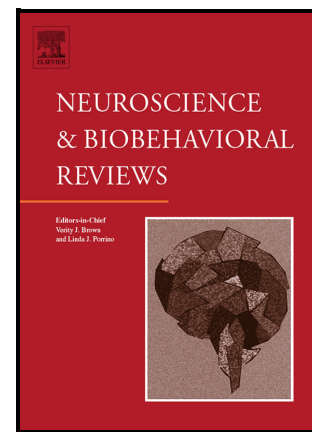# Journal Pre-proof

Learning beyond sensations: how dreams organize neuronal representations

Nicolas Deperrois, Mihai A. Petrovici, Walter Senn, Jakob Jordan

begindocument/before
shipout/backgroundshipout/foreground
file/after/translations-basic-dictionary-english.trsl

Learning beyond sensations: how dreams organize neuronal representations

Nicolas Deperrois[a,*] Mihai A. Petrovici[a] Walter Senn[a] Jakob Jordan[a,b]

[a]*Department of Physiology, University of Bern, Bern, Switzerland*
[b]*Electrical Engineering, Yale University, New Haven, CT, United States*

Abstract

Semantic representations in higher sensory cortices form the basis for robust, yet flexible behavior. These representations are acquired over the course of development in an unsupervised fashion and continuously maintained over an organism's lifespan. Predictive processing theories propose that these representations emerge from predicting or reconstructing sensory inputs. However, brains are known to generate virtual experiences, such as during imagination and dreaming, that go beyond previously experienced inputs. Here, we suggest that virtual experiences may be just as relevant as actual sensory inputs in shaping cortical representations. In particular, we discuss two complementary learning principles that organize representations through the generation of virtual experiences. First, "adversarial dreaming" proposes that creative dreams support a cortical implementation of adversarial learning in which feedback and feedforward pathways engage in a productive game of trying to fool each other. Second, "contrastive dreaming" proposes that the invariance of neuronal representations to irrelevant factors of variation is acquired by trying to map similar virtual experiences together via a contrastive learning process. These principles are compatible with known cortical structure and dynamics and the phenomenology of sleep thus providing promising directions to explain cortical learning beyond the classical predictive processing paradigm.

Keywords: semantic representations, unsupervised learning, predictive coding, sleep, virtual experiences, dreaming, adversarial, contrastive, creativity

1. Introduction

Throughout their life, animals enjoy a wide variety of unique sensory experiences. However, seemingly unaffected by this diversity, animals exhibit a remarkable degree of consistency in their behaviour and can, often effortlessly, leverage prior knowledge to generalize to novel circumstances. For example, they easily recognize which category an object belongs to (Biederman, 1987), within a fraction of a second (Thorpe et al., 1996), and despite the various conditions in which this object can be observed (DiCarlo et al., 2012). How is this possible?

Insights from neuroscience and machine learning suggest that this cognitive feat may be grounded in neuronal activity patterns in higher cortical areas that reflect the *semantic* content of sensory inputs. Thus, these "semantic neuronal representations" extract relevant factors of variation such as object categories from stimuli while remaining invariant to irrelevant factors such as pose, lighting, or partial occlusions (Barlow, 2001; DiCarlo et al., 2012). Strikingly, such an organized and invariant code is observed in recordings from the inferior temporal (IT) cortex, the highest area of the ventral visual stream (Figure 1a; Grill-Spector et al., 2001; Hung et al., 2005).

Such structure in neuronal activities arises over the course of development (Figure 1b; Rodman, 1994). Yet, the mechanisms underlying this emergence remain unclear. Computational models of the sensory cortex attempt to explain how, from sensory evoked activities (activities from lower cortical areas, e.g., in V1 cortex), neurons extract
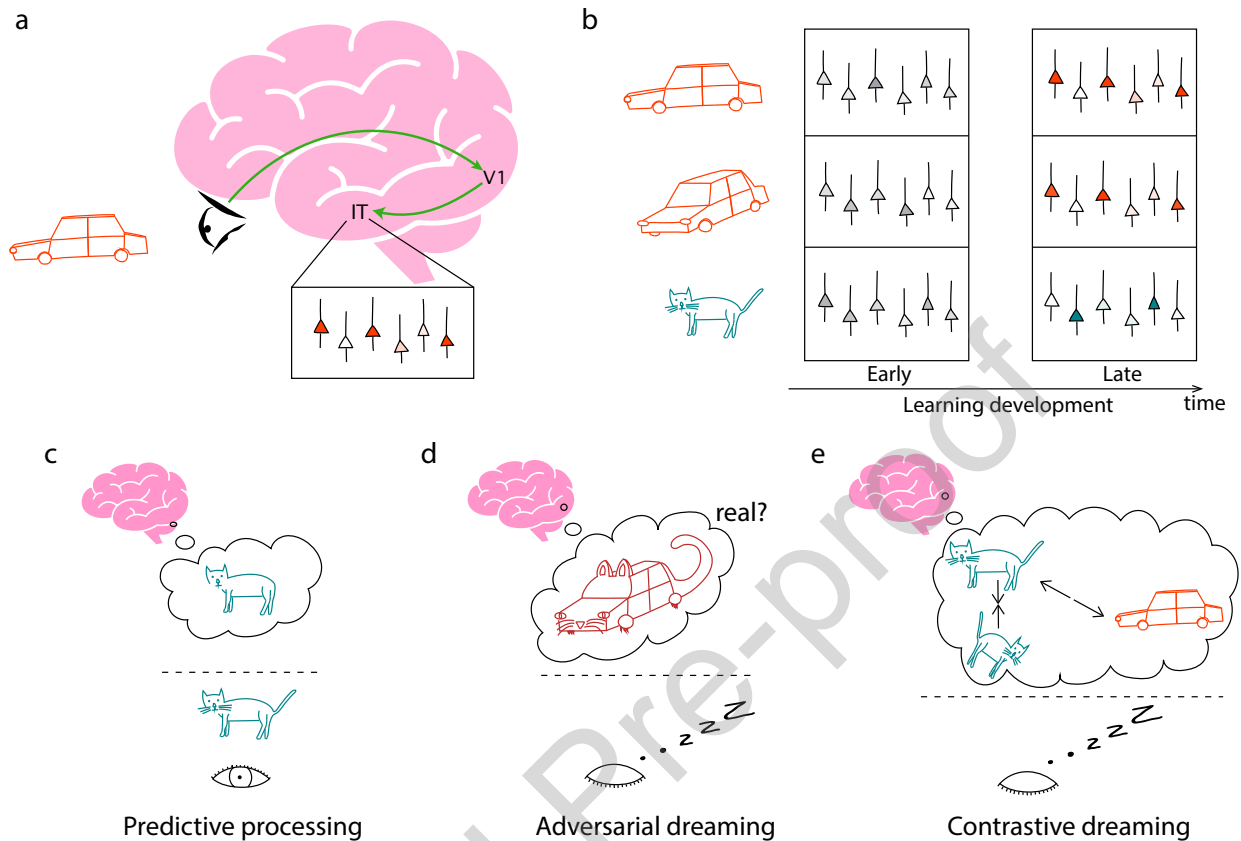
Figure 1: **Semantic representations in higher cortical areas emerge over the course of development.** (a) Sketch of time-averaged activity of neurons in IT in response to a visual stimulus. Visual stimuli activate cells in the retina and these signals are processed along the hierarchy of the visual cortex, here the ventral visual stream. (b) Typical neuronal responses to the presentation of different objects at early and late stages of development. Over the course of development, activity patterns align with the semantic category of the input (here, different patterns encode "cat" and "car" stimuli) but are invariant to semantically preserving transformations (e.g., cars from different viewing angles). (c) Common neuroscientific theories hypothesize that brains learn representations by trying to predict their sensorium (predictive processing, Rao and Ballard, 1999). (d) During offline states, e.g., sleep, brains continue to generate virtual experiences that may further contribute to learning semantic representations, for example by combining several memories into new, realistic experiences. (e) In addition, regenerating previous experiences with natural semantically preserving variations can provide additional signals for learning.

