive schemata. *Trends in Cognitive Sciences*, 15(8):343–351.

Lewis, P. A., Knoblich, G., and Poe, G. (2018). How Memory Replay in Sleep Boosts Creative Problem-Solving. *Trends in Cognitive Sciences*, 22(6):491–503.

Li, W., Ma, L., Yang, G., and Gan, W.-b. (2017). REM sleep selectively prunes and maintains new synapses in development and learning. *Nature Neuroscience*, 20(3):427–437.

Lillicrap, T. P., Cownden, D., Tweed, D. B., and Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7(1).

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., and Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*.

Lim, S., McKee, J. L., Woloszyn, L., Amit, Y., Freedman, D. J., Sheinberg, D. L., and Brunel, N. (2015). Inferring learning rules from distributions of firing rates. *Nature neuroscience*, 18(12):1–13.

Lindsay, G. W. (2021). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, 33(10):2017–2031.

Liu, X., Zhang, F., Hou, Z., Wang, Z., Mian, L., Zhang, J., and Tang, J. (2021). Self-supervised Learning: Generative or Contrastive. *arXiv:2006.08218 [cs, stat]*. arXiv: 2006.08218.

Llewellyn, S. (2016a). Crossing the invisible line: De-differentiation of wake, sleep and dreaming may engender both creative insight and psychopathology. *Consciousness and Cognition*, 46:127–147.

Llewellyn, S. (2016b). Dream to Predict? REM Dreaming as Prospective Coding. *Frontiers in Psychology*, 6.

Lutz, N. D., Diekelmann, S., Hinse-Stern, P., Born, J., and Rauss, K. (2017). Sleep Supports the Slow Abstraction of Gist from Visual Perceptual Memories. *Scientific Reports*, 7(1).

Maas, A. L., Hannun, A. Y., and Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.

Majaj, N. J., Hong, H., Solomon, E. A., and DiCarlo, J. J. (2015). Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. *Journal of Neuroscience*, 35(39):13402–13418.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. (2015). Adversarial Autoencoders. *arXiv:1511.05644 [cs]*. arXiv: 1511.05644.

Malcolm-Smith, S. and Solms, M. (2004). Incidence of threat in dreams: A response to revonsuo's threat simulation theory. *Dreaming*, 14(4):220.

Mamelak, A. N. and Hobson, J. A. (1989). Dream Bizarreness as the Cognitive Correlate of Altered Neuronal Behavior in REM Sleep. *Journal of Cognitive Neuroscience*, 1(3):201–222.

Mao, X., Su, Z., Tan, P. S., Chow, J. K., and Wang, Y.-H. (2020). Is Discriminator a Good Feature Extractor? *arXiv:1912.00789 [cs, stat]*. arXiv: 1912.00789.

Marblestone, A. H., Wayne, G., and Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Frontiers in computational neuroscience*, 10:94.

Marino, J. (2022). Predictive Coding, Variational Autoencoders, and Biological Connections. *Neural Computation*, 34(1):1–44.

Mazzarello, P. (2000). What dreams may come? *Nature*, 408(6812):523–523.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457.

McKay, B. E., Placzek, A. N., and Dani, J. A. (2007). Regulation of synaptic transmission and plasticity by neuronal nicotinic acetylcholine receptors. *Biochemical Pharmacology*, 74(8):1120–1133.

McManus, J. N. J., Li, W., and Gilbert, C. D. (2011). Adaptive shape processing in primary visual cortex. *Proceedings of the National Academy of Sciences*, 108(24):9739–9746.

Meeter, M. and Murre, J. M. J. (2004). Consolidation of Long-Term Memory: Evidence and Alternatives. *Psychological Bulletin*, 130(6):843–857.

Menéndez, M., Pardo, J., Pardo, L., and Pardo, M. (1997). The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Millidge, B., Seth, A., and Buckley, C. L. (2022). Predictive Coding: a Theoretical and Experimental Review. arXiv:2107.12979 [cs, q-bio].

Misra, I. and van der Maaten, L. (2020). Self-Supervised Learning of Pretext-Invariant Representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6706–6716, Seattle, WA, USA. IEEE.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing Atari with Deep Reinforcement Learning. *arXiv:1312.5602 [cs]*. arXiv: 1312.5602.

Moerland, T. M., Broekens, J., and Jonker, C. M. (2020). Model-based reinforcement learning: A survey. *arXiv preprint arXiv:2006.16712*.

Moore, T. and Zirnsak, M. (2017). Neural Mechanisms of Selective Visual Attention. *Annual Review of Psychology*, 68(1):47–72.

Munjal, P., Paul, A., and Krishnan, N. C. (2019). Implicit discriminator in variational autoencoder.

Murray, S. O., Schrater, P., and Kersten, D. (2004). Perceptual grouping and the interactions between visual cortical areas. *Neural Networks*, 17(5-6):695–705.

Nadasdy, Z., Hirase, H., Czurkó, A., Csicsvari, J., and Buzsáki, G. (1999). Replay and time compression of recurring spike sequences in the hippocampus. *The Journal of Neuroscience*, 19:9497 – 9507.

Nadel, L. and Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology*, 7:217–227.

Nayebi, A., Srivastava, S., Ganguli, S., and Yamins, D. L. (2020). Identifying learning rules from neural network observables. *Advances in Neural Information Processing Systems*, 33:2639–2650.

Nelson, J. P., McCarley, R. W., and Hobson, J. A. (1983). REM sleep burst neurons, PGO waves, and eye movement information. *Journal of Neurophysiology*, 50(4):784–797.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. (2011). Reading digits in natural images with unsupervised feature learning.

Nielsen, T. A. and Stenstrom, P. (2005). What are the memory sources of dreaming? *Nature*, 437(7063):1286–1289.

Nir, Y. and Tononi, G. (2010). Dreaming and the brain: from phenomenology to neurophysiology. *Trends in Cognitive Sciences*, 14(2):88–100.

Nishimura, M., Scherf, S., and Behrmann, M. (2009). Development of object recognition in humans. *F1000 Biology Reports*, 1.

Norman, K. A., Newman, E. L., and Perotte, A. J. (2005). Methods for reducing interference in the Complementary Learning Systems model: Oscillating inhibition and autonomous memory rehearsal. *Neural Networks*, 18(9):1212–1228.

Odena, A. (2016). Semi-Supervised Learning with Generative Adversarial Networks. *arXiv:1606.01583 [cs, stat]*. arXiv: 1606.01583.

O'Neill, J., Pleydell-Bouverie, B., Dupret, D., and Csicsvari, J. (2010). Play it again: reactivation of waking experience and memory. *Trends in Neurosciences*, 33(5):220–229.

Oudiette, D., Dealberto, M.-J., Uguccioni, G., Golmard, J.-L., Merino-Andreu, M., Tafti, M., Garma, L., Schwartz, S., and Arnulf, I. (2012). Dreaming without rem sleep. *Consciousness and cognition*, 21(3):1129–1140.

Palmiero, M., Nori, R., Aloisi, V., Ferrara, M., and Piccardi, L. (2015). Domain-Specificity of Creativity: A Study on the Relationship Between Visual Creativity and Visual Mental Imagery. *Frontiers in Psychology*, 6.

Pardilla-Delgado, E. and Payne, J. D. (2017). The impact of sleep on true and false memory across long delays. *Neurobiology of Learning and Memory*, 137:123–133.

Payne, J. D., Schacter, D. L., Propper, R. E., Huang, L.-W., Wamsley, E. J., Tucker, M. A., Walker, M. P., and Stickgold, R. (2009). The role of sleep in false memory formation. *Neurobiology of Learning and Memory*, 92(3):327–334.

Pearson, J. (2019). The human imagination: the cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience*, 20(10):624–634.

Penny, W. D., Stephan, K. E., Mechelli, A., and Friston, K. J. (2004). Comparing dynamic causal models. *Neuroimage*, 22(3):1157–1172.

Poe, G. R. (2017). Sleep is for forgetting. *Journal of Neuroscience*, 37(3):464–473.

Pogodin, R., Mehta, Y., Lillicrap, T. P., and Latham, P. E. (2021). Towards Biologically Plausible Convolutional Networks. *arXiv:2106.13031 [cs, q-bio]*. arXiv: 2106.13031.

Quiroga, R. Q., Kreiman, G., Koch, C., and Fried, I. (2008). Sparse but not 'grandmother-cell'coding in the medial temporal lobe. *Trends in cognitive sciences*, 12(3):87–91.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv:1511.06434 [cs]*. arXiv: 1511.06434.

Rao, R. P. N. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87.

Rasch, B. and Born, J. (2013). About Sleep's Role in Memory. *Physiological Reviews*, 93(2):681–766.