features of increasing complexity along the cortical hierarchy (Hubel and Wiesel, 1965), leading to high-level semantic representations (Richards et al., 2019). Supervised models of sensory processing suggest that cortical feedforward pathways learn to map sensory inputs to specific object categories that are externally provided, for example by a teacher. However, animals seem to learn with little to no supervision and do not require millions of category labels during development (Bergelson and Swingley, 2012; Bergelson and Aslin, 2017; Slone and Johnson, 2015; Huber et al., 2021).

To acquire semantic latent representations, cortical networks may thus leverage learning principles that do not rely on labelled data, similar to unsupervised machine learning models (Liu et al., 2021; Zhuang et al., 2021). For example, cortical feedback pathways could implicitly learn the structure of the sensorium by generating activities in lower sensory cortex that are similar to sensory-evoked responses (generative modeling, Clark, 2013). Therefore, we first discuss the predictive processing framework (Figure 1c, Rao and Ballard, 1999), that posits that the brain shapes its representations by trying to predict evoked activities in early sensory cortex. We then present "adversarial dreaming" (Figure 1d; Deperrois et al., 2022) where representations are improved through the generation of creative

2

dreams during sleep and their discrimination from actual experiences.

Another possibility is to directly shape feedforward pathways to construct relevant high-level representations, by using simple, alternative labels that can directly be inferred from data (self-supervised learning, Ericsson et al., 2022). Accordingly, we introduce "contrastive dreaming" (Figure 1e), during which neuronal activities in higher cortical areas are pulled together for semantically similar inputs, and pushed apart for dissimilar stimuli.

For each framework, we start by presenting the underlying computational principles and then discuss suggested bio-plausible implementations. Finally, we discuss experimental approaches to (in)validate the presented hypotheses.

2. Learning by predicting evoked low-level cortical activities

2.1. Principles of predictive processing

> **Box 1 - Predictive processing theories**
>
> Predictive coding (Rao and Ballard, 1999) hypothesizes that brains minimize errors between predictions generated by feedback pathways and low-level evoked activities. This serves two complementary functions: finding latent activities that are compatible with a presented stimulus (inference) and adjusting synaptic weights to improve predictions (learning).
>
> Mathematically, predictive coding can be described as a special case of variational inference (Jordan et al., 1999; Friston, 2005; Marino, 2022). Assuming a Gaussian distribution for the generative model and approximate posterior, predictive coding infers a latent activity z^* that (locally) minimizes the following loss via gradient descent
>
> $$\mathcal{L}_{\text{pred}} = \|\boldsymbol{x} - G(z)\|^2 + \|z - \boldsymbol{\mu}_z\|^2 \,, \tag{1}$$
>
> where $\boldsymbol{x}$ represents sensory input, z the latent activity of the network, G a (deep) generative network and $\boldsymbol{\mu}_z$ the mean of the latent prior. Intuitively, predictive coding thus infers latent activities z^* by minimizing the reconstruction error between the generated prediction $G(z)$ and the actual sensory input $\boldsymbol{x}$ (first term of Equation (1)). The second term reflects the prior and can be interpreted as a regularization term that can implement activity constraints, such as sparsity (Rao and Ballard, 1999).
>
> Once latent activities are inferred, a gradient step with respect to the parameters of the generator G is taken on Equation (1), further reducing the reconstruction error between actual inputs z and predicted inputs $G(z)$. These separate optimization steps assume that synaptic weight changes (and thus, learning) occur on a slower timescale than inference.

In perception, an efficient way to represent relevant aspects of the sensorium is to try to "explain away" evoked activities in lower sensory cortex through a cascade of predictions performed by cortical feedback pathways (Helmholtz, 1878; Clark, 2013). These predictions reflect what the brain already knows about the sensorium. Informally, one may consider them emerging from a set of priors acquired through experience. By trying to match the sensory evoked signal, the brain thus seeks for latent causes that would best characterize the stimulus, such as its semantic category. Through these processes, brains can learn organized semantic representations (Friston, 2010; Clark, 2013).

These ideas have been formalized by computational frameworks such as predictive coding (Figure 1c; Rao and Ballard, 1999), which describes how neuronal dynamics and synaptic plasticity are both involved in learning generative models of the environment. First, on short time scales, neuronal activities change to better predict the evoked activity. Second, on longer time scales, synaptic plasticity aims to further improve these predictions (Box 1). Over time, minimization of prediction errors through these dynamics implicitly organizes latent representations (Lotter et al., 2017). In computational models, the prediction error is usually a point-wise measure at an appropriately coarse graining of the stimulus. For example, for images, the prediction error is typically measured at a pixel level. In the brain, it is thought to be computed by subclasses of layer 2/3 pyramidal neurons (Mumford, 1992; Bastos et al., 2012; Shipp, 2016) that compare activities between predictive neurons and neurons activated by sensory inputs (Keller and Mrsic-Flogel, 2018).

Due to the hierarchical nature of the predictive coding framework, neuronal populations develop selectivity for low-level details in early areas and for high-level properties (objects, shapes, scenes) in later areas, compatible with

experimentally measured neuronal responses and receptive fields (Rao and Ballard, 1999). Moreover, building models based on these principles leads to the emergence of latent representations suitable for efficiently learning downstream tasks (Rumelhart et al., 1986; Kingma and Welling, 2013; Lotter et al., 2017; Millidge et al., 2021). Predictive processing principles thus suggest a computational model compatible with cortical structure and activity for learning in brains (but see Koch and Poggio, 1999; Murray et al., 2004) from a simple goal: predicting sensory-evoked low-level cortical activities.