Ratliff, L. J., Burden, S. A., and Sastry, S. S. (2013). Characterization and computation of local nash equilibria in continuous games. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 917–924. IEEE.

Rennó-Costa, C., da Silva, A. C. C., Blanco, W., and Ribeiro, S. (2019). Computational models of memory consolidation and long-term synaptic plasticity during sleep. *Neurobiology of Learning and Memory*, 160:32–47.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., Poirazi, P., Roelfsema, P., Sacramento, J., Saxe, A., Scellier, B., Schapiro, A. C., Senn, W., Wayne, G., Yamins, D., Zenke, F., Zylberberg, J., Therien, D., and Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770.

Richler, J. J. and Palmeri, T. J. (2014). Visual category learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(1):75–94.

Rosca, M., Lakshminarayanan, B., Warde-Farley, D., and Mohamed, S. (2017). Variational Approaches for Auto-Encoding Generative Adversarial Networks. *arXiv:1706.04987 [cs, stat]*. arXiv: 1706.04987.

Rosenbaum, R., Winocur, G., and Moscovitch, M. (2001). New views on old memories: re-evaluating the role of the hippocampal complex. *Behavioural Brain Research*, 127(1-2):183–197.

Rosenblatt, F. (1961). Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Rumelhart, D. E. and McClelland, J. L. (1987). *Information Processing in Dynamical Systems: Foundations of Harmony Theory*, pages 194–281.

Rust, N. C. and DiCarlo, J. J. (2010). Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT. *Journal of Neuroscience*, 30(39):12978–12995.

Sacramento, J. a., Ponte Costa, R., Bengio, Y., and Senn, W. (2018). Dendritic cortical microcircuits approximate the backpropagation algorithm. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved Techniques for Training GANs. *arXiv:1606.03498 [cs]*. arXiv: 1606.03498.

Sawangjit, A., Oyanedel, C. N., Niethard, N., Salazar, C., Born, J., and Inostroza, M. (2018). The hippocampus is crucial for forming non-hippocampal long-term memory during sleep. *Nature*, 564(7734):109–113.

Schapiro, A. C., McDevitt, E. A., Chen, L., Norman, K. A., Mednick, S. C., and Rogers, T. T. (2017). Sleep benefits memory for semantic category structure while preserving exemplar-specific information. *Scientific reports*, 7(1):1–13.

Schoenfeld, G., Kollmorgen, S., Lewis, C., Bethge, P., Ruess, A. M. A., Aguzzi, A., Mante, V., and Helmchen, F. (2022). Dendritic integration of sensory and reward information facilitates learning. *bioRxiv*, pages 1–31.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609.

Schwartz, S. (2003). Are life episodes replayed during dreaming? *Trends in Cognitive Sciences*, 7(8):325–327.

Seibt, J., Richard, C. J., Sigl-Glöckner, J., Takahashi, N., Kaplan, D. I., Doron, G., Limoges, D. D., Bocklisch, C., and Larkum, M. E. (2017). Cortical dendritic activity correlates with spindle-rich oscillations during sleep in rodents. *Nature Communications*, 8(684):1–13.

Senn, W. and Sacramento, J. (2015). Backward reasoning the formation rules. *Nature Neuroscience*, 18(12):1705–1706.

Shin, H., Lee, J. K., Kim, J., and Kim, J. (2017). Continual Learning with Deep Generative Replay. page 10.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48.

Siclari, F., Baird, B., Perogamvros, L., Bernardi, G., LaRocque, J. J., Riedner, B., Boly, M., Postle, B. R., and Tononi, G. (2017). The neural correlates of dreaming. *Nature Neuroscience*, 20(6):872–878.

Siegel, J. M. (2009). Sleep viewed as a state of adaptive inactivity. *Nature Reviews Neuroscience*, 10(10):747–753.

Silver, D., Hasselt, H., Hessel, M., Schaul, T., Guez, A., Harley, T., Dulac-Arnold, G., Reichert, D., Rabinowitz, N., Barreto, A., et al. (2017). The predictron: End-to-end learning and planning. In *International Conference on Machine Learning*, pages 3191–3199. PMLR.

Simons, J. S., Garrison, J. R., and Johnson, M. K. (2017). Brain mechanisms of reality monitoring. *Trends in Cognitive Sciences*, 21(6):462–473.

Sjöström, P. J. and Häusser, M. (2006). A Cooperative Switch Determines the Sign of Synaptic Plasticity in Distal Dendrites of Neocortical Pyramidal Neurons. *Neuron*, 51(2):227–238.