2.2. Beyond the prediction of sensations

As soon as we reduce external sensory inputs, for instance through unfocusing our eyes, meditation, or deep rest, we can become aware of virtual experiences continuously produced by our brain (Mildner and Tamir, 2019). These manifest in their strongest form as dreams, mostly occurring during the rapid-eye-movement (REM) phase of sleep (Nir and Tononi, 2010). While dreams may feel familiar, they often represent objects, scenes, situations that go beyond what we previously experienced (Fosse et al., 2003; Wamsley, 2014). Indeed, during REM dreams previous waking experiences are often not identically recalled but rather incorporated with other past memories into a new narrative (Fogel et al., 2018; Northoff et al., 2023).

The predictive processing framework has been previously suggested to also account for the phenomenology of virtual experiences during sleep (Hobson and Friston, 2012; Hobson et al., 2014). Accordingly, the same feedback pathways employed to generate predictions of sensory inputs during wakefulness, are "freed" from sensory inputs during sleep, allowing the generation of virtual experiences. This has been proposed to contribute to minimizing the generative model's complexity, i.e., the degrees of freedom required to describe the sensorium, for example by pruning redundant synapses. Consequently, dreaming would facilitate the ability to generalize and understand the semantics of the external world.

Here, we suggest that learning from virtual experiences has additional roles besides minimizing complexity. As we will describe in the following, learning during offlines states can further improve the predictive model by increasing its realism, can sharpen our ability to distinguish between internally generated and externally driven activities, and robustify semantic neuronal representations against perturbations, such as occlusions of parts of the visual field. To this end we discuss two complementary approaches, "adversarial dreaming" and "contrastive dreaming", which, as we will explain, are crucial for organizing neuronal representations.

3. Learning representations by creating virtual experiences

3.1. Principles of adversarial learning

Box 2 - Generative adversarial networks

Generative Adversarial Networks (GANs, Goodfellow et al., 2014) introduce a generator G that generates data samples from noise, and a binary classifier, or discriminator (D), that distinguishes these generated samples from real data. The generator G is trained to fool the discriminator D into believing that generated samples are real by creating samples that belong to the data distribution. For a sample from the data distribution $x \sim p(x)$ and a noise vector sampled from the prior distribution, e.g., $p(z) \sim \mathcal{N}(0, I)$, the objective of the discriminator D is to *minimize* the loss

$$\mathcal{L}_{\text{adv}} = -\log D(x) - \log(1 - D(G(z))), \tag{2}$$

while the objective of the generator G is to *maximize* this loss. This equation defines the cross-entropy loss for a binary classifier (D) with a sigmoid output, where the label is 1 for all data samples x, and 0 for all generated samples $G(z)$. Thus, the discriminator improves its ability to discern real from generated samples, while the generator improves the quality of its generated samples so it can fool the (improved) discriminator. After sufficiently many training steps, the generator is 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).
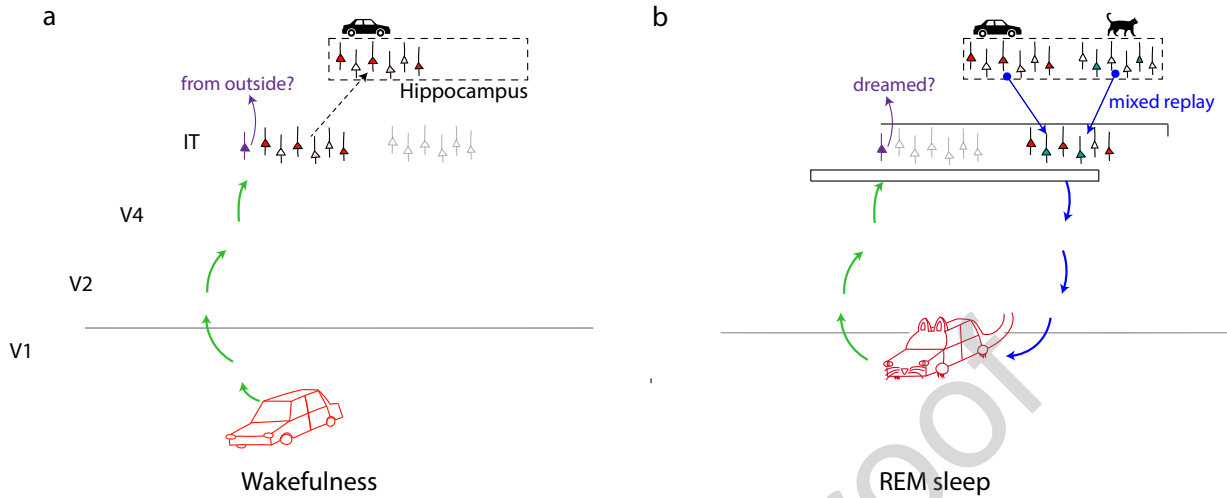

Figure 2: **Learning representations via adversarial dreaming. (a)** During wakefulness, external stimuli are processed from V1 to IT cortex along feedforward pathways (green). These learn to recognize the induced early sensory activity as coming from outside (purple neuron). Simultaneously, latent representations are stored in the hippocampus. **(b)** During REM sleep several independent memories are replayed from the hippocampus and combined in high-level areas. Feedback pathways (blue) map this latent activity to early sensory areas where virtual experiences (dreams) are generated. Following the principles of adversarial learning, feedforward pathways (green) learn to distinguish virtual from stimulus-evoked low-level activities, while feedback pathways improve the generative process to make this distinction harder.

Cortical models that aim to also learn from virtual experiences, such as dreams, cannot rely exclusively on sensory prediction errors due to the absence of ground-truth evoked activities, but have to find alternative sources of learning signals. For example, such models could learn to produce data that appears "similar" to previous stimuli without trying to exactly reproduce them. But how can we quantify this similarity to derive useful learning signals? One possibility is to expand the model with an additional module that learns to measure the similarly between generated data and actual stimuli. In this spirit, Generative Adversarial Networks (GANs, Goodfellow et al., 2014) introduce an architecture that consists of two networks: a generator producing virtual samples and a discriminator judging whether a sample is real or generated. These two networks are trained adversarially, with the discriminator learning to distinguish generated from real samples while the generator learns to fool the discriminator by improving the realism of the generated samples (Box 2). Through this adversarial game, the generator gradually learns to synthesize samples that are similar to the training data. This process can be illustrated by a student (generator) that tries to fake their parents' handwriting, and the teacher (discriminator) that detects whether the writing is real or fake. After many attempts, the student learns how to fool their teacher by writing in a way that is hardly distinguishable from their parents' handwriting.

The goal of this process is to find a balanced optimum in which the generator produces a large variety of samples with sufficient similarity to real samples. For example, in the context of natural image generation, generated samples contain colors, shapes, and objects that are typically present in real images. However they can also be distorted or combined versions of these objects, as a consequence of adversarial learning, as even after convergence the generator keeps the freedom to generate creative samples (Brock et al., 2019).

Furthermore, GANs are known to extract semantic latent representations from data (Radford et al., 2015; Donahue et al., 2016; Donahue and Simonyan, 2019). Intuitively, this originates from the optimization-induced organization of the GANs' latent space where nearby points lead to images that are semantically similar (Brock et al., 2019). This smoothness generalizes when interpolating between distant points in the latent space: generated samples exhibit smooth transitions from one sample to another, creating new objects that can combine features from multiple distinct objects (Berthelot et al., 2018; Brock et al., 2019). Exploiting this learned structure, several models invert the generative process of GANs and demonstrate that their latent space contains semantic representations that can be useful to perform downstream tasks (Makhzani et al., 2015; Dumoulin et al., 2017; Donahue and Simonyan, 2019).

3.2. Adversarial dreaming

Adversarial learning principles have been hypothesized to allow the brain to learn semantic representations from virtual experiences, such as creative dreams, typically occurring during REM sleep (Deperrois et al., 2022). In this study, the authors[1] propose a cortical architecture with a feedback pathway that generates activity in early sensory cortex from high-level representations. Additionally, they introduce a feedforward pathway that determines whether activity in lower sensory cortices is externally driven or internally generated. Feedforward pathways thus assume the role of the discriminator in GANs and are additionally shaped through predictive learning , being simultaneously trained to infer latent representations from low-level activities (Figure 2a). Latent representations inferred during wakefulness are stored in a simple hippocampus model allowing storage and replay. When a hippocampal memory is retrieved, feedback pathways are reactivated and generate the associated sensory input.

Learning in this model is organized across three different physiological phases, wakefulness, non-rapid eye movement (NREM) and rapid eye movement (REM) sleep, each characterized by a different objective. During REM sleep, two different representations from previously observed stimuli are retrieved and together with cortical background activity (Spanò et al., 2020) generate a creative dream through feedback pathways. These dreams thus contain elements from both stored memories (Figure 2b). To improve the realism of these virtual experiences, feedback pathways are trained to adversarially fool the FF discriminator into believing that the activity in early sensory areas is externally driven. This process defines "adversarial dreaming". Formally, adversarial dreaming is minimizing the classical objective functions of GANs (Box 2) during Wake and REM phases via synaptic weight changes implementing stochastic gradient descent.

The results from this model suggest that REM creative dreams, generated through adversarial dreaming, become more realistic over learning, but still remain different and novel as compared to external sensory inputs (Deperrois et al., 2022), in line with dream phenomenology (Nir and Tononi, 2010; Scarpelli et al., 2019). Crucially, generating these virtual experiences through both memory combinations and adversarial learning improves the quality of the learned cortical representations. Indeed, the authors show that object categories can easily be extracted from the latent activity using a linear classifier. Additionally, they demonstrate that this ability is significantly impaired when they artificially inhibit REM sleep during training. The authors thus conclude that creative dreams are a key ingredient for the acquisition of semantic latent representations.

3.3. Neuronal and behavioral correlates

The principle of adversarial dreaming leads to neuronal and behavioral consequences that can be investigated experimentally.

A central feature of the framework is that creative dreams during offline states, such as REM sleep, are crucial for the emergence of organized cortical representations. This could be tested by recording neuronal population activity in high-level areas using multielectrode arrays. From these recordings, one could quantify how well neuronal representations separate object categories, either by training a linear classifier on these representations (Hung et al., 2005) or by computing the representation dissimilarity matrices between stimuli (Yamins et al., 2014). We expect that in subjects who are chronically deprived of REM sleep, such as with antidepressant drugs (Palagini et al., 2013), optogenetic inhibition (Boyce et al., 2016; Aime et al., 2022), or that lack the ability to form mental images (aphantasia, Zeman et al., 2015; Pearson, 2019), representations are less semantically organized than for control subjects. Behaviorally, this would translate as a slower learning speed of novel object classification tasks.

Considering the similarities between mental imagery, imagination and dreaming (Kahan et al., 1997; Llewellyn, 2016a; Pearson, 2019), one could use mental imagery as a practical alternative to dreaming for studying the impact on learning and representation of novel objects in humans. A potential experiment would involve asking human subjects to classify novel 3D objects and monitoring their learning progress. Human subjects could be asked to perform mental imagery training sessions following the presentation of novel objects, for instance by mentally rotating them. We predict that participants who performed these mental tasks would perform better at categorizing these novel objects than the control participants.

Furthermore, in adversarial dreaming, internal activity in early sensory areas becomes more similar to evoked activity over the course of learning, which suggests that dreams should become more realistic with age. This correlates

---

[1] when referring to Deperrois et al. (2022), the authors, although identical to the present paper, are referred to in third person

with dream reports over different stages of life, that are initially unstructured and plain, and gradually become more meaningful, narrative and less bizarre (Nir and Tononi, 2010; Scarpelli et al., 2019). According to the theory, this may reflect that older persons know more about the structure of the world and its limitations, and thus become more conservative and less prone to exploration, reducing their capacity to learn new concepts. On a neuronal level, this corresponds to an increasing similarity between stimulus evoked and REM generated activity in lower sensory areas. In this line, previous work has demonstrated that spontaneous activity, potentially driven by creative daydreaming, indeed becomes more similar to evoked activity in ferret visual cortex over the course of development (Berkes et al., 2011).

Finally, in terms of cortical structure, adversarial dreaming predicts a functional organization into two effectively separate feedforward and feedback streams. If the information is not forced to go up and down the whole hierarchy, shortcuts between higher cortical areas will prevent lower cortical areas to learn useful features. Even though cross-projections between feedforward and feedback pathways are observed experimentally (Gilbert and Li, 2013), adversarial dreaming predicts that those are effectively gated off during essential periods for organizing neuronal representations.

3.4. Creativity and adversarial dreaming

As a consequence of adversarial dreaming, new virtual experiences can be generated by randomly combining different memories (Fig. 2b). This thus leads to the generation of low-level activities that are unlikely to have been evoked by previously experienced stimuli, but that nevertheless may be part of the external world. While the main focus of this article is to suggest a role of virtual experiences on learning, such a phenomenon suggests two additional functional benefits that are important to mention.

First, by learning to encode these novel experiences, the system prepares for a future where these imagined sensations are encountered in the wild, such as simulating a dangerous situation offline to escape from it faster when it actually occurs (cf. Hobson, 2009; Llewellyn, 2016b). Furthermore, generating "semantic superpositions" and exposing the feedforward pathways to these may equip the agent with the ability to quickly recognize new stimuli as a composition of known components, making its reaction to them significantly simpler, such as an electric bike leveraging our knowledge about engines and bikes. In a behavioral experiment, one could investigate whether participants viewing novel stimuli composed of known parts identify their related categories faster after REM sleep.

Second, novel adversarially generated experiences could provide an unexpected solution to a specific problem the agent is facing. During REM sleep, the agent may hence experience an "insight" suggesting how to solve a complex problem (also see Friston et al., 2017), such as the Benzene structure that was discovered through a dream by Kekulé (Mazzarello, 2000). In this line, generative models are now used in the field of drug discovery to circumvent the limitation of traditional approaches relying upon domain knowledge from physics and chemistry to construct synthesis rules. In particular, GAN-based frameworks such as adversarial autoencoders have been used to extend the search of possible molecules for drug design, generating compounds with desired molecular properties (Guimaraes et al., 2017; De Cao and Kipf, 2018; Blanchard et al., 2021). Naturally, not all creative combinations experienced during dreaming are useful, and their usefulness is ultimately determined by how compatible they are with the actual external world. This suggests that additional steps may be necessary, such as testing experimentally the existence of suggested compounds.

More broadly, creative dreaming closely relates to concepts of how to trigger creative thoughts (Llewellyn, 2016a): After being intensively exposed to a certain topic, one needs periods of rest, or "incubation" periods, to freely let the generator produce samples. Experimentally, this could be tested by evaluating the performance of participants at a creative synthesis task (Finke et al., 1996; Palmiero et al., 2015), consisting of combining different visual patterns into a new, potentially useful object. Subjects that have chronically impaired REM sleep would be less likely to synthesize useful/realistic objects (Giancola et al., 2022). The adversarial dreaming framework thus expands the predictive processing view to the offline generation of creative, virtual experiences that could facilitate the acquisition of semantic representations. We next introduce another unsupervised learning principle that offline states could leverage.

4. Learning representations by contrasting sensory experiences

4.1. Principles of contrastive learning

Box 3 - Contrastive learning

Contrastive learning algorithms use an encoder that is trained to compare (latent) representations of data samples. These representations are shaped by pulling together representations of semantically similar inputs and pushing apart those of dissimilar inputs (Jaiswal et al., 2020; Le-Khac et al., 2020). Similar (positive) examples are usually obtained by applying a series of (semantically preserving) data augmentations such as cropping, resizing, blur, color distortion to a given sample (Chen et al., 2020), and negative examples are simply other samples from the dataset. This comparison can be learned with a loss function $\mathcal{L}_{\text{contr}}$ defined on a single positive pair (i, j) and a large number of negative pairs $(i, k)_{k \neq i}$ such as:

$$\mathcal{L}_{\text{contr}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1, k \neq i}^{2N} \exp(\text{sim}(z_i, z_k)/\tau)}, \tag{3}$$

where where N is the number of examples in a minibatch ($2N$ because all examples are augmented), z_i is the representation of the sample k, $\text{sim}(u, v) = u^T v / (\|u\| \|v\|)$ denotes the dot product between l_2 normalized u and v and τ denotes the temperature parameter (Chen et al., 2020), i.e., this loss function computes the cosine of the angle between u and v. Through this learning objective, the network aims to reduce the distance between the representations of positive pairs (z_i, z_j) and increase the distance between the representations of negative pairs $(z_i, z_k)_{k \neq i}$.

Ultimately, the idea of semantic latent representations is to have similar latent neuronal responses to semantically similar stimuli (DiCarlo et al., 2012). Instead of learning representations implicitly via generative models, one could directly train feedforward pathways to map semantically similar inputs to similar latent representations, and dissimilar inputs to different regions of the latent space (e.g., Le-Khac et al., 2020). In this context, one often refers to similar ("positive") examples as being "pulled together" and dissimilar ("negative") examples as being "pushed apart" from each other during the training process.

How are positive examples obtained during training, before the network has had the chance to organize its representations? Typically, positive examples are created by transforming an existing sample through so-called data augmentations, consisting of cropping, color distorting, or blurring the sample (Chen et al., 2020). Through this transformation process, the input remains semantically similar to the original input, while its sensory structure can be vastly different.

Negative examples serve to prevent trivial solutions, such as mapping all samples to the same latent vector, often referred to as "representational collapse" (Le-Khac et al., 2020; Bardes et al., 2021) . For a given sample, all the other samples from the data set are typically considered negative examples. Even though this broad definition includes samples from the same category, which can not be excluded in the absence of labels, the majority of negative samples will come from a different category for typical datasets. Note that recent work suggests that negative examples may not be required for efficient contrastive learning. Alternative methods include ensuring that representations are variable enough (Bardes et al., 2021) or breaking the symmetry between projections of positive examples (Grill et al., 2020; Chen and He, 2021). Through simple yet effective principles, contrastive learning led to models that are currently state-of-the-art at learning semantic representations useful for downstream tasks in an unsupervised manner (Liu et al., 2021; Ericsson et al., 2022).

4.2. Contrastive dreaming

Contrastive learning principles may be leveraged by the brain to enhance and robustify neuronal representations during imagination and dreaming. Just like for adversarial dreaming, the generative model learned by feedback pathways from the prediction of sensory inputs during wakefulness can be leveraged for learning during sleep. In contrastive dreaming, only a single hippocampal memory serves as the basis for subsequent generation of activity in early sensory cortex. Instead of combining multiple stored memories, the virtual experiences thus represent previously
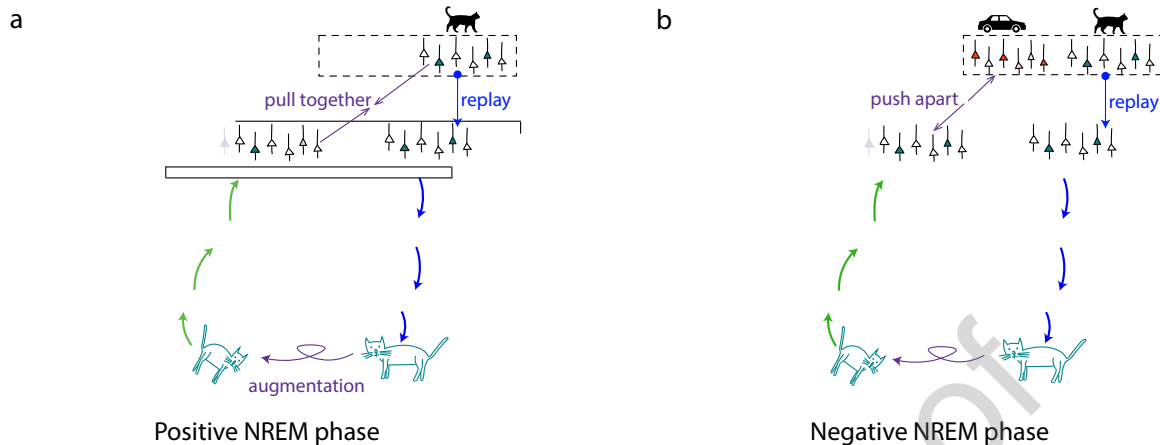
Figure 3: **Learning representations via contrastive dreaming. (a)** During NREM single memories are replayed and generate activities in V1, which are modified in a semantically preserving way, e.g., rotating or squeezing the input. Following the principles of contrastive learning, feedforward pathways learn to map this "augmented input" to the same latent representations as the initially replayed hippocampal representation. **(b)** To implement the contrastive step during NREM, feedforward pathways learn to push apart the inferred representation to a different hippocampal memory, serving as a negative example.

observed sensory inputs that are altered by a series of augmentations. These augmentations need to be strong enough to change the low-level details of the virtual experience, but not so strong as to change its semantic content, e.g., adding noise to, blurring, cropping, rotating or distorting an image (Figure 3a). These augmentations could be applied by leveraging an additional cortical module, or by directly influencing generation through modulation of feedback pathways at different hierarchical levels (Karras et al., 2018; Wybo et al., 2023). The goal of feedforward pathways then consists of mapping this altered input to its original hippocampal representation, thus pulling together positive pairs (Figure 3a).

Negative examples are provided by older memories, with feedforward pathways learning to map the augmented experience away from these (Figure 3b). Cortically, this could occur after the positive phase, by maintaining the inferred latent representation and comparing it to other hippocampal memories.

This approach was partly explored by Deperrois et al. (2022). During the NREM phase of the model, virtual experiences generated from single hippocampal memories were partly occluded. This process made the feedforward network more robust to similar perturbations during perception. We hypothesize that by extending the model to additional augmentations, and contrasting it with negative examples (Figure 3b), such a phase could further improve the semantic organization of the model's latent space. In summary, we propose that the efficiency of contrastive learning objectives can be exploited by offline states through contrastive dreams of previous experiences.

4.3. Neuronal and behavioral correlates

The contrastive dreaming framework can be experimentally investigated. First, it makes predictions about dream phenomenology. One can assess the diversity of internally generated experiences by waking up sleeping participants at different physiological stages, such as NREM, REM, hypnagogic or day-dreaming states (Waters et al., 2016), and asking them to report the content from their dreams, or possibly by directly communicating with them while dreaming (Konkoly et al., 2021). We predict that dreams reported from the hypothesized "contrastive" states, such as within sharp-wave ripples during NREM sleep (Kudrimoti et al., 1999), tend to contain individual previous experiences. Depending on the detail of the dream reports, they may even reveal the suggested augmentations, for example in the form of distorted colors or reversed directions. In contrast, dreams from adversarial dreaming during REM sleep would be dissimilar to previous experiences but rather combine diverse elements from them (Fogel et al., 2018). These predictions are line with experimental data showing that NREM dream reports have more episodic memory sources (Baylor and Cavallero, 2001) and exhibit less complexity (Martin et al., 2020).

Second, contrastive dreaming makes prediction at the neuronal level. During wakefulness, as sensory inputs that are nearby in time usually involve the same object under different views (Illing et al., 2021), one could expect that over

9

the course of development, activities in high sensory areas of the ventral stream of the visual cortex become increasingly stable during the observation of a moving object. This could be measured by the representation dissimilarity matrix approach (Yamins et al., 2014) tracking high-level activities over time. We would thus expect impaired NREM sleep to lead to more variable neuronal representations. More generally, the contrastive dreaming principle predicts that while activities from lower areas are very different across stimulus categories and augmentations, high-level activities should become robust against augmentations but remain sensitive to stimulus categories. Additionally, one could compare low-level and high-level activities during NREM sleep. We predict that while high-level activities resemble waking activities closely due to hippocampal replay, low-level activities vary significantly due to augmentations (Figure 3a).

5. Learning beyond the shackles of direct experiences

5.1. Summary

To explain the emergence of semantic neuronal representations in an autonomous, unsupervised manner, influential neuroscientific theories suggest that the brain minimizes prediction errors between its expectations and stimulus-evoked activities (Rao and Ballard, 1999; Friston, 2005; Millidge et al., 2021; Mikulasch et al., 2022). Models emerging from these frameworks are successful at describing various properties of cortical networks and can solve complex computational tasks. However, the rich, sometimes bizarre world of non-sensory related phenomena our brains experience on a nightly basis appear only to reduce the complexity of the generative model. To complement predictive processing theories, here we discussed two computational frameworks through which brains can benefit even more from their internally generated virtual experiences.

First, adversarial dreaming combines several stored memories with cortical noise and pits feedback and feedforward pathways against each other in a creative game of generating and discriminating low-level activities. This process thereby implicitly learns an organized latent structure. Second, contrastive dreaming explicitly trains feedforward networks to map semantically similar inputs to similar high-level cortical representations by dreaming up previously observed inputs with semantically preserving augmentations. Both principles are compatible with the bidirectional architecture of sensory cortices and could be implemented in network models relying on biologically plausible credit assignment algorithms and learning rules (Richards et al., 2019; Lillicrap et al., 2020).

While our principles primarily pertain to dreaming, we anticipate their applicability to other forms of spontaneous, virtual experiences like mental imagery (Pearson, 2019), meditation (Cooper et al., 2022), and spontaneous thoughts (Mildner and Tamir, 2019). The key distinction lies in the nature of these experiences: adversarial dreaming involves the creative recombination of memory elements, whereas contrastive dreaming reenacts "augmented" past experiences. We propose that virtual experiences during dreamlike states generally align with one of these two mechanisms.

5.2. Outlook: learning from predictive, adversarial and contrastive principles

Despite their algorithmic differences, the three presented learning principles can be implemented by the same cortical architecture. They however require different physiological phases, in line with previous theories (Hinton et al., 1995; Giuditta et al., 1995; Hobson and Friston, 2012; Lewis et al., 2018).

An interesting direction would be to explore whether the combined optimization of different learning objectives could have a synergistic effect on the acquisition of semantic representations. While the combination of predictive and adversarial learning has been previously explored (Makhzani et al., 2015; Brock et al., 2017; Ulyanov et al., 2017), the benefits of combining contrastive and adversarial principles remain to be elucidated (but see Chen et al., 2019; Deperrois et al., 2022).

Finally, humans develop under *some* supervision, for example in the form of explicit verbal instructions about object categories. It is hence natural to explore the combination of the unsupervised learning principles described so far with sparse labels to further improve the learned latent structure (see also Deperrois et al., 2022).

Since these principles are in many ways complementary, experimentally the influence of each may be challenging to tease apart. However, despite their similarities, they have different functional goals. Predictive processing allows the brain to predict upcoming stimuli, adversarial dreaming aims to prepare the brain for previously unobserved stimuli, and contrastive dreaming aims to make latent representation invariant to irrelevant factors of variation. Through

10

carefully designed experiments, for instance analyzing the individual effects of NREM and REM sleep on cortical dynamics (Tamaki et al., 2020), or by analyzing the effect of different learning tasks on NREM and REM activity patterns (Fogel and Smith, 2006; Fogel et al., 2007) their different functional goals may hence be exploited to tease apart their influences on neuronal representations.

5.3. Relation to previous work

Gershman (2019) proposed an adversarial framework for brain function in view of psychological and neural evidence. In particular, he discusses the consequences of a dysfunctioning discriminator on the perception of hallucinations, leading to potential delusions observed in mental disorders. The framework discussed here may serve as a suggestion for implementing a mechanistic model of these ideas and further elucidate the consequences of dysfunction in specific modules.

Previous work suggested an alternative explanation for the creative aspect of dreams during REM sleep. The pioneer activation-synthesis theory from Hobson and McCarley (1977) suggests 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". Adversarial dreaming provides a concrete instantiation of this idea by forming coherent REM dreams from incoherent signals. From replaying a random mixture of episodes out of the hippocampus to the cortex, the discriminator network provides the feedback to increase the realism of the generated dream imagery. Other authors attribute this creative phenomenon to a shift of topographical neural activity towards the Default Mode Network (Domhoff and Fox, 2015), encouraging external inputs from wakefulness to be integrated with internally generated imagery, jointly manifesting as bizarre dream content (Northoff et al., 2023).

Generative modeling's early developments, notably the Wake-Sleep algorithm (Hinton et al., 1995), previously emphasized the importance of offline processes in optimizing latent representations. First, the sleep phase in the Wake-Sleep algorithm, while conceptually different, is algorithmically akin to the positive phase of 'contrastive dreaming,' where generated inputs are aligned with their originating latent activities via the feedforward network. Second, this algorithm introduced a bidirectional structure of cortical projections, where feedforward pathways encode sensory inputs and feedback pathways generate sensory predictions or fictive inputs. This concept later influenced the development of variational autoencoders (Kingma and Welling, 2013); their biological plausibility was recently examined in (Marino, 2022). The introduced framework here also leverages a bidirectional organization necessary to implement both adversarial and contrastive dreaming paradigms. In this view, backward projections serve a dual role of making predictions during wakefulness, while generating virtual experiences during offline states.

Previous work on predictive processing suggests that its principles extend beyond waking experiences (Hobson and Friston, 2012; Hobson et al., 2014). This theory rests on the free-energy principle, which formulates processing and learning as a variational optimization problem. Intuitively speaking, a good model should provide a good explanation of observed data ("model accuracy"), while maintaining a minimal set of assumptions ("model complexity"), together maximizing "model evidence" (Friston, 2010). Accordingly, wakefulness provides an opportunity to optimize both of these components, while dreams, or more generally offline states, specifically allow for the reduction of model complexity. Indeed, such ideas have been successfully employed for machine learning models (Ponnapalli et al., 1999; Simoncelli and Olshausen, 2001; Williams, 1995) and suggested to provide a functional explanation for synaptic homeostasis during sleep (Tononi and Cirelli, 2014): minimizing the brain's model complexity may improve generalization abilities.

Similar to predictive processing, adversarial dreaming also aims to maximize model evidence, though implicitly with the help of feedforward pathways, rather than explicitly (Huszár, 2017). Nevertheless, this similarity in spirit suggests that adversarial dreaming too could benefit from the reduction of complexity as suggested by Hobson et al. (2014). Vice versa, predictive processing could benefit from the ability of feedforward pathways being able to distinguish between internally generated and externally driven activities in sensory cortex, learned via adversarial dreaming. While predictive processing learns semantic representations implicitly, contrastive dreaming explicitly optimizes these behaviorally relevant variables. Nevertheless, the neuronal representations emerging from contrastive learning may also help generative models to maximize model evidence. These observations suggest an intimate relation between these theories, jointly highlighting the importance of virtual, non-sensory, experiences.

Two recent studies suggested how the brain could benefit from constrastive objectives. In Illing et al. (2021), the authors propose that positive examples are obtained from the observation of a moving object, while negative examples

11

appear through saccades towards new objects. In contrast, Halvagal and Zenke (2022) argue that the brain does not need negative examples, as long as latent activities are encouraged to remain sufficiently variable (through variance maximization, Bardes et al., 2021). Through this mechanism, networks learn invariant representations for stimulus features which change slowly in time. A downside from these models is that positive examples, and thus augmented inputs, are assumed to be obtained through the observation of moving objects, leading to limited augmentation. However, a series of strong augmentations are needed to obtain strong semantic representations via contrastive learning objectives (Chen et al., 2020). To avoid interference of such strong augmentations with perception, hosting them during offline states as suggested by the contrastive dreaming principle thus provides a beneficial alternative.

5.4. Conclusion: the necessity of virtual experiences for learning

We proposed that essential processes shaping our cortical function arise from brains generating virtual experiences during sleep. Learning from an imagined world may thus be just as important as learning from sensations. This view significantly expands our perspective on perception and learning. Do we need to be constantly focused on our sensorium to learn optimally, or can we finally justify sometimes having our heads in the clouds?

Bibliography

Aime, M., Calcini, N., Borsa, M., Campelo, T., Rusterholz, T., Sattin, A., Fellin, T., and Adamantidis, A. (2022). Paradoxical somatodendritic decoupling supports cortical plasticity during REM sleep. *Science*, 376(6594):724–730.

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

Barlow, H. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12(3):241.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711.

Baylor, G. W. and Cavallero, C. (2001). Memory sources associated with rem and nrem dream reports throughout the night: a new look at the data. *Sleep*, 24 2:165–70.

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]*.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94(2):115.

Blanchard, A. E., Stanley, C., and Bhowmik, D. (2021). Using GANs with adaptive training data to search for new molecules. *Journal of Cheminformatics*, 13(1):1–8.

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.

Brock, A., Donahue, J., and Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. *arXiv:1809.11096 [cs, stat]*.

Brock, A., Lim, T., Ritchie, J. M., and Weston, N. (2017). Neural photo editing with introspective adversarial networks. *arXiv:1609.07093 [cs, stat]*.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR.

Chen, T., Zhai, X., Ritter, M., Lucic, M., and Houlsby, N. (2019). Self-supervised GANs via auxiliary rotation loss. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12146–12155, Long Beach, CA, USA. IEEE.

Chen, X. and He, K. (2021). Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, Nashville, TN, USA. IEEE.

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

Cooper, A. C., Ventura, B., and Northoff, G. (2022). Beyond the veil of duality—topographic reorganization model of meditation. *Neuroscience of Consciousness*, 2022(1):niac013.

De Cao, N. and Kipf, T. (2018). MolGAN: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*.

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.

Domhoff, G. W. and Fox, K. C. (2015). Dreaming and the default network: A review, synthesis, and counterintuitive research proposal. *Consciousness and Cognition*, 33:342–353.