Slone, L. K. and Johnson, S. P. (2015). Infants' statistical learning: 2- and 5-month-olds' segmentation of continuous visual sequences. *Journal of Experimental Child Psychology*, 133:47–56.

Solms, M. (2000). Dreaming and rem sleep are controlled by different brain mechanisms. *Behavioral and Brain Sciences*, 23(6):843–850.

Spanò, G., Pizzamiglio, G., McCormick, C., Clark, I. A., De Felice, S., Miller, T. D., Edgin, J. O., Rosenthal, C. R., and Maguire, E. A. (2020). Dreaming with hippocampal damage. *eLife*, 9.

Spoerer, C. J., McClure, P., and Kriegeskorte, N. (2017). Recurrent Convolutional Neural Networks: A Better Model of Biological Object Recognition. *Frontiers in Psychology*, 8.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Steriade, M. (2006). Grouping of brain rhythms in corticothalamic systems. *Neuroscience*, 137(4):1087–1106.

Stickgold, R., Malia, A., Maguire, D., Roddenberry, D., and O'Connor, M. (2000). Replaying the game: hypnagogic images in normals and amnesics. *Science*, 290(5490):350–353.

Stickgold, R., Scott, L., Rittenhouse, C., and Hobson, J. A. (1999). Sleep-induced changes in associative memory. *Journal of cognitive neuroscience*, 11(2):182–193.

Subramaniam, K., Luks, T. L., Fisher, M., Simpson, G. V., Nagarajan, S., and Vinogradov, S. (2012). Computerized Cognitive Training Restores Neural Activity within the Reality Monitoring Network in Schizophrenia. *Neuron*, 73(4):842–853.

Takahashi, N., Ebner, C., Sigl-Glöckner, J., Moberg, S., Nierwetberg, S., and Larkum, M. E. (2020). Active dendritic currents gate descending cortical outputs in perception. *Nature Neuroscience*, 23(10):1277–1285.

Tang, H., Li, H., and Yan, R. (2010). Memory Dynamics in Attractor Networks with Saliency Weights. *Neural Computation*, 22(7):1899–1926.

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., Hardesty, W., Cox, D., and Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, 115(35):8835–8840.

Tarr, M. J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin & Review*, 2(1):55–82.

Tefft, B. C. (2018). Acute sleep deprivation and culpable motor vehicle crash involvement. *Sleep*, 41(10). zsy144.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, 331(6022):1279–1285.

Terada, Y. and Luxburg, U. (2014). Local ordinal embedding. In *International Conference on Machine Learning*, pages 847–855. PMLR.

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582):520–522.

Tononi, G. and Cirelli, C. (2014). Sleep and the Price of Plasticity: From Synaptic and Cellular Homeostasis to Memory Consolidation and Integration. *Neuron*, 81(1):12–34.

Tononi, G. and Cirelli, C. (2020). Sleep and synaptic down-selection. *European Journal of Neuroscience*, 51(1):413–421.

Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. (2020). On mutual information maximization for representation learning.

Tulving, E. (1972). *Episodic and semantic memory.*, pages xiii, 423–xiii, 423. Academic Press, Oxford, England.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology*, 53(1):1–25. PMID: 11752477.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2017). It Takes (Only) Two: Adversarial Generator-Encoder Networks. *arXiv:1704.02304 [cs, stat].* arXiv: 1704.02304.

Urbanczik, R. and Senn, W. (2014). Learning by the Dendritic Prediction of Somatic Spiking. *Neuron*, 81(3):521–528.

van de Ven, G. M., Siegelmann, H. T., and Tolias, A. S. (2020). Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, 11(1).

van Engelen, J. E. and Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440.

Voigts, J. and Harnett, M. T. (2020). Somatic and Dendritic Encoding of Spatial Variables in Retrosplenial Cortex Differs during 2D Navigation. *Neuron*, 105(2):237–245.e4.

Walker, M. (2017). *Why we sleep: Unlocking the power of sleep and dreams.* Simon and Schuster.

Walker, M. P. (2009). The Role of Sleep in Cognition and Emotion. *Annals of the New York Academy of Sciences*, 1156(1):168–197.

Walker, M. P., Liston, C., Hobson, J. A., and Stickgold, R. (2002). Cognitive flexibility across the sleep–wake cycle: Rem-sleep enhancement of anagram problem solving. *Cognitive Brain Research*, 14(3):317–324.