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

Donahue, J. and Simonyan, K. (2019). Large scale adversarial representation learning. *Advances in Neural Information Processing Systems*, 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]*.

Ericsson, L., Gouk, H., Loy, C. C., and Hospedales, T. M. (2022). Self-Supervised Representation Learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62.

Finke, R. A., Ward, T. B., and Smith, S. M. (1996). *Creative Cognition: Theory, Research, and Applications*. MIT Press.

Fogel, S. M., Ray, L. B., Sergeeva, V., De Koninck, J., and Owen, A. M. (2018). A novel approach to dream content analysis reveals links between learning-related dream incorporation and cognitive abilities. *Frontiers in Psychology*, 9:1398.

Fogel, S. M. and Smith, C. T. (2006). Learning-dependent changes in sleep spindles and stage 2 sleep. *Journal of Sleep Research*, 15(3):250–255.

Fogel, S. M., Smith, C. T., and Cote, K. A. (2007). Dissociable learning-dependent changes in rem and non-rem sleep in declarative and procedural memory systems. *Behavioural Brain Research*, 180(1):48–61.

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.

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.

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., and Ondobaka, S. (2017). Active inference, curiosity and insight. *Neural Computation*, 29(10):2633–2683.

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

Giancola, M., Palmiero, M., Piccardi, L., and D'amico, S. (2022). The relationships between cognitive styles and creativity: The role of field dependence-independence on visual creative production. *Behavioral Sciences*, 12(7):212.

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.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, 27.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284.

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.

Guimaraes, G. L., Sanchez-Lengeling, B., Outeiral, C., Farias, P. L. C., and Aspuru-Guzik, A. (2017). Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. *arXiv preprint arXiv:1705.10843*.

Halvagal, M. S. and Zenke, F. (2022). The combination of hebbian and predictive plasticity learns invariant object representations in deep sensory networks. *bioRxiv*.

Helmholtz, H. (1878). The facts in perception. In Kahl, R., editor, *Selected Writings of Hermann Helmholtz*. Wesleyan University Press.

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

Hobson, J. and Friston, K. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology*, 98(1):82–98.

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*.

Hubel, D. H. and Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, 28(2):229–289.

Huber, L. S., Geirhos, R., and Wichmann, F. A. (2021). A four-year-old can outperform ResNet-50: Out-of-distribution robustness may not require large-scale experience. In *SVRHM 2021 Workshop @ NeurIPS*.

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.

Huszár, F. (2017). Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*.

Illing, B., Ventura, J., Bellec, G., and Gerstner, W. (2021). Local plasticity rules can learn deep representations using self-supervised contrastive predictions. *Advances in Neural Information Processing Systems*, 34:30365–30379.

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.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.

Kahan, T. L., LaBerge, S., Levitan, L., and Zimbardo, P. (1997). Similarities and differences between dreaming and waking cognition: An exploratory study. *Consciousness and Cognition*, 6(1):132–147.

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

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

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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

Konkoly, K. R., Appel, K., Chabani, E., Mangiaruga, A., Gott, J., Mallett, R., Caughran, B., Witkowski, S., Whitmore, N. W., Mazurek, C. Y., et al. (2021). Real-time dialogue between experimenters and dreamers during rem sleep. *Current Biology*, 31(7):1417–1427.

Kudrimoti, H. S., Barnes, C. A., and McNaughton, B. L. (1999). Reactivation of Hippocampal Cell Assemblies: Effects of Behavioral State, Experience, and EEG Dynamics. *The Journal of Neuroscience*, 19(10):4090–4101.

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

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.

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

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]*.

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.

Lotter, W., Kreiman, G., and Cox, D. (2017). Deep predictive coding networks for video prediction and unsupervised learning. In *International Conference on Learning Representations*.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. (2015). Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

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

Martin, J. M., Andriano, D. W., Mota, N. B., Mota-Rolim, S. A., Araújo, J. F., Solms, M., and Ribeiro, S. (2020). Structural differences between REM and non-REM dream reports assessed by graph analysis. *PLOS ONE*, 15(7):e0228903.

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

Mikulasch, F. A., Rudelt, L., Wibral, M., and Priesemann, V. (2022). Where is the error? hierarchical predictive coding through dendritic error computation. *Trends in Neurosciences*.

Mildner, J. N. and Tamir, D. I. (2019). Spontaneous thought as an unconstrained memory process. *Trends in Neurosciences*, 42(11):763–777.

Millidge, B., Seth, A., and Buckley, C. L. (2021). Predictive coding: a theoretical and experimental review. *arXiv preprint arXiv:2107.12979*.

Mumford, D. (1992). On the computational architecture of the neocortex: Ii the role of cortico-cortical loops. *Biological Cybernetics*, 66(3):241–251.

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.

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

Northoff, G., Scalabrini, A., and Fogel, S. (2023). Topographic-dynamic reorganisation model of dreams (trod)–a spatiotemporal approach. *Neuroscience & Biobehavioral Reviews*, page 105117.

Palagini, L., Baglioni, C., Ciapparelli, A., Gemignani, A., and Riemann, D. (2013). REM sleep dysregulation in depression: state of the art. *Sleep Medicine Reviews*, 17(5):377–390.

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.

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

Ponnapalli, P., Ho, K., and Thomson, M. (1999). A formal selection and pruning algorithm for feedforward artificial neural network optimization. *IEEE Transactions on Neural Networks*, 10(4):964–968.

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

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.

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.

Rodman, H. R. (1994). Development of inferior temporal cortex in the monkey. *Cerebral Cortex*, 4(5):484–498.

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

Scarpelli, S., Bartolacci, C., D'Atri, A., Gorgoni, M., and De Gennaro, L. (2019). Mental sleep activity and disturbing dreams in the lifespan. *International Journal of Environmental Research and Public Health*, 16(19):3658.

Shipp, S. (2016). Neural elements for predictive coding. *Frontiers in Psychology*, 7:1792.

Simoncelli, E. P. and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216.

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.

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.

Tamaki, M., Wang, Z., Barnes-Diana, T., Guo, D., Berard, A. V., Walsh, E., Watanabe, T., and Sasaki, Y. (2020). Complementary contributions of non-REM and REM sleep to visual learning. *Nature Neuroscience*, 23(9):1150–1156.

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.

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

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

14

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.

Williams, P. M. (1995). Bayesian regularization and pruning using a laplace prior. *Neural Computation*, 7(1):117–143.

Wybo, W. A., Tsai, M. C., Tran, V. A. K., Illing, B., Jordan, J., Morrison, A., and Senn, W. (2023). Nmda-driven dendritic modulation enables multitask representation learning in hierarchical sensory processing pathways. *Proceedings of the National Academy of Sciences*, 120(32):e2300558120.

Yamins, D. L., 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.

Zeman, A., Dewar, M., and Della Sala, S. (2015). Lives without imagery–congenital aphantasia. *Cortex*, 73:378–380.

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.