Wamsley, E. J. (2014). Dreaming and Offline Memory Consolidation. *Current Neurology and Neuroscience Reports*, 14(3).

Wamsley, E. J., Hirota, Y., Tucker, M. A., Smith, M. R., and Antrobus, J. S. (2007). Circadian and ultradian influences on dreaming: A dual rhythm model. *Brain Research Bulletin*, 71(4):347–354.

Wamsley, E. J., Perry, K., Djonlagic, I., Reaven, L. B., and Stickgold, R. (2010a). Cognitive replay of visuomotor learning at sleep onset: temporal dynamics and relationship to task performance. *Sleep*, 33(1):59–68.

Wamsley, E. J. and Stickgold, R. (2019). Dreaming of a learning task is associated with enhanced memory consolidation: Replication in an overnight sleep study. *Journal of sleep research*, 28(1):e12749.

Wamsley, E. J., Tucker, M., Payne, J. D., Benavides, J. A., and Stickgold, R. (2010b). Dreaming of a learning task is associated with enhanced sleep-dependent memory consolidation. *Current Biology*, 20(9):850–855.

Waters, F., Blom, J. D., Dang-Vu, T. T., Cheyne, A. J., Alderson-Day, B., Woodruff, P., and Collerton, D. (2016). What Is the Link Between Hallucinations, Dreams, and Hypnagogic–Hypnopompic Experiences? *Schizophrenia Bulletin*, 42(5):1098–1109.

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

Whittington, J. C. and Bogacz, R. (2019). Theories of Error Back-Propagation in the Brain. *Trends in Cognitive Sciences*, 23(3):235–250.

Wierzynski, C. M., Lubenov, E. V., Gu, M., and Siapas, A. G. (2009). State-dependent spike-timing relationships between hippocampal and prefrontal circuits during sleep. *Neuron*, 61(4):587–596.

Williams, J., Merritt, J., Rittenhouse, C., and Hobson, J. (1992). Bizarreness in dreams and fantasies: Implications for the activation-synthesis hypothesis. *Consciousness and Cognition*, 1(2):172–185.

Winocur, G. and Moscovitch, M. (2011). Memory Transformation and Systems Consolidation. *Journal of the International Neuropsychological Society*, 17(05):766–780.

Winocur, G., Moscovitch, M., and Bontempi, B. (2010). Memory formation and long-term retention in humans and animals: Convergence towards a transformation account of hippocampal–neocortical interactions. *Neuropsychologia*, 48(8):2339–2356.

Xie, L., Kang, H., Xu, Q., Chen, M. J., Liao, Y., Thiyagarajan, M., O'Donnell, J., Christensen, D. J., Nicholson, C., Iliff, J. J., et al. (2013). Sleep drives metabolite clearance from the adult brain. *science*, 342(6156):373–377.

Yamins, D. L. K. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624.

Yee, E., Chrysikou, E. G., and Thompson-Schill, S. L. (2013). *Semantic Memory*. Oxford University Press.

Yuste, R. (2015). From the neuron doctrine to neural networks. *Nature Reviews Neuroscience*, 16(8):487–497.

Zadra, A., Desjardins, S., and Marcotte, E. (2006). Evolutionary function of dreams: A test of the threat simulation theory in recurrent dreams. *Consciousness and Cognition*, 15(2):450–463.

Zadra, A. and Stickgold, R. (2021). *When brains dream: Exploring the science and mystery of sleep*. WW Norton.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*.

Zeiler, M. D. and Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. *arXiv:1311.2901 [cs]*. arXiv: 1311.2901.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., and Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118.

Declaration of Authorship

Last name, first name: Nicolas Rouben Pascal Deperrois
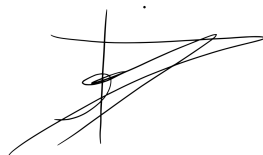
Matriculation number: 18-131-664

I hereby declare that this thesis represents my original work and that I have used no other sources except as noted by citations.

All data, tables, figures and text citations which have been reproduced from any other source, including the internet, have been explicitly acknowledged as such.

I am aware that in case of non-compliance, the Senate is entitled to withdraw the doctorate degree awarded to me on the basis of the present thesis, in accordance with the Statut der Universität Bern (Universitätsstatut; UniSt), Art. 69, of 7 June 2011.

Place, Date: Bern, 01/11/2022

Signature: