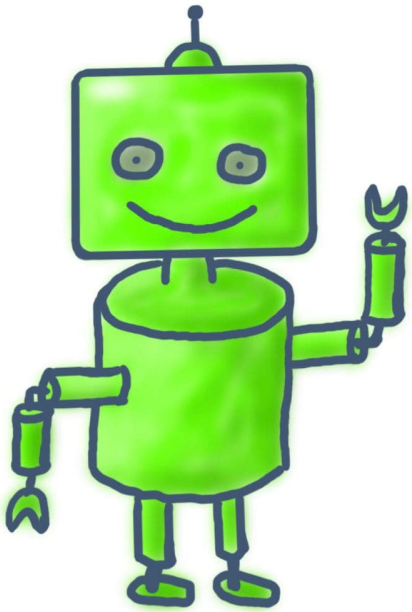
THE ROLE OF TASK AND ENVIRONMENT IN BIOLOGICALLY INSPIRED ARTIFICIAL INTELLIGENCE: LEARNING AS AN ACTIVE, SENSORIMOTOR PROCESS

VIVIANE CLAY



SCIENTIFIC SUPERVISORS:
Gordon Pipa
Kai-Uwe Kühnberger
ADDITIONAL SCIENTIFIC SUPERVISOR:
Peter König
NON-SCIENTIFIC SUPERVISOR:
Sabine König

The Role of Task and Environment in Biologically Inspired Artificial Intelligence: Learning as an Active, Sensorimotor Process

Dissertation

zur Erlangung des Grades eines Doktors der Naturwissenschaften
eingereicht am Fachbereich Humanwissenschaften
der Universität Osnabrück

vorgelegt von

Viviane Clay

Osnabrück, Oktober 2021

ABSTRACT

The fields of biologically inspired artificial intelligence, neuroscience, and psychology have had exciting influences on each other over the past decades. Especially recently, with the increased popularity and success of artificial neural networks (ANNs), ANNs have enjoyed frequent use as models for brain function. However, there are still many disparities between the implementation, algorithms, and learning environment used for deep learning and those employed by the brain, which is reflected in their differing abilities. I first briefly introduce ANNs and survey the differences and similarities between them and the brain. I then make a case for designing the learning environment of ANNs to be more similar to that in which brains learn, namely by allowing them to actively interact with the world and decreasing the amount of external supervision. To implement this sensorimotor learning in an artificial agent, I use deep reinforcement learning, which I will also briefly introduce and compare to learning in the brain.

In the research presented in this dissertation, I focus on testing the hypothesis that the learning environment matters and that learning in an embodied way leads to acquiring different representations of the world. We first tested this on human subjects, comparing spatial knowledge acquisition in virtual reality to learning from an interactive map. The corresponding two publications are complemented by a methods paper describing eye tracking in virtual reality as a helpful tool in this type of research. After demonstrating that subjects do indeed learn different spatial knowledge in the two conditions, we test whether this transfers to artificial agents. Two further publications show that an ANN learning through interaction learns significantly different representations of the sensory input than ANNs that learn without interaction. We also demonstrate that through end-to-end sensorimotor learning, an ANN can learn visually-guided motor control and navigation behavior in a complex 3D maze environment without any external supervision using curiosity as an intrinsic reward signal. The learned representations are sparse, encode meaningful, action-oriented information about the environment, and can perform few-shot object recognition despite not knowing any labeled data beforehand. Overall, I make a case for increasing the realism of the computational tasks ANNs need to solve (largely self-supervised, sensorimotor learning) to improve some of their shortcomings and make them better models of the brain.

PUBLICATIONS

Clay, Viviane, Peter König, and Sabine König (2019). "Eye tracking in virtual reality." In: *Journal of Eye Movement Research* 12.1. ISSN: 19958692. DOI: 10.16910/jemr.12.1.3.

Clay, Viviane (2020). "Data from Neural Network Training in the Obstacle Tower Environment to Investigate Embodied, Weakly Supervised Learning." In: *Mendeley Data*. DOI: 10.17632/ZDH4D5WS2Z.2.

Clay, Viviane, Johannes Schrumpf, Yannick Tessenow, Helmut Leder, Ulrich Ansorge, and Peter König (2020). "A quantitative analysis of the taxonomy of artistic styles." In: *Journal of Eye Movement Research* 13.2. DOI: 10.16910/jemr.13.2.5.

Clay, Viviane, Peter König, Kai-Uwe Kühnberger, and Gordon Pipa (2021a). "Learning sparse and meaningful representations through embodiment." In: *Neural Networks* 134, pp. 23–41. DOI: 10.1016/j.neunet.2020.11.004.

Clay, Viviane, Peter König, Gordon Pipa, and Kai-Uwe Kühnberger (2021b). "Fast Concept Mapping: The Emergence of Human Abilities in Artificial Neural Networks when Learning Embodied and Self-Supervised." In: *arXiv preprint arXiv:2102.02153*.

König, Sabine U, Viviane Clay, Debora Nolte, Laura Duesberg, Nicolas Kuske, and Peter König (2019). "Learning of spatial properties of a large-scale virtual city with an interactive map." In: *Frontiers in human neuroscience* 13, p. 240. DOI: 10.3389/fnhum.2019.00240.

König, Sabine U., Ashima Keshava, Viviane Clay, Kirsten Rittershofer, Nicolas Kuske, and Peter König (2021). "Embodied Spatial Knowledge Acquisition in Immersive Virtual Reality: Comparison to Map Exploration." In: *Frontiers in Virtual Reality* 2, p. 4. ISSN: 2673-4192. DOI: 10.3389/frvir.2021.625548.

And, when you want something, all the universe conspires in helping you to achieve it.

King of Salem, The Alchemist — (Coelho, 1998)

ACKNOWLEDGEMENTS

First, I would like to thank my supervisors Gordon Pipa, Kai-Uwe Kühnberger and Peter König. You have guided me for years up to this point and supported me at every stage of this dissertation. Thank you for all the meetings, fruitful discussions, and knowledge you have passed on to me.

I would like to especially thank Peter König, who has already supervised me throughout my Bachelor's and Master's and taught me all the foundations of scientific work. There is no one else that I have learned so much from about clearly presenting and constantly questioning the results. I would not be where I am now without countless meetings with you and your excellent supervision over the past five years.

I would also like to thank Pascal Nieters and, again, Gordon Pipa for supervising my Master's thesis and preparing me through many insightful meetings for the Ph.D. years to come.

Thank you also to the Numenta research team for many inspiring scientific discussions. Even though it was late in the evening, I never got tired of those conversations and always looked forward to our meetings. A special thanks also to Lucas, whom I met with most frequently and who showed me many tips and tricks around writing code and collaborating on coding projects. I definitely learned a lot from you.

Thank you, Sabine, for your guidance during my Bachelor's and Master's and as my non-scientific supervisor for the past three years. It was always encouraging to talk to you about any difficulties I had. Thank you for listening and for all your advice.

Next, I'd like to thank my family for being there for me and also for supporting me in all my endeavors throughout my life. Thank you for encouraging my curiosity about everything, for giving me a safe environment to experiment and find my passions, and for always believing in me.

Most importantly, I would like to thank my husband, Joshua, who has been suffering much more for this dissertation than I have. Thank you for being here with me and for supporting me in the many ways that you did. You always knew just how to get my spirits up. Thank you for making me take a break and for reminding me to take it easy. And also for helping me get my mind off of my research sometimes. Going outside to work on a movie, scout locations, travel, or just hanging out together was always a great counterbalance. I could not have done this without your support.

CONTENTS

Part I

INTRODUCTION AND BACKGROUND

1

INTRODUCTION

"Instead of trying to produce a program to simulate the adult mind, why not rather try to produce one which simulates the childs? If this were then subjected to an appropriate course of education one would obtain the adult brain."

— Computing Machinery and Intelligence (Turing, 1950)

Even though AI outperforms us on many challenging tasks such as playing chess, there is currently no AI that can match the general capabilities of a toddler.

Over the history of artificial intelligence research, one strange phenomenon emerged: It seems easier to teach machines tasks that are deemed very intellectually challenging, such as multiplying large numbers or playing chess, than to teach them basic skills any toddler possesses and adults perform without consciously thinking, such as action and perception. This is also known as Moravec's paradox (Moravec, 1988). All the abilities that five-year-old children already have, like visually guided motor behavior, general problem-solving abilities, language understanding, or theory of mind, seem to be locked away from current artificial intelligence systems. To date, there is no general AI system that comes even remotely close to the broad intelligence and general abilities of a human (Hole and Ahmad, 2021).

To give AI more human-like capabilities, we could give them stronger inductive biases that resemble ours.

The no-free-lunch theorem states that averaged over all possible data-generating distributions, every classification algorithm has the same error rate when classifying previously unobserved points (Wolpert and Macready, 1997). Therefore, to generalize well on a specific set of tasks, there always has to be some inductive bias (Mitchell, 1980). If we want to solve similar tasks to the ones that humans solve well, such as vision or motor control, it is reasonable to start with an inductive bias similar to the one used in the brain.

How learning and information processing in the brain is defined and constrained is still unclear. However, the knowledge we do have can help inform how the ideal artificial learning system should be designed to match or supersede human abilities. Since the inductive biases of humans and other intelligent species are encoded in a relatively small genome, this information bottleneck suggests that a few rules might underlie the brains optimized learning structures and that they could be applied to machine learning to facilitate more "human-like," fast learning (Zador, 2019).

There are several levels on which current AI algorithms could incorporate more biological inductive biases and learning mechanisms.

Hole and Ahmad (2021) argue that further breakthroughs on the path to artificial general intelligence (AGI) will not originate simply from engineering tricks alone but require more brain-like mechanisms in AI algorithms. They make the case that it is crucial to look at neuroscientific evidence about the neocortex when designing intelligent machines. Sinz et al. (2019) argue along similar lines and propose three elements that could help give ANNs an inductive bias more suitable for general intelligence: Training on multiple tasks, optimizing ANN representations to be similar to encodings of the brain,

and using network architectures with more structural similarity to the brain.

Sinz et al. (2019) point out that the processing power of ANNs is not their limiting factor and that simply making them larger and using more extensive datasets might ultimately be a dead end. The brain seems to implement more useful inductive biases to obtain the general and robust abilities on many tasks and stimulus distributions we know from even small children. Ultimately the way to more generally intelligent systems may be to add better inductive biases in the form of architecture, learning rules, and learning environment (Sinz et al., 2019).

The inductive biases and principles of the brain may be simple but give rise to complex behavior.

Similarly, Richards et al. (2019) propose that using three concepts from AI as a framework could help understand the brain: architectures, learning rules, and objective functions. These three factors, combined with a rich and informative environment, can lead to learning very complex representations and behavior. This optimization-based framework may be better than using explicit models on a circuit level and coming up with human interpretable processes. They argue that one can view *"neural responses as an emergent consequence of the interplay between objective functions, learning rules and architecture"* (Richards et al., 2019, p.1764).

I will focus on the computational level of information processing, which is the computational problem that needs to be solved.

These three aspects align nicely with the three levels of understanding information processing proposed by David Marr: Hardware implementation, representation and algorithm, and computational theory (Marr, 2010; Richards et al., 2019). This dissertation will focus on the importance of looking at Marr's highest level on which information processing must be understood; the computational level. This level deals with the computational task and how a system can solve it. Marr describes the importance of this level the following way:

"Although algorithms and mechanisms are empirically more accessible, it is the top level, the level of computational theory, which is critically important from an information-processing point of view. The reason for this is that the nature of the computations that underlie perception depends more upon the computational problems that have to be solved than upon the particular hardware in which their solutions are implemented. To phrase the matter another way, an algorithm is likely to be understood more readily by understanding the nature of the problem being solved than by examining the mechanism (and the hardware) in which it is embodied.

In a similar vein, trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: It just cannot be done. In order to understand bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds' wings make sense." (Marr, 2010, Chapter 1, p.27)

I will also summarize similarities and differences between the brain and ANNs on the level of implementation and algorithm.

The level of representation and algorithm and the level of hardware implementation enjoy a lot of interest in the literature on biologically inspired machine learning. As these aspects are, of course, also relevant in the big picture, I will shortly summarize them in the following background sections. In each chapter, I will emphasize the overlap and the differences between biological and artificial learning systems.

However, much work on making ANNs realistic models of cognitive function focuses on these lower-level aspects and disregards the learning environment and task (Pulvermüller et al., 2021). Even the work that does focus on making data sets more ecological still implies a very unnatural way of learning, namely fully supervised learning from a labeled dataset of static images (Mehrer et al., 2021). I argue that changing the learning setup to be more natural can lead to learning vastly different representations and skills.

Letting AI systems learn more like children learn could also lead to more capabilities as we know them from children.

To make new breakthroughs in artificial intelligence and to achieve more child-like learning capabilities, it may be helpful to take inspiration from what is known about learning in children, particularly the field of developmental psychology (Smith and Slone, 2017). Therefore, I will present several aspects of learning in children that are very different or completely disregarded in artificial learning systems. Among these are embodied learning through interaction, self-supervised learning through prediction and curiosity, and gradually acquiring knowledge that can build on itself and generalize to various skills.

I focus on the effect of embodied learning with little to no external supervision and argue that this has a large influence on what is being learned.

In the included publications, I will first look at the effect of one of these aspects, embodiment, on learning in humans. Then, after demonstrating the effect that embodiment has on spatial knowledge acquisition, I take this and several other aspects and apply them to artificial neural networks. I show that ANNs that learn through curious, self-supervised interaction with the world learn different representations than conventionally trained ANNs. These representations are sparse and robustly encode action-relevant aspects of the environment. Additionally, the representations can be used to efficiently learn a new task such as object recognition with very few labeled examples. Overall, I demonstrate the effect that task and environment have on learning and argue for more realistic training of artificial learning systems to reach more human-like capabilities.

2

DEEP LEARNING - FOUNDATIONS

Ptolemy (king of Egypt): Is there no shorter road to geometry than studying your 13 books 'The Elements'?
Euclid (greek mathematician): There is no royal road to geometry.

— (Proclus, 1992)

Deep learning is a way of approximating the underlying function in a data set using ANNs.

In deep learning, artificial neural networks (ANN) are used as function approximators to model the input-output relationship of a given data set. The data set is usually a set of input-output pairs $\{(x_0, y_0), (x_1, y_1), ..., (x_n, y_n)\}$ which is used to train a model parameterized by θ. For example, a dataset may contain a collection of images as input x and corresponding labels of the image content as the target output y.

The mapping is inferred from the data provided instead of being manually implemented using a rule-based approach. The goal is to approximate a mapping from x to $\hat{y}$ that minimizes the difference between the approximated output $\hat{y}$ and the target output y. This means to adapt parameters θ such that $f(x, \theta) \approx f^*(x) = y$ using inputs x and target outputs y. Using an ANN to model this relationship makes it possible to generalize to new, previously unseen data. This ability to generalize gives the ANN the edge over using more straightforward solutions such as a look-up table of the data.

Figure 1: The learning setup.

If the target outputs corresponding to the inputs are available in a dataset, supervised learning can be used.

When the training data contains input-output mappings, this is called a *labeled dataset*. Using a labeled dataset for training a model is known as *supervised learning*. If such a dataset is not available, *semi-supervised learning* or *unsupervised learning* can be used. For instance, the data structure of an unlabeled dataset can be learned by reusing the input data X as the target outputs Y. This is, for example, done when training an autoencoder (Hinton and Salakhutdinov, 2006).

2.1 THE PERCEPTRON

The perceptron is a single processing unit, taking multiple inputs and returning a binary ouput.

The perceptron is a computational unit introduced by Rosenblatt (1958). The output of a perceptron is defined as

$$\hat{y} = \text{stepF}\left(\sum_{i=1}^{m} x_i W_i + b\right) \tag{1}$$

where stepF is a step function with value one if $x^T \cdot w + b > 0$ and zero otherwise. The step function is applied to the weighted sum of

the inputs. The value of these weights W determines how the input of the perceptron is transformed such that they fall either above or below the threshold of the step function. b is the bias term and can shift the transformed data and thereby the decision boundary (Bishop, 2006; Rosenblatt, 1958).

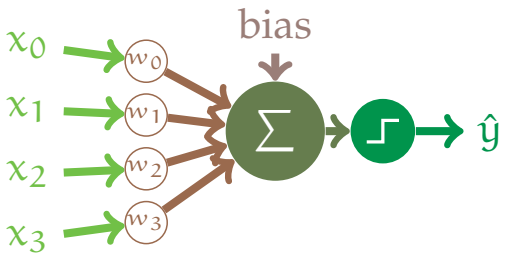


Figure 2: The perceptron implements a weighted sum of inputs, gated by a step function. A bias term can shift the entire weighted sum.

2.2 THE MULTILAYER PERCEPTRON (MLP)

For deep learning, multiple perceptrons are grouped into a layer. These layers can, in turn, be stacked, making the network "deep".

The multilayer perceptron (MLP) takes the idea of a perceptron one step further by arranging multiple perceptrons (also called neurons) in a layer. These layers can then be concatenated. The first layer is called the *input layer*, where each neuron represents one element out of the input vector x. The following layers are called *hidden layers* and are responsible for the information processing. The last layer is the *output layer*, it's activations should approximate y. Two consecutive layers are connected by a weight matrix W which specifies the weight between each neuron pair of the two layers.

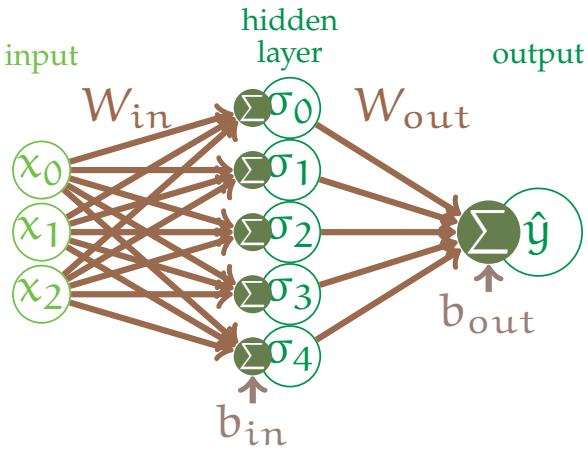


Figure 3: The Multi Layer Perceptron (MLP) combines multiple perceptrons into a layer. Multiple layers can be stacked before the final output is computed. The connectivity between two successive layers is defined by a weight matrix.

The output of an MLP is the weighted sum of the input, transformed by the activation function for each layer.

The output of an MLP with one hidden layer connected to the input layer with W_{in} and to the output layer with weights W_{out} is therefore

$$\hat{y} = \sigma(\underbrace{\sigma(\overbrace{x^T \cdot W_{in} + b_{in}}^{ws_{out-1}})}_{act_{out-1}} \cdot W_{out} + b_{out}). \tag{2}$$

Deep neural networks can approximate any continuous function.

Calculating $\hat{y}$ from the input is also called a *forward pass*. The step function of the perceptron is now replaced by a non-linear, differentiable activation function (for example a sigmoid function σ). Using at least one hidden layer with a bounded, non-polynomial activation function (such as the sigmoid function) makes the MLP a universal function approximator (Cybenko, 1989; Hornik, 1991; Leshno et al., 1993). Making the activation function differentiable also allows for calculating a gradient of the loss using partial derivatives and using this gradient for learning.

The ANN learns by using the gradient of the loss to adapt its network parameters.

The *loss* can be defined as the mean squared error between the network output $\hat{y}$ and the target y. Error-backpropagation can be used to calculate the gradient of this loss with respect to the network parameters W (Rumelhart, Hinton, and Williams, 1986). Starting with the last layer, the partial derivatives of the neuron activations are calculated with respect to their incoming weights and then used to calculate the partial derivatives of the previous layer. This is also called a *backward pass* and solves the credit-assignment problem, namely which weights should be updated how much to improve the performance of the whole network.

Backpropagation uses the chain rule of calculus to efficiently calculate the partial derivative of the loss with respect to the weights of the last layer W_{out} (Goodfellow, Bengio, and Courville, 2016).

$$\frac{\partial loss}{\partial W_{out}} = \frac{\partial loss}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial ws_{out}} \frac{\partial ws_{out}}{\partial W_{out}} = f'(\hat{y}) f'(ws_{out}) f'(W_{out}) \tag{3}$$

Where ws_{out} is the weighted sum $act_{out-1} \cdot w_{out} + b_{out}$ before applying the activation function and $\hat{y}$ is $\sigma(ws_{out})$. act_{out-1} are the activations of the neurons in the previous layer after applying the activation function.

The partial derivative of the weights in the previous layer W_{out-1} (in a network with one hidden layer, this would be W_{in}) then reuses the first two partial derivatives of equation 3.

$$\frac{\partial loss}{\partial W_{out}} = \frac{\partial loss}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial ws_{out}} \frac{\partial ws_{out}}{\partial act_{out-1}} \frac{\partial act_{out-1}}{\partial ws_{out-1}} \frac{\partial ws_{out-1}}{\partial W_{out-1}} \tag{4}$$

To update the network parameters, the gradient of the loss (multiplied by a learning rate) is subtracted from the parameters.

This method is repeated for any additional layers as well as the biases. The calculated gradient indicates the direction in parameter space into which the error increases the most. To improve the network, the parameters W need to be updated in the opposite direction performing *gradient descent*. This is done by subtracting the gradient (partial derivative) of layer l multiplied with a *learning rate* α from the corresponding weights W_l of the layer.

$$W_l = W_l - \alpha \frac{\partial \text{loss}}{\partial W_l} \tag{5}$$

Usually, updates are performed on a *batch* of data points which is a random subset of the entire dataset. This procedure is called *stochastic gradient descend (SGD)* and is done for all layers and for weights and biases alike. In practice, small tricks in the optimization are often used, such as adding momentum to accelerate learning and avoid getting stuck in a local optimum (Sutskever et al., 2013). Other advanced optimizers add a parameter-wise, adaptive learning rate to make bigger or smaller updates of the weights as needed (Duchi, Hazan, and Singer, 2011; Kingma and Ba, 2014; Zeiler, 2012).

2.3 CONVOLUTIONAL NEURAL NETWORKS (CNNS)

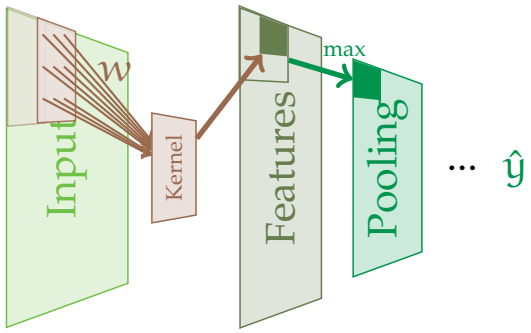


Figure 4: A convolutional layer in a CNN consists of several kernels that move over the input and extract features into a feature map. This feature map can then be reduced in size using a pooling operation.

CNNs are often used for image processing as they make efficient use of the image structure.

CNNs are a specific type of ANNs, containing at least one convolutional layer. The convolutional layer is an efficient way to dealing with data such as images or time series because it makes use of their gridlike structure (Goodfellow, Bengio, and Courville, 2016).

As shown in figure 4, a convolutional layer usually consists of several *kernels* that extract *feature maps* from their input. This is the convolutional operation, and it is followed by applying a non-linear activation function to the kernel outputs. The last operation of a convolutional layer is called *pooling* where multiple neighboring features are combined into one, for example, by taking their maximum or average value. This shrinks down the layer output size and leads to some translation invariance (Goodfellow, Bengio, and Courville, 2016).

CNNs share the same parameters by using small kernels that are moved across the input space.

The kernels usually have a smaller size than the input, which means they will be moved across the input until they have covered it all. The kernel processes the input at each position by multiplying it with its weights and calculating a sum, similarly to how it is done in the fully connected MLP. The weighted sum is then added as one element in the resulting feature map. The difference to the fully connected layer is that when the kernel is moved to the next neighboring

input patch, it uses the same weights as previously. This means that much fewer weights need to be learned.

For example, a kernel that processes patches of 3x3 pixels would have nine weights (and one bias) that are moved across the whole image. If the size of the image increases, the number of weights stays the same. However, if a fully connected layer would be used, the number of weights is the image $height x width x channels x next layer size$, which can quickly explode with larger images. The convolutional kernel solves this problem by sharing its weights and applying the same pattern detector to every part of the image (Bishop, 2006).

2.4 OUTLOOK

RNNs have neurons that connect to themselves and take into account previous activations. This allows a form of memory.

ANNs can also incorporate a time component by performing transformations of the input over multiple time steps. Those kinds of networks are called *recurrent neural networks (RNNs)* because the connectivity of their neurons involves recurrent connections. This means that a neuron can connect to itself and thereby influence its future state by its current state. Based on this principle, more complex forms of memory can be encoded, for instance by learning a gating that decides which information should stay in memory and which should be forgotten as is done in the *long short term memory (LSTM)* architecture (Hochreiter and Schmidhuber, 1997).

Countless other architectural decisions can be made for modeling different types of data. For instance, one can use attentional mechanisms to selectively "attend to" and process relevant parts of the input (Vaswani et al., 2017; Veličković et al., 2017). Feed-forward layers also do not need to be fully connected. Sparse connectivity (either by defining sparse weights or by applying *dropout*, zeroing out a random subset of weights in each update) between two layers can lead to more noise robustness and generalization (Ahmad and Scheinkman, 2019; Srivastava et al., 2014). Additionally, regularization and normalization can be applied to further improve generalization and stabilize learning (Goodfellow, Bengio, and Courville, 2016). Overall, deep learning is a very active field of research with many promising directions such as unsupervised representation learning, language processing, complex reasoning, combinations with reinforcement learning and recurrent processes, and many others (Lecun, Bengio, and Hinton, 2015).

3

ANNS AS A MODEL OF THE BRAIN?

"BEWARE THE PREACHERS
Beware the Knowers."

— The Genius of the Crowd (Bukowski, 1966)

ANNs draw from some biological principles but still have many differences to brains in their implementation, algorithm, and computational abilities.

The fields of machine learning and neuroscience have had decades of interactions and bidirectional influences on each other. Nevertheless, there are still many differences between biological motivated machine learning methods and the brain (Cox and Dean, 2014; Gerven, 2017; Hole and Ahmad, 2021; Kriegeskorte, 2015; Lake et al., 2017; Marblestone, Wayne, and Körding, 2016; Marcus, 2018; Pulvermüller et al., 2021; Smith and Slone, 2017). Richards et al. (2019) argue that artificial models of the brain should fulfill the following criteria: Be able to solve complex tasks that brains can solve, incorporate knowledge about the brain's anatomy and plasticity, and learn representations similar to those found in the brain. Unfortunately, ANNs are still falling short on all of these fronts. This chapter will review some of the current similarities and differences as well as the challenges of today's deep learning implementations and possible biologically inspired solutions.

3.1 SIMILARITIES AND DIFFERENCE ON THE ARCHITECTURAL AND ALGORITHMIC LEVELS

3.1.1 *The Neuron*

Artificial neurons are a highly simplified version of biological neurons.

Although properties of biological neurons initially inspired the perceptron (Rosenblatt, 1958) the comparison of the artificial neuron to the biological neuron is highly simplified. Biological neurons have a much more complex and versatile structure, such that a single biological neuron can perform sophisticated computations on its own already (London and Häusser, 2005). For example, it takes a two-layer MLP to roughly predict the firing rate of a pyramidal neuron (Poirazi, Brannon, and Mel, 2003) and approximately five to eight layers to incorporate the effects of multiple NMDA-based synapses and higher temporal resolution (Beniaguev, Segev, and London, 2021). On the other hand, it can be argued that an artificial neuron models a population of biological neurons because it can produce continuous output values while the output of a biological neuron is binary (spikes). However, the biological neuron has more leeway in adjusting the frequency of the binary spikes on a continuous spectrum and encoding information using precise spike timings and temporal codes (Gerstner et al., 1997). Therefore, comparing artificial neurons to biological neurons is a strong oversimplification, leaving out many of the properties of biological neurons.

Additionally to the differences at the level of the individual neuron model, the neuron population of an ANN is also highly homogeneous. The brain's neurons are very heterogeneous in their structure and the way they code information (even neurons of the same cell type) (Cembrowski and Spruston, 2019; Koch and Laurent, 1999). For example, meta-analyses identified 122 different types of neurons in the rodent hippocampus alone (Wheeler et al., 2015). For the brain, this heterogeneity may be an advantageous property, making information processing more robust (Perez-Nieves et al., 2020).

Lengler, Jug, and Steger (2013) suggest that robustness, sensitivity, and information processing efficiency may be further increased by the unreliability and noisiness of biological neurons. High variability is often seen as a problem in artificial information transmission systems, but it may actually be advantageous and an essential feature of the brain (Lengler, Jug, and Steger, 2013).

The complex structure of biological neurons, including dendrites, leads to a single neuron being a powerful computational unit by itself.

The brain's computational abilities do not only originate from the connectivity between neurons but also from computations that are performed within a neuron. The complex structure of the biological neuron, including dendrites, leads to each neuron being its own computational unit (London and Häusser, 2005). Neuron models that incorporate dendrites are, in general, more biologically plausible, and the effect of NMDA spikes at the dendrites can add important nonlinearities in the neuron's information processing (Major, Larkum, and Schiller, 2013). Going beyond the point neuron and using more complex neuron models by incorporating dendritic properties can lead to increased abilities to deal with computational complexity, such as adding more robustness on a temporal time scale (Leugering, Nieters, and Pipa, 2020, 2021), modulating activity by different contexts, and implementing a powerful sequence memory (Hawkins and Ahmad, 2016).

The brain may additionally employ apical dendrites of pyramidal neurons as a potential solution to the credit assignment problem by using top-down signals to distinguish credit-related information from other inputs (Richards and Lillicrap, 2019). Guerguiev, Lillicrap, and Richards (2017) show that dendritic mechanisms can be applied to machine learning and help with more biologically plausible credit assignment that avoids the weight transport problem (for gradient backpropagation, the backward weights and forward weights need to be symmetric. More on this in section 3.1.5). Sacramento et al. (2018) demonstrate that using local dendritic prediction errors can help solve the credit assignment problem and approximate backpropagation. The studies above suggest that working with a more complex, biologically-inspired neuron model can have various computational benefits.

Similar to ANNs, the brain's input-output relationships may also not be describable with compact formulas.

3.1.2 *Number of Parameters and Connectivity*

One criticism of ANNs is that they are heavily overparameterized and do not learn straightforward, human-interpretable rules. How-

ever, the brain also has a massive amount of parameters, and trying to find simple computational solutions that can be expressed in a short formula may be a misleading path in psychology and neuroscience. Over-parameterized systems have some counterintuitive mathematical properties that lead to good modeling and generalization of complex input and could be a more likely mechanism behind the impressive capabilities of the brain (Hasson, Nastase, and Goldstein, 2020).

The human brain still has many more parameters than today's ANNs.

Today's biggest ANN has 175 billion parameters (Brown et al., 2020) which, despite pushing current computational limits, is still relatively small compared to the human brain with an estimated 86 billion neurons (Azevedo et al., 2009), connected by many more synapses. Synapses could be roughly comparable to the parameters of an ANN. The number of synapses in the neocortex alone is estimated to be around 165 trillion, with an average of 6.900 synapses per neuron (Tang et al., 2001). This dwarfs the number of parameters of even large-scale ANNs.

The brain is not just a feed-forward network. It includes many feed-back and recurrent connections that have important functions for cognition.

The idea that the brain can be compared to a feed-forward neural network that processes raw sensory input in the first layers and successively more complex structures in the higher layers is rather simplistic. There is a plethora of evidence for top-down effects such as prediction, imagery, attention, emotion, and task on cognition and sensory processing (Betz et al., 2010; Dijkstra et al., 2017; Gilbert and Li, 2013; Moran and Desimone, 1985). The brain uses feed-forward, feed-back, and recurrent connections (Hupé et al., 1998; Siegel, Körding, and König, 2000) and Lamme and Roelfsema (2000) even suggest that the latter are necessary for conscious visual awareness. The activity of V1 neurons when viewing natural scenes is strongly influenced by complex network dynamics around them and not just by the sensory input, spike history, or local field potentials suggesting already rather complex and cross-cortically influenced representations in this early visual area (Haslinger et al., 2012). Also, on the level of cortical regions, most connections are bidirectional (Felleman and Van Essen, 1991), and synchronized patterns have been found between regions, encoding top-down influences on perception (Stein, Chiang, and König, 2000).

ANN architectures often do not reflect the complex bi-directional, recurrent and long-range connectivity of the brain. Incorporating recurrent and long-range connections in an ANN in the right way (implementing both gating and bypassing) can significantly improve object recognition accuracy while using fewer parameters (Nayebi et al., 2021). On the other hand, ANNs without recurrent connections can not sufficiently capture the dynamics of the human visual cortex, which means that using models with bi-directional information flow are necessary to model visual processing in the brain (Kietzmann et al., 2019). Overall, moving away from feed-forward neural network models by incorporating recurrency has many engineering and modeling advantages (van Bergen and Kriegeskorte, 2020).

The brain's connectivity is much more specialized and sparse than an ANN, possibly leading to better inductive biases for fast and efficient learning.

Also concerning connectivity sparsity, the connectivity in classical ANNs is very different from that of the brain (Felleman and Van Essen, 1991; Pulvermüller et al., 2021). While in most feed-forward

neural networks, all neurons in one layer connect to all neurons in the next layer (fully-connected), neurons in the brain only connect to a small subset of all possible connections they could form.

Overall, the brain's structural connectivity was optimized over millions of years of evolution, inducing very good inductive biases to enable fast and efficient learning in the type of environment the organism needs to act in (Zador, 2019). The connectivity matrix of ANNs, however, contains very little specialized structure, and practitioners usually initialize the weights randomly, which requires the network to learn entirely from scratch (Goodfellow, Bengio, and Courville, 2016). Thus, taking a look at general principles of the connectivity of the brain could help induce better inductive biases into ANNs.

Both ANNs and brains seem to use a few simple principles and repeating, unspecialized computational elements that, when combined, can perform a wide array of complex tasks.

One commonality in the learning approach is that the brain, similar to ANNs, may use only one or a few basic algorithms and building blocks across the whole cortex to solve many different tasks depending on the kind of input received. This is, for instance, indicated by experiments showing that ferrets can learn to see (to a lesser extent) with the auditory cortex if the auditory cortex gets 'rewired' shortly after birth to receive visual input (Melchner, Pallas, and Sur, 2000). The idea of cortical columns as a repeating structure all over the neocortex also fits into this principle (Mountcastle, 1997). However, the cortical column has a much more intricate wiring structure than building blocks of artificial neural networks and more powerful individual modeling abilities (Hawkins, Ahmad, and Cui, 2017).

3.1.3 *Activations*

Biological neural networks are much more sparse than ANNs, both in their connectivity and activations.

Additionally to the sparse connectivity, the brain also has an extremely high activation sparsity with only a few neurons at a time selectively responding to complex input patterns (Kloppenburg and Nawrot, 2014; Olshausen and Field, 2004; Quiroga et al., 2008). Such a sparse encoding has several advantages such as lowered energy consumption as well as increased robustness and representational capacity (Olshausen and Field, 2004) which also translate to ANNs (Ahmad and Scheinkman, 2019; Kurtz et al., 2020; Numenta, 2021). Furthermore, optimizing representations for sparsity can be beneficial to capture natural image statistics efficiently and leads to learning receptive fields similar to those found in mammals (Olshausen and Field, 1996). However, to date, the advantages of sparsity are usually only implicitly made use of by applying dropout (sparser connectivity) and the ReLU activation function (sparser activation).

The brain uses spikes and temporal codes to transmit information between neurons. This may have several advantages for continuous and efficient information processing.

As the gradient calculation in ANNs requires differentiable activation functions, they produce continuous output values. The brain, on the other hand, transmits information using discrete spikes, which can be modeled with Spiking neural networks (SNNs). SNNs have some beneficial properties, such as being able to determine causal influence, which helps to approximate gradients locally (Lansdell and Körding, 2019). Zylberberg, Murphy, and DeWeese (2011) demonstrated that SNNs learn V1 receptive fields using only local plasticity

rules. As summarized by Pfeiffer and Pfeil (2018); even though current hardware is not optimized for spiking neural networks, these networks can unlock great advantages on neuromorphic hardware such as more energy- and data-efficient learning based on instantaneous and local processing of incoming sparse activation bursts. As computations are event-driven, not much needs to be computed when there is little activity in the input. Additionally, spiking neural networks can produce initial output estimates based on incomplete information and improve their estimates over time (Pfeiffer and Pfeil, 2018). This makes them excellent solutions to the kind of real-time problem solving that brains need to perform on a day-to-day basis.

Due to their mechanics being more suited for continuous input and output streams, SNNs currently do not perform very well on conventional deep learning benchmarks (Pfeiffer and Pfeil, 2018). However, as I argue in this dissertation, these benchmarks are somewhat artificial and probably not the ultimate measure of intelligence. Despite the benchmarks not being optimal to capture the strengths of SNNs, recent advances in the field have made deeper SNNs possible and lifted them to a competitive level on popular visual benchmarks (Sengupta et al., 2019), language and music modeling (Woźniak et al., 2020). Furthermore, new methods such as backpropagation specialized for sparse, spiking neural networks make them more and more viable alternatives to conventional ANNs with the advantage of increased speed and memory efficiency (Nieves and Goodman, 2021).

Another difference in information processing is that the brain performs many computations in parallel while computations in ANNs are usually executed serially, and each layer has to wait for the computations in the previous layer to complete. However, this must not be a fundamental difference as any parallel computation can be rewritten to be performed serially (Marr, 2010).

3.1.4 *CNNs and the Visual Cortex*

CNNs have better inductive biases than ANNs for image processing. Their learned response properties are similar to those found in the visual cortex.

As discussed in Kriegeskorte (2015), convolutional neural networks now reach human-level classification performance on several tasks. Their architecture is brain-inspired, and the rough concepts behind them could be implemented with similar mechanisms in nature. Several properties of the visual system in the brain also arise in CNNs, and their learned visual representations compare to representations found in the brain (Kriegeskorte, 2015).

Hubel and Wiesel famously discovered orientation selective cell responses in columns of the primary visual cortex in cats and primates (macaques) (Hubel and Wiesel, 1959, 1968). Their model of simple and complex cells and their response properties and hierarchical structure inspired the design of the neocognitron (Fukushima, 1980) which in turn inspired convolutional neural networks (Goodfellow, Bengio, and Courville, 2016). The learned response properties in CNNs have some commonalities with early visual areas in the brain. For instance, the orientation-selective response of V1 neu-

rons to bar-shaped stimuli found by Hubel and Wiesel also appears in the first layer of convolutional neural networks. Successive layers then respond to more and more complex features (Zeiler and Fergus, 2014), similar to some observed response properties of cells in the visual pathway of the brain. Furthermore, performance-optimized object detection networks can model neural responses at higher levels of the cortical hierarchy, such as V4 and IT (Khaligh-Razavi and Kriegeskorte, 2014; Pospisil, Pasupathy, and Bair, 2018; Yamins et al., 2014; Yamins and DiCarlo, 2016). Also, when trying to predict spiking V1 activations in response to natural stimuli, CNNs are currently the best model (Cadena et al., 2019).

There are some approaches to measure how close current ANNs are to what we know so far about the brain. Schrimpf et al. (2020) developed a BrainScore which approximates an ANN's similarity to the ventral visual stream of primates according to several benchmarks, using current biological knowledge and experimental recordings. They show that ANNs that perform better on ImageNet, a popular image classification benchmark, tend to have a higher BrainScore. However, this correlation weakens with the more recent state-of-the-art networks, suggesting that some of those advances do not stem from increased biological plausibility (Schrimpf et al., 2018).

Even though there is a lot of research effort going into understanding how the visual system works in the brain, there are still many open questions (Olshausen and Field, 2005). Therefore, even if one wanted to make a biologically accurate artificial model of the brain, it would not be feasible yet due to a lack of knowledge about the biological processes involved in learning, knowledge representation, and perception in the brain.

3.1.5 *Backpropagation*

The error backpropagation used in deep learning can not be implemented by the brain. However, the brain can use local approximations of the gradient for learning.

The brain fulfills some of the requirements for backpropagation, namely being able to make changes on the synaptic level (Bliss and Lømo, 1973; Markram et al., 1997) and possessing feed-back connections that could transmit error-related information and modulate activations (Gilbert and Li, 2013). However, as it is usually implemented in deep learning, some of the assumptions of backpropagation do not hold for the brain, such as symmetric feed-back weights, magnitude-based gradients (not spike-based, gradients can be negative too), and separate forward and backward passes of information (Lillicrap et al., 2020). Therefore many believe that deep learning, specifically using the strict formulation of backpropagation, is not what is happening in the brain (Crick, 1989) but may be approximated with more local methods (Lillicrap et al., 2020). As using gradients for learning has shown to be very effective, there are several proposals for how the brain may be learning using gradient estimates that are biologically more plausible (Bartunov et al., 2018; Bengio et al., 2015; Cho et al., 2011; Guerguiev, Lillicrap, and Richards, 2017; Guerguiev, Körding, and Richards, 2019; Jabri and Flower, 1992; Lansdell and Körding,

2019; Lillicrap et al., 2016b; Roelfsema and Ooyen, 2005; Rowland, Maida, and Berkeley, 2006; Sacramento et al., 2018; Whittington and Bogacz, 2017). However, many suffer from problems like a higher bias or variance compared to exact gradient calculation using error backpropagation (Richards et al., 2019) and finding more biologically plausible alternatives is still an active area of research.

3.1.6 *Changes in Representation from Increased Biological Similarity*

When moving closer to biology on any of these aspects listed above, the learned representations become closer to those found in the brain.

Several studies show that perceptual and processing mechanisms comparable to those found in nature seem to develop when ANNs learn under similar conditions. For instance, when training recurrent CNNs using a biologically inspired predictive coding framework, they become susceptible to perceptual motion illusions (Lotter, Kreiman, and Cox, 2020; Watanabe et al., 2018). Other emerging phenomena include, for example, frequency selectivity and temporal tuning properties of the primary auditory cortex (Singer et al., 2018), error patterns and predictions matching the human auditory cortex (Kell et al., 2018) and grid cells (Banino et al., 2018). When using a model that simulates many aspects of the macaque visual cortex (a multi-area spiking network with similar connectivity), the model activity is similar to resting-state activity measured in the macaque brain (Schmidt et al., 2018).

Not just making the model more brain-inspired but also using more ecological data sets can lead to increased similarities between the object representations learned in an ANN compared to those found in the brain (Mehrer et al., 2021). One major difference between most CNNs and the human visual system is that the former often receive full-resolution images as input. The latter only perceives the most central part of the visual field (fovea) in full resolution and then performs multiple saccades across the visual stimulus. First approaches incorporating saccades and foveation into CNNs have been promising by reducing computational cost while mostly preserving accuracy (Akbas and Eckstein, 2017; Daucé, 2018; Jaramillo-Avila and Anderson, 2019). Also training CNNs on egocentric observations of children leads to much faster learning than with an adult dataset. Analysis of the dataset reveals that it contains a great diversity of rare viewpoints of objects (Bambach et al., 2018). These studies show that there are many promising possibilities and interventions on all of Marr's three levels to make ANNs a better model of the brain.

3.1.7 *Conclusion*

Many current advances in deep learning are motivated mainly by engineering considerations and focus on optimizing a specific problem instead of modeling the brain like the field of computational neuroscience does (Goodfellow, Bengio, and Courville, 2016). However, using knowledge about the brain as inspiration for machine learning algorithms has been useful in many cases in the past. Additionally,

Level of understanding	The Brain	Most ANNs (at level of simulation)
Hardware implementation	Complex neurons with dendrites	Point neuron
	Many different types of neurons, each neuron has a unique structure	Homogenous neuron structure
	Sparse connectivity	Fully-connected layers
	Weak hierarchical structure (Long-range connections, feedback connections)	Strong hierarchical structure
	Inductive bias optimized by evolution	Tabula rasa weight initialization, architecture optimized by engineering considerations
	Parallel, asynchronous computations	Serial, synchronous execution of layers
Representation & Algorithm	Binary output (spikes)	Continuous valued output
	Sparse temporal code (firing frequency)	Dense static output
	Local learning rules (Hebbian plasticity, SDTP, LTP)	Gradient backpropagation (global)
Computational (functional)	Active, action-oriented learning (data actively sampled over time)	Inferring statistics from big, static data sets (i.i.d.)
	Continuous, life-long learning, gradual knowledge acquisition	Separation between learning and inference, starting out with the final task
	Open-ended, multi-task learning, transfer knowledge and generalize quickly	Optimized for one task
	Weak supervision (much self-supervised, not always task-driven but also curiosity-driven, observational learning, social interactions, learning from few examples)	Fully supervised
	Noisy, low resolution sensors (e.g. foveated vision), communication (language) and attention	Full-resolution information channels
	Multi-sensory Processing	Usually one modality
	Probabilistic, causal modeling abilities/Bayesian reasoning, compositional knowledge representations (language, creativity, abstract reasoning)	Strong pattern recognizers, finding statistical correlations

Table 1: Differences between the brain and most ANNs sorted by Marrs levels of understanding information processing systems (list is not exhaustive).

neural networks can be valuable as models to understand cognition, especially when adding more biologically plausible elements such as more complex neuron models, sparse local connectivity patterns, and long-range connections, including inhibition, and implementing brain-like synaptic plasticity rules (Pulvermüller et al., 2021).

3.2 DIFFERENCES IN COMPUTATIONAL ABILITIES

"It's the last ditch effort, makes no sense to try
Put more gas on the tires, put more wood on the fire
Warm your hands up 'cause we'll probably be here a while
If you want the honest truth"

— Dance and Sing - Down in the Weeds, Where the World once was (Bright Eyes, 2020)

ANNs trained on images often over-rely on texture, can be tricked by small perturbations of the input, and in general make non-human mistakes.

3.2.1 *Adversarial Attacks and Texture Bias*

ANNs trained on object recognition are vulnerable to adversarial attacks (Kurakin, Goodfellow, and Bengio, 2016; Qiu et al., 2019; Szegedy

et al., 2013). This means that tiny perturbation to an image (invisible to the human eye) can make the network misclassify the image content with very high confidence. Even specific natural images can trick these networks into making mistakes that humans would never make (Hendrycks et al., 2019). There are some methods to make ANNs more robust to these adversarial attacks. For instance, making the neural network more similar to the brain by using a first processing block modeled after the macaque V1 makes the ANN more resistant against adversarial attacks (Dapello et al., 2020). However, even on an extremely simple dataset like MNIST (recognizing handwritten digits), adversarial defense tactics often fail when testing perturbations that they were not optimized for, and one can always find perturbations that would not trick the human perceptual system (Schott et al., 2018).

One reason for ANN's vulnerability to adversarial attacks may be that ANNs trained on visual tasks often exhibit a texture bias. This means they put more emphasis on low-level details of an image and disregard more global properties such as shape (Baker et al., 2018; Brendel and Bethge, 2019; Geirhos et al., 2019). This is contrary to human visual perception, which exhibits a stronger shape than texture bias (Landau, Smith, and Jones, 1988). For the narrow tasks given to most visual ANNs, such as recognizing objects in static images, it is easier to fit a function on local features than to take into account long-range dependencies such as global object shapes (Sinz et al., 2019). This strategy seems sufficient to solve those benchmarks but may not be sufficient in a complex multi-task, real-world environment such as the one in which we find ourselves.

Getting a network to exhibit less texture and more shape bias improves accuracy, especially on out-of-distribution and distorted images (Geirhos et al., 2019). First experiments by Hermann, Chen, and Kornblith (2019) show that this bias seems to be mainly influenced by the type of data the models are trained on and only to a small part by the model architecture or learning objective. Using architectures more similar to the human visual system, measured by the BrainScore, does not decrease the texture bias here. Using more realistic data augmentation such as noise and blur instead of excessive random cropping can reduce the texture bias and increase the shape bias (Hermann, Chen, and Kornblith, 2019). It seems an interesting possibility that increasing the realism of the learning setup, such as moving away from static images to enactive exploration or having a more open-ended learning objective, would further discourage this texture bias.

In general, current supervised and unsupervised ANNs make non-human mistakes such as overclassifying one class and over-relying on texture when confronted with noisy or distorted stimuli (Geirhos et al., 2020). Humans are better at dealing with distorted and noisy images than CNNs, and even though CNNs can deal with these challenging images if they are included in the training data, they fail to generalize to other out-of-distribution images (Geirhos et al., 2018).

3.2.2 *Continual and Multi-Task Learning*

ANNs have difficulties learning on dynamic datasets where the input distribution or the task changes.

The performance of ANNs hinges on the assumption of the data points being independent and identically distributed (i.i.d.), meaning that each data point is randomly drawn from an underlying probability distribution, independent of the previous data point. If the data distribution or the sampling shifts during training or afterward, this assumption is violated, and ANNs often fail. The data sampled by humans when interacting with the world is inherently non-i.i.d. and our remarkable ability to generalize experiences to novel situations, to transfer knowledge to a new task, or to extrapolate general rules to new domains shows that we must have a better mechanism to deal with these kinds of problems. Not only that, Smith and Slone (2017) propose that this unbalanced, ordered, and dynamic data distribution from which humans learn may actually be what allows us to robustly recognize objects under all kinds of circumstances.

One common problem in ANNs that learn on changing input distributions over time is a phenomenon called *catastrophic forgetting* where they completely discard already learned knowledge when starting to learn a new task (French, 1999). Ideally, learning about what a horse looks like should not make you forget previously acquired knowledge about the appearances of cats and dogs. However, if you give this new task of recognizing horses to a neural network without continuing training on the already learned skill of recognizing cats and dogs, the weights that encoded the previous knowledge will be changed entirely after only a few updates. This is significantly different from the kind of continual learning humans effortlessly perform over a lifetime (Parisi et al., 2019).

In general, most AI systems have a clear separation between learning and inference. Once a neural network is trained, it usually does not learn anymore, and to incorporate new information such as a novel object class or to adapt to a new input distribution, the whole network has to be retrained from scratch. With humans, this is obviously very different.

Current engineering solutions are often not sufficient. Learning more like children, using a curriculum or multi-task learning can be advantageous but still comes with problems.

There has been a lot of interest in continual learning in ANNs over the past years, and several methods have been introduced to learn new tasks without forgetting the previously learned knowledge (Delange et al., 2021; Kaushik et al., 2021; Kirkpatrick et al., 2017; Masse, Grant, and Freedman, 2018). However, most methods assume that the task ID (which task should be performed) is explicitly given to the network making it easier to recruit different network parts for different tasks specifically. Additionally, much of the current research focuses on continual learning setups where distinctly separate tasks (often image classification) are introduced sequentially, and many solutions are very tailored to this specific setup, leaving much room for further research (Delange et al., 2021).

The increasing motor abilities of an infant create an implicit curriculum of different object classes that become more available over time (Smith and Slone, 2017). Same as children, machines can benefit from gradually learning more and more complex tasks. For this, cur-

riculum learning can be used, which is known to improve learning speed (sample efficiency), performance, and generalization capabilities in deep learning (Bengio et al., 2009; Soviany et al., 2021) and reinforcement learning (Narvekar et al., 2020; Portelas et al., 2020).

Learning multiple tasks at once can help with learning each individual task by reusing learned inductive biases from one task for another, particularly when the tasks are related (Caruana, 1993). Furthermore, the inductive bias picked in a multi-task learning environment also leads to more stable generalization to new tasks from the same domain (Baxter, 2000). Even when the tasks are automatically generated without human labeling, the representations resulting from this unsupervised meta-learning setup generalize well to new test tasks and aid in supervised few-shot learning (Hsu, Levine, and Finn, 2018).

Training multiple tasks together can often be problematic, but training the right tasks together such that they can maximize the use of common knowledge can lead to significant performance and speed gains (Fifty et al., 2021). Research on methods for knowledge transfer between tasks and domains is active and ongoing, and the direction could provide some exciting solutions to learning with less labeled data as well as acquiring more general representations of input (Zhuang et al., 2020).

In most setups, ANNs optimize one global cost function while the brain may learn from more local, dynamic and complex cost functions. Marblestone, Wayne, and Körding (2016) show that using layer-specific or temporally changing cost functions aids ANNs with generalization and performance. Being able to learn multiple tasks and generalize knowledge from previous tasks to new ones through learning-to-learn mechanisms is an essential step towards more human-like artificial intelligence (Lake et al., 2017).

3.2.3 *Causality vs. Correlation*

ANNs are great at extracting correlations from large amounts of data but lack a causal understanding of it. Active interaction with the world and Bayesian models may help with this.

Looking at the previously mentioned problems, it seems that ANNs lack a true understanding of the world and the task at hand and can achieve good performances simply because of their powerful function approximation capabilities. A big, underlying problem is that these ANNs only learn correlations in the data, not causation (Vasudevan et al., 2021). The fact that conventional ANNs can only learn superficial statistical patterns from the i.i.d. data they are presented with and lack a causal model is a large contributing factor for their failure to generalize and transfer knowledge to new situations or instances outside of the training distribution (Schölkopf et al., 2021).

Young children do not just learn about the causal structure of the world from observations but also through interaction and experimenting with the outcomes of conditional interventions (Gopnik and Schulz, 2004; Schulz, Gopnik, and Glymour, 2007). They systematically build up hypotheses about the world and actively test them through targeted experiments that can produce evidence for or against their hy-

potheses, just like scientists do (Gopnik and Wellman, 2012; Gopnik, Meltzoff, and Kuhl, 1999).

Humans can infer the physical properties of objects by observing their interaction with other objects. They learn physical properties through experience and can use mental simulation to make predictions and inferences (Hamrick et al., 2016). To make predictions and inferences about physical interactions of objects, humans may use an intuitive, probabilistic physics model of the world to simulate outcomes using a form of noisy Newtonian physics (Battaglia, Hamrick, and Tenenbaum, 2013). Such an "intuitive physics engine" is an essential ingredient for efficient learning but currently lacking in ANNs (Lake et al., 2017).

Children also seem to use Bayesian reasoning to learn from probabilities about causality, and they do not just correlate actions with outcomes but understand causal and probabilistic relationships between events (Gopnik and Wellman, 2012). It may be crucial for machine learning methods to incorporate Bayesian methods for effectively learning the causal structure of the world (Vasudevan et al., 2021)

Being able to build causal models and extrapolate rules is a crucial skill in humans and animals, and building this into artificial learning systems could unlock many new capabilities.

Extrapolating rules and generalizing knowledge can be found all over the animal kingdom. For instance, in an experiment by Vaughan (1988) pigeons learn stimulus-behavior contingencies through reinforcement, where in some sessions the contingencies are reversed. After some training, the pigeons learn that they can infer from the first trial which way around the contingencies are in the current session. Also rats have been shown to learn rules and to be able to transfer them to new stimuli and determine whether a new sequence matches the learned rule (Murphy, Mondragón, and Murphy, 2008). For humans, already seven-month-old infants can extract general grammatical rules from only two minutes of speech and generalize them to new speech samples (Marcus et al., 1999).

Overall, current machine learning approaches often try to emulate intelligence by building powerful pattern recognizers. Lake et al. (2017) suggest that the key to learning and intelligence may be building good models of the world instead of pattern recognizers and that these models should be compositional and easily combined to form new models. This can make learning much more sample efficient, as knowledge can easily be generalized using infinitely many possible new combinations of old knowledge. Compositional models seem to be a crucial aspect for many elements of human cognition, such as understanding the world made up of compositional objects and object parts, language, creativity, and innovation.

3.2.4 *Supervision and Data Efficiency*

The brain requires much less labeled data to learn a specific task than current ANNs do.

Learning in ANNs can happen supervised or unsupervised. In the supervised setup, input-output mappings need to be provided to calculate the error and corresponding gradient for each of the network's predictions. Obtaining such a labeled dataset can be costly and some-

times infeasible. The rule of thumb goes that to match human performance using supervised learning, at least 10 million labeled examples are needed (Goodfellow, Bengio, and Courville, 2016).

The brain requires orders of magnitudes less labeled examples to learn tasks such as object recognition, and the majority of human observations are "unlabeled" (Zador, 2019). A child does not have a teacher at its side at all times, naming everything it sees. In fact, very few experiences of a child receive an explicit label, and most of the stream of observations contain no external supervision signal.

Even when children hear the names of objects, they still need to apply complex social knowledge to ground the language in the real world and infer which object the label refers to (Briganti and Cohen, 2011; Gopnik, Meltzoff, and Kuhl, 1999). For instance, a child may look at one object when an adult names a different object that the child is not attending to. If it would learn words simply by co-occurrence, the child should associate the new word with the object it was looking at when it heard the word. Experiments show, however, that infants can pick up on the adults' referential cues (like gaze or pointing) and correctly deduce which object the new name refers to (Baldwin, 1991; Baldwin, 1993).

Being able to learn from one or few examples is a crucial skill of humans, currently lacking in ANNs.

Impressively, children can sometimes learn new words from as little as one example and recall it several days later (Carey and Bartlett, 1978). This instant association between novel words and objects is also called *fast mapping* (Carey and Bartlett, 1978) and is not just exclusive to language learning in humans (Kaminski, Call, and Fischer, 2004). The quick association between word and meaning has been shown in many subsequent studies, however, for retaining the newly acquired words over a longer time, memory aids such as repetition seem to be required (Vlach and Sandhofer, 2012).

This fast mapping ability is still present in adulthood. For example, you can watch Star Wars once and will be able to recognize and name any lightsaber afterward. I can also tell you once that a short lightsaber is called a slaber, and you will be able to associate this new name and the concept instantly. Also regarding grammatical structures, the feedback provided by adults is too sparse and noisy to be reliably used for language learning (Marcus, 1993). Overall, children must be employing more efficient mechanisms besides supervised learning to acquire language.

There are several different approaches in machine learning for learning from small amounts of labeled data, also called *one-shot* or *few-shot learning* (O'Mahony et al., 2019; Wang et al., 2020). However, many of these approaches still require large amounts of labeled data to learn a suitable encoding that can generalize to new classes with only a few examples.

Multi-modal learning could simplify learning each modality on its own and integrating them in a coherent world model. Yu and Ballard (2004) show very early experiments on using language and visual perception in a multimodel learning setup where language helps cluster objects in visual space and objects are associated with words through

temporal co-occurrence. They show that using cross-modal learning can simplify both tasks and help ground language in perception.

It may be the defining factor of our intelligence that we are limited in the amount of time and computational resources that we have available, and solving those types of problems in machines could bring us closer to understanding our own intelligence (Griffiths, 2020).

3.2.5 *Learning through Interaction*

Children gradually build up knowledge through interacting with the world. ANNs often start out learning the final task, fully supervised with no embodiment or interaction.

Brains seem to have a different strategy than most ANNs to learn about the world. Following Piaget's observations, cognitive development is often described in four stages (Ahnert, 2014) (see figure 5) with some apparent differences to how machines learn. For one, there are different stages (even though they are not discrete but rather a continuous transition), meaning that knowledge is gradually acquired and builds on top of previous learning. For example, the child does not start out doing abstract reasoning or solving any specific task such as playing chess, driving a car, or predicting the stock market. Many years pass before the child graduates to attempting these challenging activities, whereas ANNs usually start with learning the final task. For instance, an artificial chess-playing agent did not previously learn how to walk, and an object recognition system did not have to figure out object permanence (the fact that objects continue to exist when they are out of sight).

The child, in contrast, starts out learning through interaction with the world. In the first two years, it learns about sensorimotor contingencies, object permanence, and other physical properties of the world. This happens largely without external supervision by moving in and interacting with the world. Especially the first stage, the sensorimotor stage (further divided into six substages), is focused on learning through interacting with the world, and Piaget views behavior as a crucial aspect of developing intelligence (Piaget, Aebli, and Seiler, 1992).

Being able to move around may simplify many computational challenges of perception.

Ballard (1991) introduces the notion of *animate vision* which is the notion that the human visual system is inherently active and our visual understanding of the world is closely tied to and simplified by our ability to actively sample it through movements of the gaze and body. *"The visual system is not completely general, nor does it have an arbitrary set of capabilities: it is a particular embodiment of the relationship of the organism to the world."* (Ballard and Brown, 1992, p.5). Animate vision systems embed perception in a behavioral framework and thereby constrain the required computations to make them much easier to solve. For instance *"the ability to control the camera's gaze, particularly the ability to fixate targets in the world while in motion, allows a robot to choose external coordinate frames that are attached to points in the world. Behaviors based on fixation point relative coordinates allow visual computations to be done with less precision."* (Ballard, 1989, p. 1636). What is required for animate vision is not an explicit internal representation

of the world but rather a set of adaptive behaviors that can precisely sample the information necessary to solve tasks (Ballard, 1989).

Sensorimotor learning seems to play an important role in animals and may also help in artificial agents. When including sensorimotor contingencies in the learning system of a robot, reliable and stable object recognition, even under unreliable action execution and perception, can be achieved (Maye and Engel, 2011).

As will be elaborated more in the research part, an agent learning through interaction with the world can acquire meaningful representations of its environment (Clay et al., 2021a) that can then be used to fast map names to different objects (Clay et al., 2021b). Also, Hill et al. (2021) demonstrate that an agent that learns language in an interactive environment, with a teacher providing an object name whenever the agent inquires about it, can ground the meaning of words and acquire fast mapping capabilities.

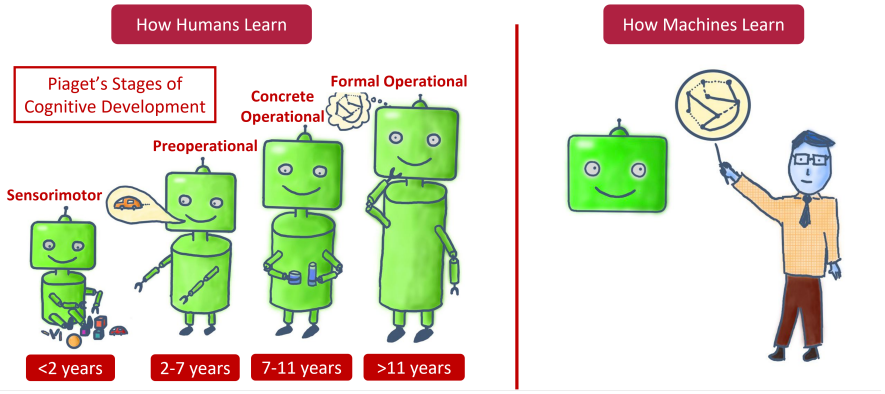


Figure 5: Piagets stages of cognitive development (left) (Ahnert, 2014) contrasted against the classical learning framework of ANNs (right). Humans learn enactive, gradually, open-ended and weakly-supervised.

3.2.6 *Conclusion*

All the problems of deep learning do not mean that it is inherently flawed or wrong; it is just a tool for a specific set of problems. This tool can be extremely powerful for applications where large amounts of data are available, and a static mapping needs to be extracted. However, it is not designed for the sort of problems biological brains need to solve on a daily basis. Even though they have inspirations from neuroscience, ANNs on their own are not the best solution to learning in a dynamic world and acquiring robust, generalizable knowledge beyond statistical correlations (Marcus, 2018).

To tackle some of the challenges on the computational level, the interdisciplinary field of *embodied AI* has started to focus on learning through interaction in a dynamic environment and how this may work in brains and machines (Duan et al., 2021; Iida et al., 2004). The following chapters will address aspects of enactive learning with weak external supervision and the role of these ideas in machine intelligence.

4

EVIDENCE FOR ACTIVE, SENSORIMOTOR LEARNING

"Well, what do you think you understand with? With your head? Bah!"

Zorba the Greek — (Kazantzakis, 1996)

Cognition seems to be strongly influenced by interaction with the environment.

There is accumulating evidence that cognition is grounded in sensorimotor coupling and should be studied not as a general model of the world but one that serves action (Engel et al., 2013; Foglia and Wilson, 2013). Instead of equating cognition with logical, step-wise computation as it has been conventionally the case in computational cognitive science, proponents of this idea assume a crucial role of the interaction with the environment in perception (not just a necessity to sample input). The idea is that cognition is not just about information-processing but about interaction and the body's relation to the environment (Rupert, 2019). Evidence for embodied cognition and related ideas has become eminent and of broad interest over the past decades, and even computationalists are now updating their theories to incorporate aspects of it (Scheutz, 2003).

4.1 THEORIES AND SUBFIELDS

Several different sub-ideas fall under the category of seeing cognition as being affected by interaction with the environment. Four of these are commonly summarized as the 4Es of cognition: enacted, embodied, embedded, and extended cognition. Closely linked and overlapping is also situated cognition. I will focus here on enacted [1] and embodied theories of cognition even though there is considerable overlap between all of these concepts.

Enactive cognition describes the idea that perception is an active process facilitated by sensorimotor coupling.

Enactive cognition was first introduced by Varela, Thompson, and Rosch (1991) stating that cognition is not an objective mirror of the world but shaped by our history of interactions and the actions that an agent can perform in a given environment. Varela, Thompson, and Rosch (1991) state that *"In a nutshell, the enactive approach consists of two points: (1) perception consists in perceptually guided action and (2) cognitive structures emerge from the recurrent sensorimotor patterns that enable action to be perceptually guided."* (p. 173). Proponents of enactivism view interaction with the environment as crucial for cognition. With the idea that *"experience is not something that happens in us but is something we do"* (O'Regan and Noë, 2001, p.99) conscious perception is seen as inseparably interwoven with action.

Embodied cognition describes the idea that cognition is influenced by the physical constraints of the body.

Embodied cognition is defined in the Stanford Encyclopedia of Philosophy as follows: *"Cognition is embodied when it is deeply dependent*

1 Specifically, I focus on sensorimotor enactivism. There is also autopoietic or radical enactivism with slightly different claims, but all of them emphasize the importance of sensorimotor interactions

upon features of the physical body of an agent, that is, when aspects of the agent's body beyond the brain play a significant causal or physically constitutive role in cognitive processing." (Wilson et al., 2021) This means that cognition is not just a product of the brain but the whole body, its sensors, and effectors. The most important function of the brain is to control the body to interact well with the environment. Focusing on abstract thought and logic for explaining cognition overlooks underlying principles of the workings of the brain (Clark, 2017). *"Minds make motions, and they must make them, fast - before the predator catches you, or before your prey gets away from you. Minds are not disembodied logical reasoning devices."* (Clark, 1997, p.1).

Theories of embodied cognition have this underlying idea of the body's role in cognition but come in various flavors and are not always clearly defined and disentangled (Wilson, 2002). There is often a substantial overlap with other action-oriented theories. For instance, situated cognition can be seen as a sub idea of embodied cognition (Wilson, 2002) and enactivism is a natural consequence of embodiment. Ward argues that *"if cognition is enactive, then it is also embodied, embedded, affective and potentially extended"* (Ward and Stapleton, 2012). However, the inclusion of extended cognition in the 4Es is debated, and Scarinzi (2020) argue that it is not compatible with the ideas of enactivism and should therefore be reduced to the 3Es of cognition.

Sensorimotor contingencies are the laws that describe how an action changes perception.

An important concept when investigating action-oriented theories of cognition are *sensorimotor contingencies (SMCs)*. They describe the way that motor actions affect sensory perceptions. To learn these SMCs (how actions change perceptions), the agent needs to interact with the environment. Often named under the umbrella of enactivism, *Sensorimotor Contingency Theory (SCT)* suggests that our conscious perception is not based on the raw sensory inputs but instead on the SMCs we have learned about the world. This would explain phenomena such as change blindness, where we do not perceive objectively visible changes in a scene. Our experience, in this case, seems to be more determined by the expected sensory changes following movement than on the raw sensory input. It was first suggested by O'Regan and Noë (2001) who made the radical proposal that *"seeing is a way of acting. It is a particular way of exploring the environment. Activity in internal representations does not generate the experience of seeing. The outside world serves as its own, external, representation. The experience of seeing occurs when the organism masters what we call the governing laws of sensorimotor contingency."* Since its proposal, the theory has been widely debated, partially due to different definitions of consciousness and experience, and alternative proposals and adaptations have been made (Bishop and Martin, 2014).

4.2 THE ROLE OF SELF-GENERATED MOVEMENTS

Before diving deeper into embodied and enactive theories and the supporting evidence, I would like to present a very early study by Held and Hein (1963). It is an excellent illustrative example of the

role of action, specifically self-generated movements, in learning and perception. It is closely related to O'Regan and Noë's enactive approach and the role of SMCs in perception.

Held and Hein demonstrated the importance of self-generated movements for perceptual learning in 1963.

In the study, Held and Hein took 8-12 week old kittens and divided them into two groups; an active group and a passive group. The kittens were all raised in the dark, so previous to the experiments they had no visual experiences. For the first phase of the experiment, they put the kittens pairwise into an apparatus where the kitten from the active group was able to walk freely around a pole in circles. The passive kitten was placed in a box on the opposite side of the circle. The movements of the active kitten were translated to the box of the passive kitten such that they both had nearly the same visual inputs. The main difference was that the active kitten was influencing its visual experiences directly by its movements, while the passive kitten was only observing the visual experiences without being able to influence them.

After several hours of daily visual exposure in the apparatus, the kittens are tested on their ability to show visually guided behavior. For instance, they are placed on an elevated platform with a deep drop on one side and a shallow one on the other. The active kittens all descended to the shallow side. The passive kittens did not seem to be able to distinguish the deep from the shallow side. Even when the passive kittens have eight more weeks of passive visual exposure, they fail at the task. However, when letting the passive kittens explore an illuminated room freely for 48 hours, they catch up to the performance of their active counterparts. Held and Hein's experiments highlight the importance of self-generated movements to perform visually guided tasks. Self-generated movements were crucial for these kittens to discriminate the deep from the shallow platform and could not be compensated by more passive visual exposure.

The importance of self-generated movements has been shown in many following experiments.

Of course, attention could have confounding effects in this study, and also the distinction between active and passive is difficult (passive kitten could still move their head) (Bermejo, Hüg, and Di Paolo, 2020). However, there were many follow-up studies on the role of self-generated movements in learning and perception with similar findings. For example, humans who wear a wedge-prism that distorts their visual field adapt much better under active conditions (walking) than passive conditions (pushed in a wheelchair) (Held and Rekosh, 1963; Mikaelian and Held, 1964). Active head movements have been shown to affect space and depth perception in humans (Wexler, 2003; Wexler and Van Boxtel, 2005). Even on a neuronal level, the responses to visual stimuli in early visual areas (V1 and V4) of mice and macaque monkeys are different between actively viewing the stimuli through self-generated movements compared to being passively presented with them (Mazer and Gallant, 2003; Niell and Stryker, 2010). Additionally, the kitten experiment setup was tested in an evolutionary robot where researchers found differences in the learned receptive fields between active and passive learning (Suzuki, Floreano, and Di Paolo, 2005). Overall these studies show the positive effect of self-generated movements on perceptual learning.

The type of action performed seems to also matter for perception. The ability of children to avoid falling off a (small) cliff depends on the way they move and how familiar they are with this movement. For instance, children who recently learned to walk stepped off the cliff while experienced walkers or crawlers did not. The height avoidance did not seem to transfer from crawling to walking (Kretch and Adolph, 2013). A similar effect was shown for the difference between a sitting and a crawling posture (Adolph, 2000).

Self-generated movements play a role in other sensory modalities besides vision and are important to learn SMCs.

The importance of self-generated movements has not only been shown for vision but also for other sensory modalities. Learning how to localize sounds can be achieved by using self-generated movements (Aytekin, Moss, and Simon, 2008). Touch and other somatosensory perceptions are already difficult to imagine without any interaction, as most of our somatosensory perceptions are elicited through our actions. Ostry et al. (2010) show that even short periods of interaction (10min) can affect perception for 24 hours. This effect can not be shown when passively moving someone's limbs (Ostry et al., 2010). Finally, even completely new senses can be best acquired by learning new sensorimotor contingencies through interaction. Using sensory augmentation devices allows for learning new SMCs and shows changes in the person's perception as well as in the representations in the brain (Kaspar et al., 2014; Kieliba et al., 2021; König et al., 2016; Nagel et al., 2005).

Embodied and enactive theories are additionally supported by studies into language processing. For instance, processing of spatial words requires sensorimotor processes (Ansorge et al., 2010). Imaging studies further validated that language understanding is accompanied by activations in motor and premotor regions and can be influenced by transcranial magnetic stimulation or lesions in these motor areas (Pulvermüller and Fadiga, 2010).

4.3 SENSORIMOTOR PROCESSING IN THE BRAIN

Before looking at some more related theories and evidence, I will give a quick overview of some of the neurological bases and evidence for sensorimotor processing.

Action-related signals can be found all over the brain.

Action processing seems to happen all over the cortex, with many "perceptual" neurons being gain modulated by factors such as eye velocity and head direction to keep a stable and translation invariant perception of objects in the world (Salinas and Sejnowski, 2001). There are body-centric visual receptive fields in the premotor cortex, which respond for example to stimuli close to the arm and shift in synchrony with arm movements. These types of cells may be important for sensorimotor control (Graziano, Yap, and Gross, 1994). Corollary discharge, which is an efference copy of the executed motor command that goes to the sensory regions, plays a crucial role in all animals for reflexes and a stable perception of the world. They enable the animal to distinguish self-generated movements from externally generated movements and to predict future sensory inputs.

This is also crucial for behavioral learning and planning (Crapse and Sommer, 2008). Interestingly, there is evidence for a role of the somatosensory cortex (S1) in imagining relative movements of limbs, challenging the common conception that only sensory processing is taking place in this early sensory area (Jafari et al., 2020). Furthermore, all areas of the cortex have connections to lower motor control centers and transmit motion-related information, not just the motor cortex (Sherman and Guillery, 2013).

The efference motor copies in the brain could be used to learn predictive models. A forward predictive model predicts the sensory observations following a motor command and may be located in the Cerebro-cerebellum (Kawato et al., 2003; Miall et al., 1993; Tanaka et al., 2020). There is also evidence of predictive mismatch signals in layer 2/3 of the mouse primary visual cortex. This predictive modeling capability is dependent on sensorimotor coupling during development (Attinger, Wang, and Keller, 2017). The inverse model is proposed to estimate the required motor command to get from the current state to the desired state and to be implemented in the Purkinje cells of the cerebellum (Kawato and Gomi, 1992; Kitazawa, Kimura, and Yin, 1998; Shidara et al., 1993). These predictive models, by design, need to combine action and sensory information to model the world.

There is neural evidence against the classical view of separated and hierarchical sensory and motor areas in the brain. So far, every sensory area in the cortex has been found to also receive motor signals and send out motor projections to the thalamus and subcortical motor regions. *"Information enters [the] cortex and leaves as motor instructions at each cortical level, and copies of these instructions are present at each level of input to [the] cortex. In this scheme, such differences between what is commonly known as sensory and motor cortex are quantitative (e.g., stronger motor outputs via layer five projections for "motor" areas) rather than qualitative. All cortical areas appear to function as sensorimotor regions."* (Sherman and Guillery, 2011, p.1073).

4.4 THE EFFECT OF ACTION ON PERCEPTION AND COGNITION

Perception is needed for action but action also influences perception.

There is a lot of evidence that not only action can be influenced by perception, but that object perception is also influenced by executed or planned action. For instance, processing visual stimuli that are congruent with the planned action is faster than stimuli that are incongruent (Craighero et al., 1999; Fagioli, Hommel, and Schubotz, 2007). Several theories were born from this general idea of perception being influenced by action.

Very early on, Gibson proposed that what we see is not merely the light reflecting off an object and hitting our retina but rather the interactions associated with this object. He coined the term *affordances* for these interactions that are naturally afforded by the objects in the world and argues for an experience-based perception that is determined by our relation to our environment (Gibson, 1979). In

early work, he showed that recognizing shapes by touch is not just a function of the shape arrangement on the skin, but that accuracy can be dramatically improved (49% to 95%) when actively exploring the shape by moving your fingers. He argues that actively touching objects produces a different sensory experience than passively being touched (Gibson, 1962). His work is influential in the field of psychology and perception and had a considerable impact on research into embodied and situated cognition (Lobo, Heras-Escribano, and Travieso, 2018).

When looking at an object that was learned to be used as a tool, not only visual regions activate but also motor regions associated with the tool use (Weisberg, Turennout, and Martin, 2006). These affordances are learned through active exploration of the environment (Adolph, Eppler, and Gibson, 1993) and seem to support visual object recognition (Roberts and Humphreys, 2011). For example, a right-handed person perceives a coffee cup rotated for a right-handed person to use (handle on the right side) earlier than a cup rotated for a left-handed person to use (Ariga, Yamada, and Yamani, 2016). This *affordance effect* may be caused by affordance modulated faster early visual processing (affordance is detected preattentive), leading to an earlier perception of affordance-matched stimuli.

Also, more higher-level cognitive functions such as spatial navigation and spatial knowledge acquisition are affected by the type of method employed during learning and the degree of embodiment. As will be elaborated more in the research section, learning by exploring an interactive map leads to representing knowledge in more global, allocentric reference frames (relative to north) (König et al., 2019). In contrast to this, actively exploring a town by walking through it leads to more egocentric reference frames (relative to the body). This means that after map exploration, it is easier to identify the orientation of a house relative to north, while exploration in VR or the natural world makes it easier to perform action-oriented tasks such as pointing from one house to another (König et al., 2019; König et al., 2021).

The theory of action-specific perception encompasses the idea that our perception reflects our ability to act in an environment, not objective reality.

The theory of *action specific perception* claims that *"Perception is not an objective representation of the environment but instead reflects the relationship between the environment and the perceiver's ability to act within it."* (Witt, 2011, p. 205). There is a plethora of evidence that our perception is not only a bottom-up reflection of reality but can be influenced top-down by our intended actions and our ability to execute these actions (Witt, 2017). For example, athletes with a higher scoring average perceive the ball as bigger (Witt and Proffitt, 2005) and better performance or increased ease due to a larger bat are correlated with perceiving the ball as slower (Witt and Sugovic, 2010). When asked to throw a ball to a target, the perceived distance to the target is influenced by how heavy the ball is (a heavier ball makes the target appear further away) (Witt, Proffitt, and Epstein, 2004). The amount of effort required also influences the perception of incline and distance when walking. For instance, wearing a heavy backpack, being fatigued, or in poor physical condition makes hills look steeper and targets appear further away (Bhalla and Proffitt, 1999; Proffitt et al.,

2003). Additionally, Witt, Proffitt, and Epstein (2005) demonstrate that objects seem further away when we intend to reach for them and they are too far away. If provided with a tool that enables you to reach for it, the object (at the same distance) seems closer. This perceptual effect is not present without an intention to reach (Witt, Proffitt, and Epstein, 2005). Therefore, our perception seems to be a mixture of physical reality and our intended actions, as well as the ability and effort associated with these actions.

The premotor theory equates attention with action planning.

The *premotor theory* postulates that attention can be equated to action planning. Meaning, that if you are attending to something in your right visual field, this attention is the result of the planned eye movement to the object to your right and is a weak version of the activations you would get when performing the movement (Craighero and Rizzolatti, 2005; Rizzolatti et al., 1987). While there is evidence for visual attention and saccade movements being coupled (Beauchamp et al., 2001; de Haan, Morgan, and Rorden, 2008), their causal relationship or a shared neural substrate have not been proven, and the claim that motor preparation is necessary and sufficient for attention can not be supported entirely by recent studies. According to the current state of knowledge, the theory has to be limited to bottom-up, stimulus-driven attention (exogenous) since top-down, goal-directed attention (endogenous) has been shown to exist without motor preparation. Alternative explanations for the coupling between motor planning and attention have also been proposed, such as a biased competition model in which attention can be biased by action preparation but is not determined by it (Smith and Schenk, 2012).

The common coding theory claims that action and perception share a neural code.

A related idea is the *common coding theory* which hypothesizes that action and perception share a common encoding in the brain. It underlines the strong, overlapping nature of action and perception. The theory suggests that acting shares a neural code with observing the action or mentally simulating it. It also suggests that action is not encoded as raw motor commands but rather as the perceptional changes they elicit (Prinz, 1990). This idea is supported by several experimental findings (Dayan et al., 2007; Etzel, Gazzola, and Keysers, 2008; Sommerville and Decety, 2006; Tye-Murray et al., 2013) and is closely intertwined with the field of embodied cognition as well as research on mirror neurons.

Whether perception is impossible without action is unclear, but action certainly has a strong influence on our perception.

Overall, the role of action in cognition is still a highly debated topic, and although there is increasing interest and scientific evidence, the strict necessity of action for perception is not clear. We know that all intelligent beings on this earth interact with the world in some way or another, starting from the day they are born. This naturally will have some effect, as can be seen from the overwhelming amount of evidence listed above. Whether action is merely a means to collect information or if it determines how reality is experienced and how cognition works is unclear. Additionally, the sometimes vague definitions and overlapping ideas in the field make it challenging to investigate the various claims systematically (Goldinger et al., 2016). However, as interaction with the environment plays such a central role for

all biological life, it should not be a factor quickly discarded when developing theories of cognition or building artificial intelligence.

In this dissertation, I will demonstrate that interaction strongly influences the learned knowledge in both humans and machines. As the definition of what constitutes a body is a bit murky, I focus on the aspect of interactive learning in the experiments with simulated AI agents.

5

DEEP REINFORCEMENT LEARNING - FOUNDATIONS

"This man has conquered the world! What have you done?",
The philosopher replied without an instant's hesitation,
"I have conquered the need to conquer the world."

— Alexander - The Virtues of War (Pressfield, 2005)

Before looking at the biological parallels and problems of deep reinforcement learning, I will briefly cover the underlying principles and algorithms. I will introduce all foundations necessary to understand proximal policy optimization, the algorithm used in two publications included in this dissertation.

The information presented in this section can mostly be found in the textbook *"Reinforcement Learning: An Introduction"* by Sutton and Barto (2018). It is a great comprehensive resource on reinforcement learning and the source of information here whenever it is not explicitly marked otherwise.

5.1 MARKOV DECISION PROCESSES

Reinforcement learning is used to solve markov decision processes with the goal of finding an optimal control strategy in a sequential task.

Reinforcement learning is applied to solve sequential decision-making problems. Contrary to the classic deep learning framework introduced in chapter 2, there is always a time component involved. The learner does not learn static input-output mappings anymore but instead interacts with the environment and learns through this interactive loop. Now, the input determines not only the output, but the output also influences the following input. The learning signal is not the difference between output and target anymore. Instead, one can use a temporal difference error or a policy gradient for learning from a reward signal.

An MDP consists of states, actions, transition probabilities, and rewards. It can be used to model the interaction of an agent with an environment.

The two main components of the interactive loop are called the *environment* and the *agent*. The environment at time step t is in state s_t. This state is used by the agent to determine the next action a_t which in turn influences the next state s_{t+1} of the environment. For this, the environment has transition probabilities $T(s,a,s')$ which are defined as the probability that the next state is s' given that the agent is in state s and performs action a which is $P(s_{t+1} = s' | s_t = s, a_t = a)$. To select the next action based of the current state, the agent uses a policy $\pi(a_t \mid s_t)$ which defines, often probabilistically, which action should be executed corresponding to which state. Finally, the environment also emits rewards which are defined by a reward function with a value for each transition $r = F_{reward}(s,a,s')$.

Together, states S, actions A, transition probabilities T and rewards R make up a Markov Decision Process (MDP) which is used to model sequential-decision making processes. Importantly, an MDP needs to

Figure 6: The optimal control setup. There is a bidirectional interaction between the agent and the environment over time.

fulfill the Markov property which means that s_{t+1} is dependent on state s_t, but conditionally independent from all previous states. This means that given the current state, the future is independent of the past. Certain RL algorithms make it possible to relax this property in practice.

The state s_t can either be the direct input to the agent or be indirectly observed by the agent. In the second case, the environment is called *partially observable* which means that the agent only has access to observations based on the state of the environment but not directly to the underlying state s_t. For instance, an agent may only see the part of the world directly in front of it. This does not mean that the rest of the world does not exist; it is just not observable at the moment. The current observation o_t is based on the overall state of the environment s_t.

5.2 SOLVING MARKOV DECISION PROCESSES

An RL agent tries to find the optimal policy to select actions that maximizes the overall reward the agent receives.

The goal in the family of *optimal control* problems is to find the optimal policy π^* that maximizes the overall reward that the agent receives from the environment over time. This can be defined as an expected discounted sum of rewards over infinite time.

$$\mathbb{E}[\sum_{t=0}^{\infty} \gamma r(s_t, a_t, s_{t+1})] \text{ with } a_t \text{ chosen from } \pi(s_t) \tag{6}$$

$\gamma \epsilon [0, 1)$ is the discount factor. The discount factor ensures that rewards do not sum up to infinity and provides a way of weighting immediate rewards against long-term rewards. A policy with a small γ will favor short-term, immediate rewards. When γ is close to one, the policy will forgo small immediate rewards to obtain larger rewards in the long term.

This sum of rewards is also called the discounted return G.

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{7}$$

The value $v(s)$ expresses how good it is to be in state s. It is defined as the expected, discounted, future reward when following policy π starting in state s.

The value of a state is the expected discounted sum over future rewards when following a given policy.

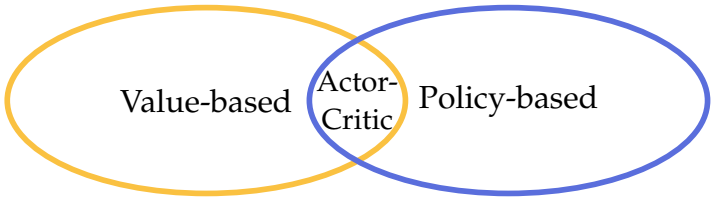
$$V_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s] \tag{8}$$

If the optimal value function V^* is known, the optimal policy π^* can be deduced by taking the action with the highest reward plus the discounted future return from all possible next states, weighted by the transition probabilities.

$$\pi^* = \arg\max_a \sum_{s',r} \underbrace{P(s',r|s,a)}_{\substack{\text{transition} \\ \text{probability}}} [r + \underbrace{\gamma V^*(s')}_{\substack{\text{discounted} \\ \text{future reward}}}] \tag{9}$$

To solve an MDP, one wants to find V^* and π^*. There is a large zoo of approaches for doing this. I will focus on *model-free* learning and will not detail *model-based* approaches such as *dynamic programming* and *exhaustive search* which require knowledge of all environment dynamics (transition probabilities and rewards). Model-based approaches are often used when the environment dynamics can be easily defined, such as the rules of a chess game. The defined model can then be used to calculate the outcome of different action sequences without actually having to play the game. In many real-world applications, these dynamics are not given, and the agent needs to learn by interacting with the world. Here one can differentiate between *offline learning* (e.g., Monte Carlo methods) where the agent performs updates based on a pre-collected batch of experiences and *online learning* (e.g., TD-learning) where the agent can perform updates simultaneously with every step in the environment. Another major division between reinforcement learning methods is whether one learns by updating an estimate of the value function (*value-based learning*) or whether one uses policy-gradients to update the policy directly (*policy-based learning*). I will go through all of these briefly to arrive at *actor-critic methods* which are a combination of value-based and policy-based methods and can be used online and offline. Finally, I will introduce one specific state-of-the-art actor-critic method, called *proximal policy optimization (PPO)* as it is the learning method applied in the research chapter.

Some common distinctions between RL approaches are whether they are model-free or model-based, online or offline, and on-policy or off-policy.



Monte Carlo approaches learn by interacting with the environment (as opposed to calculating optimal action trajectories using a given model).

5.3 MONTE CARLO LEARNING

The Monte Carlo approach uses experience sampling to solve the

optimal control problem when the underlying MDP is unavailable. One difficulty that comes with not knowing the transitions of the environment is that the policy cannot be inferred easily from the value function anymore. Previously one could do a one-step look-ahead from the current state and see which action gets the agent to the next state with the highest value estimate. However, without knowing the underlying MDP, the agent does not know what s_{t+1} will be when executing action a_t. The agent has to act to gather this information and can only do this for one out of all the possible actions in state s_t. To solve this problem, V can be replaced with Q which now estimates the value of state-action pairs instead of states.

$$Q(s,a) = \mathbb{E}_\pi[G_t|S_t = s, A_t = a] \tag{10}$$

$$\begin{aligned} q^*(s,a) &= \mathbb{E}[r + \gamma \max_a q^*(s',a')|S_t = s, A_t = a] \\ &= \sum_{s',r} \underbrace{P(s',r|s,a)}_{\substack{\text{transition} \\ \text{probability}}} [r + \gamma \underbrace{\max_a q^*(s',a')}_{\substack{\text{best q-value of} \\ \text{next state-action pair}}}] \end{aligned}$$

MC is used in episodic environments which have a terminal state. At the end of an episode, the agent performs an update, and the environment is reset.

As previously noted, the transition probabilities to calculate this expectation are not available and need to be learned by sampling experiences. Therefore, the entry $\hat{q}(s,a)$ of $\hat{Q}$ is estimated by using the average returns that followed being in (s, a) during past collected experiences. Experiences are collected in episodic environments, meaning that the environment contains terminal states which reset the agent. The agent interacts with the environment following a policy π until it reaches such a terminal state. The terminal state marks the end of an episode, at which point the agent calculates the returns and updates the q-value estimates (see pseudocode 1).

Pseudocode 1 : Monte Carlo (with exploring starts)

```
Initialize Q(s,a), π(s), returns(s,a) for all s ϵ S , a ϵ A
while π ≠ π* do
    choose s_0 and a_0 such that all pairs have p > 0 of being
     selected;collect experiences for an episode starting from
     (s_0, a_0), following π
    foreach (s, a) in the episode do
        append G_t that follows from the first/every occurance
         of (s, a) to returns(s, a)
        Q(s, a) = mean(returns(s, a))
    end
    foreach s in the episode do
        π(s) = arg max_a Q(s, a)
    end
end
```

Off-policy learning uses a different policy to collect experiences than the one that is being optimized. Epsilon-greedy is a stochastic on-policy method to ensure good exploration.

The policy that is used to collect experiences can either be the same policy that is being optimized (*on-policy*) or a separate *behavior policy*

b (*off-policy*). Off-policy learning usually has more variance and therefore takes longer to converge. One use-case of off-policy learning is if the agent cannot generate the data itself but can learn from already collected data, for example, by humans. For this to work, all actions taken by the learned policy π must at least sometimes be taken by b.

When learning on-policy, a proper exploration of the state-action space needs to be ensured. This can either be done by resetting each episode in a random position out of all possible state-action pairs (exploring starts) or by using a stochastic policy. A common choice for the latter is using an *epsilon-greedy* policy. This means that at each step, there is a small probability ϵ by which the agent samples a random action; otherwise, the agent picks the best action greedily under the current policy. The choice of ϵ weights how much the agent should explore the environment to collect novel experiences against how much it should exploit its already acquired knowledge.

5.4 TEMPORAL-DIFFERENCE LEARNING

TD learning performs updates online at every step by using a temporal-difference error between the current value estimate and the next.

Temporal-Difference learning (TD-learning) is an online-RL method which means that learning can happen instantaneously while collecting experiences. Therefore, no episodes are required since only the one-step return is being used instead of the return of a whole episode, like in Monte Carlo learning. The update is simple, requires minimal computation, and can be expressed in a single equation. At the core of this method is an incremental update of the value estimate using the TD error. This error is the difference between the current value estimate of state s and a better value estimate which is obtained after observing the next reward and the next state.

$$V(s) = V(s) + \alpha[\overbrace{r' + \gamma V(s')}^{\text{better}} - \overbrace{V(s)}^{\text{current}}] \underbrace{\qquad\qquad}_{\text{TD-error } \delta} \tag{11}$$

The magnitude of the value update is scaled by α, also called the step size. In a nutshell, the agent interacts with the environment and after every step the above update is performed on the value function. This can happen on-policy (e.g. *SARSA*) or off-policy (e.g. *Q-learning*) and like Monte Carlo also this method usually uses Q-values instead of V. In the one-step TD-learning, also called TD(0), all that is required for one learning step is a quintuple of experiences: (s, a, r', s', a').

In the space between TD and MC learning is n-step TD which uses n steps to calculate an update.

The one-step TD error only takes into account the most recent experience. This means that it may take a while for reward information to spread to further away states. For example, imagine an agent in a grid world where one field gives a reward of one, and the other states give no reward. If a TD(0) agent reaches this rewarding state, it will only update the state that it was in immediately before reaching the reward. However, it may also be helpful to remember the rest of the path used to get to this state. In TD(0), this will slowly be achieved by the value estimate tickling through to the neighboring fields on each

Pseudocode 2 : SARSA

```
Initialize Q(s,a) for all s ϵ S , a ϵ A
while Q ≠ Q* do
    initialize s choose a from s using policy derived from Q
    while s not terminal do
        take a, observe r' and s'
        choose a' from s' using policy derived from Q
        Q(s, a) = Q(s, a) + α[r'+γ Q(s', a') - Q(s, a)]
        s = s', a = a'
    end
end
```

successive visit. Nevertheless, there is a more efficient way to do this called n-step TD.

In n-step TD, the agent uses the next n states to update a value estimate. For this update, it has to calculate the n-step return, which considers the n-1 future discounted rewards plus the value estimate of the state n steps into the future.

$$
\begin{aligned}
G_{t:t+1} &= R_{t+1} + \gamma V_t(S_{t+1}) && \text{1-step return} \\
G_{t:t+2} &= R_{t+1} + \gamma R_{t+2} + \gamma^2 V_{t+1}(S_{t+2}) && \text{2-step return} \\
G_{t:t+n} &= R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{n-1} R_{t+n} \\
&\quad + \gamma^n V_{t+n-1}(S_{t+n}) && \text{n-step return}
\end{aligned}
$$

The n-step value update is then

$$V_{t+n}(S_t) = V_{t+n-1}(S_t) + \alpha[G_{t:t+n} - V_{t+n-1}(S_t)] \tag{12}$$

This means that an update can only be made after n steps since the computation of the n-step return requires r_{t+n} and $V_{t+n-1}(S_{t+n})$. N-step TD learning is an interpolation between TD(0) and Monte Carlo learning where the choice of n can be made dependent on the task demands. A useful extension of this idea is to weigh a sum of different n returns by λ^{n-1} such that states further away from a reward are updated less. This is, for instance, used in the λ-*return algorithm* and *TD(λ)* and can lead to significantly faster learning, especially in environments with delayed rewards.

Using function approximators (like deep neural networks) to represent value functions allows for larger state and action spaces and generalizing knowledge to new experiences.

5.5 VALUE-FUNCTION APPROXIMATION (DEEP RL)

So far, the methods introduced here have been using a look-up table for the V or Q estimates with one entry for each possible s or (s, a) respectively. This approach can quickly become impractical when dealing with large state spaces. For example, a simple Atari game like breakout produces visual states with 84x84 pixels for four consecutive frames. Assuming a greyscale version of the game with 256 grey values, this would mean $256^{84*84*4}$ possible states, which is far

more than the number of atoms in the universe and impossible to represent in a look-up table. Even if a giant look-up table could be constructed, it would require massive amounts of experience to be filled because almost every state encountered would be unique from all the other states. Therefore we need a scalable method that can generalize to unseen states.

A great solution for this are function approximation methods. The idea is to use some type of function approximator such as linear functions, polynomials, Fourier basis functions, radial basis functions, multi-layer perceptrons or CNNs, parameterized by features w to estimate $\hat{V}$ or $\hat{Q}$. Parameters w can then be adjusted, for example by stochastic gradient descent, such that $\hat{v}(s,w) \approx v_\pi(s)$ (same for $\hat{Q}$). Using deep neural networks for this value-function approximation is called *deep reinforcement learning*.

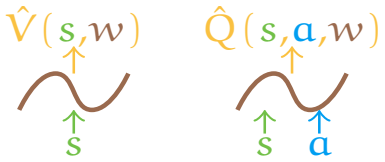


Figure 7: Function approximation of V (left) or Q (right) with parameters w.

Using function approximators makes it possible to deal with sizeable state-action spaces and with partially observable environments. However, the combination with RL introduces new challenges to deep learning that make stable learning more difficult. For one, non-stationary targets need to be approximated. There is not a stable target value for each state, and the training data keeps changing. While supervised function approximation usually assumes a fixed training set, function approximation in online-RL has to deal with a constantly changing policy generating new data distributions. Additionally, learning in RL often has to deal with delayed targets, and the training data is usually non-i.i.d. (where the agent is right now is highly correlated with where it will be next). Despite all these added difficulties, the combination of deep learning and reinforcement learning can be powerful and solve many novel problems.

5.6 POLICY GRADIENT METHODS

Policy gradient methods optimize the policy directly by performing gradient ascend on its parameters.

Monte Carlo and TD-learning are two value-based methods where the agent learns the values of states and then selects actions based on these values. The policy is, therefore, a byproduct of the learned action-value estimates. Another approach is to learn the policy directly by parameterizing it and performing gradient ascend on the policy parameters. This removes the need for a value function estimate and can be helpful in environments with high-dimensional or continuous action and state spaces.

Directly estimating the policy can often be more straightforward to learn than a value function. For instance, in the Atari game "Pong," it is easier to learn how to move the pedal relative to the ball position

than to predict the future reward and infer actions from this. Additionally, policy gradients have stronger convergence guarantees because the action probabilities change smoothly. When in value-based methods, the value of one action becomes the maximum value in that state, it leads to a big change in the policy resulting from a possibly tiny change in the value function. Lastly, policy-based methods can deal better with uncertainty and partial observability and make it possible to find stochastic policies. In some settings, such as playing rock-paper-scissors, the opponent can easily exploit a deterministic policy. Value-based methods have no way of finding an optimal stochastic policy here, while policy-based methods can learn a probabilistic solution.

On the downside, policy-based methods are often harder to optimize in practice. The gradients are noisy and have high variance, which means that stable learning requires large batch sizes and a well-tuned learning rate, and learning can be slow. In addition, since learning requires samples collected under the current policy (on-policy), older samples cannot be reused, and learning is sample-inefficient. This is problematic when data collection is expensive but can be partially alleviated by using importance sampling.

To optimize the policy directly, we first need to define a performance measure. In the episodic case, this is defined as

$$J(\theta) = V_{\pi_\theta}(s_0) \tag{13}$$

To optimize this, we need to find $\Delta J(\theta)$ which is the gradient of the performance. This gradient has been proven to be proportional to the sum of q-values of all actions multiplied by the policy gradient and weighted by the on-policy distribution of states $\mu(s)$ under the policy π. This is called the policy-gradient theorem and can be rewritten as an expectation dependent on π.

$$\begin{aligned}\Delta J(\theta) &\propto \sum_s \mu(s) \sum_a q_\pi(s,a) \Delta_\theta \pi(a|s,\theta) \\ &\propto \mathbb{E}_\pi[\sum_a q_\pi(s_t,a) \Delta_\theta \pi(a|s_t,\theta)]\end{aligned}$$

The gradient update of the policy parameters θ would then be

$$\theta_{t+1} = \theta_t + \alpha \sum_a \hat{q}_\pi(s_t,a) \Delta_\theta \pi(a|s_t,\theta_t) \tag{14}$$

REINFORCE is a practical implementation of PG that uses an episode of experiences to perform a policy gradient step.

This update rule involves all possible actions in a state in order to perform an update. REINFORCE is a practical implementation of policy-gradient optimization which performs a gradient update using only the action a_t that was actually taken in-state s_t (Williams, 1992). To do this, everything needs to be weighted by the probability of taking action a_t under policy π. Additionally, the sum of discounted returns G_t of the current episode of experiences is used instead of the q-value. The gradient ascend update of REINFORCE, therefore, looks like this:

$$\theta_{t+1} = \theta_t + \alpha G_t \frac{\Delta_\theta \pi(a_t|s_t,\theta_t)}{\pi(a_t|s_t,\theta_t)} \tag{15}$$

The fraction is also sometimes referred to as the "eligibility vector" and can be written as $\Delta_\theta \ln \pi(a_t|s_t,\theta_t)$. It can be interpreted as the direction that most increases the probability of a under π on future visits of s (numerator) divided by the current probability of taking action a under π. This means that the parameters θ are increased in the gradient direction weighted by the return and negatively proportional to the action probability. Including the action probability in this weighting of the update makes sure that frequently sampled actions do not have an automatic advantage because they are updated more often and not because they are actually better. In REINFORCE the parameter updates in equation 15 are made after the end of an episode because all rewards are used to calculate G_t. Then the next episode is collected using π_θ with the updated parameters θ.

5.7 ACTOR-CRITIC METHODS

Actor-Critics learn both a policy (actor) and a value estimate (critic). To do this they use a TD error.

Actor-Critic methods are a combination of value-based and policy-based learning. The agent learns both a value function and a policy. The value function is used for the gradient-based policy optimization in place of the episode return but not for action selection. It neatly combines all the concepts introduced earlier and can be used online with step-wise updates.

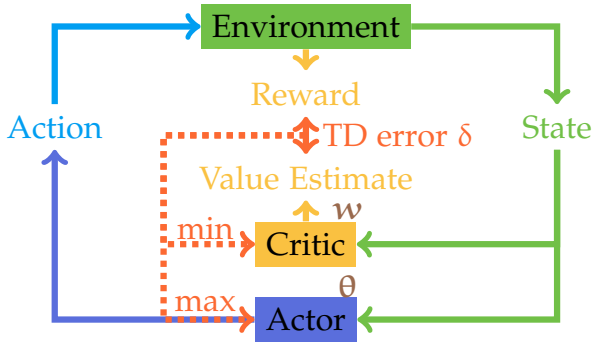


Figure 8: The Actor-Critic setup. Both actor and critic receive the state as input and use the TD error as learning signal. The actor produces actions that try to maximize the TD error. The critic produces values estimates and tries to minimize the TD error.

The actor-critic agent is made of two components, the actor and the critic, and optimizes two different objectives. The actor has policy parameters θ which are used to produce actions and optimized using policy gradients. The critic has value parameters w which are used to produce value estimates and optimized using TD(0) or n-step TD. As learning signal, the TD-error δ (see equation 11) is being used. The actor now tries to maximize the TD error instead of the discounted returns, which means it tries to adapt the policy at every step such that it leads to higher rewards than expected. The critic tries to keep

the TD error close to zero, meaning that it tries to make its value estimates as accurate as possible.

Pseudocode 3 : 1-step actor-critic

Define a differentiable policy $\pi(a \mid s, \theta)$ and value function $\hat{v}(s, w)$ and initialize θ and w
Set step sizes $\alpha^\theta > 0$, $\alpha^w > 0$
while $\pi \neq \pi^*$ **do**
  initialize s_0
  while s_t *not terminal* **do**
    take a_t according to $\pi(s_t)$, observe r_{t+1} and s_{t+1}
    $\delta = r_{t+1} + \gamma \hat{v}(s_{t+1}, w) - \hat{v}(s_t, w)$
    $w = w + \alpha^w \Delta_w \hat{v}(s_t, w)$
    $\theta = \theta + \alpha^\theta \delta \Delta_\theta \ln \pi(a_t | s_t, \theta)$
    $s_t = s_{t+1}$
  end
end

In practice, w and θ are often partially shared. For instance, when working with images as observations, the same convolutional layers can extract features for the value estimate and the policy. Both parameter updates can include an optional discount factor. The 1-step TD error can also be replaced by an n-step error as shown in equation 12.

To reduce variance, the agent can estimate the advantage instead of the value, which is the difference between the q and the v estimate of a state.

A small tweak to reduce variance during learning is to estimate the advantage instead of the value function. The advantage function is defined as $A(s,a) = Q_w(s,a) - V_v(s)$ and expresses how much better it is to take action a in state s compared to the general value of state s. Since now two value estimates are required (the Q-value and the V-value), there are two sets of parameters (w and v) that need to be optimized in the critic update. In practice, the advantages can be estimated using only the value function and therefor only one set of parameters (Schulman et al., 2015b). This is called the *advantage actor-critic (A2C)*.

To make sure that update steps are not too large in policy space, a trust-region constraint can be added.

One major problem with policy gradient methods is that $\Delta J(\theta)$ does not contain information about the second-order curvature of the reward function. This means that the performance can easily collapse if the step size is too large and finding the correct step size is very difficult (Dong, Ding, and Zhang, 2020). Unfortunately, actor-critics are not immune to this problem. To solve this, Schulman et al. (2015a) proposed the idea of a *trust region* which makes sure that the update on θ in parameter space does not lead to a big step in policy space. To measure this, they calculate the KL-divergence between the old policy and the new policy. *Trust region policy optimization (TRPO)* then computes the Hessian of the average KL-divergence in a sample to enforce a KL-divergence constraint. The step size is picked as the smallest number possible under the KL-divergence constraint that still improves the policy gradient loss (Achiam, 2018).

PPO is a simplified version of the strict trust-region constraint that still prevents large updates in policy space.

Using a Hessian matrix is complicated and computationally expensive. Another method called *proximal policy optimization (PPO)* uses a

little trick to simplify TRPO by removing the second-order approximation of the KL-constraint and still enforcing similarity between π_{old} and π_{new}. Instead of computing the Hessian of the KL-divergence, the KL-divergence is directly added into the objective function. This is either done by adding a regularization coefficient (*PPO-penalty*) or by clipping the ratio between the old and the new policy into a range between 1-ϵ and 1+ϵ (*PPO-clip*) (Schulman et al., 2017). PPO-clip is a stable and straightforward approach to constrain the policy updates. It seems to work equally well as TRPO and is commonly used in practice (Achiam, 2018).

Pseudocode 4 : PPO-clip

Define a differentiable policy $\pi(a \mid s, \theta)$ and value function $\hat{v}(s, w)$ and initialize θ and w
Set step sizes $\alpha^{\theta} > 0$, $\alpha^{w} > 0$ and clip factor ϵ
while $\pi \neq \pi^*$ **do**
  Collect N trajectories of experiences following π_θ
  Compute returns G_t
  Calculate advantage estimates $\hat{A}_t$ using G_t and $\hat{V}_t$
  for *M epochs* **do**
    $$ratio_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$$
    $$L_t(\theta) = \min \begin{cases} ratio_t(\theta) \cdot \hat{A}_t \\ clip(ratio_t(\theta), 1-\epsilon, 1+\epsilon) \cdot \hat{A}_t \end{cases}$$
    Optimize policy parameters θ w.r.t. $L_t(\theta)$
    Optimize value parameters w w.r.t. the value loss
  end
  $\theta_{old} = \theta$
end

5.8 OUTLOOK

The field of reinforcement learning is broad, and there are many exciting sub-fields and approaches that I will not detail here. However, for completeness, I will list a few more big ideas around this field.

There are many popular extensions of the concepts introduced so far, such as A3C, DDPG, TD3, and SAC.

First of all, there are many good extensions and variations of the algorithms introduced above. To name a few, there is the *asynchronous advantage actor-critic (A3C)*, which performs asynchronous updates between multiple experience collectors (Mnih et al., 2016); *(deep) deterministic policy gradient (DDPG)* where a deterministic policy is calculated instead of a stochastic one and off-policy learning improves sample efficiency (Lillicrap et al., 2016a; Silver et al., 2014); *twin delayed DDPG (TD3)* adds clipped double Q-learning (a variation on Q-learning), delayed updates and target policy smoothing to deal with the overestimation bias in the value function (Fujimoto, Hoof, and Meger, 2018); and the *soft actor-critic (SAC)* enforces a stochastic policy in an actor-critic framework to combat convergence problems in off-policy learning (Haarnoja et al., 2018). Combining several of

those approaches can lead to an even better performance (Hessel et al., 2017). Interestingly, Chen et al. (2021) recently suggested that one can also interpret RL problems as a sequence modeling task such that a transformer can be applied instead of learning a policy and value function.

Model-based approaches also show promising potential, especially on rule-based games such as chess and go.

As briefly mentioned, there are also model based RL approaches such as *dynamic programming* (Bellman, 1966) and *Monte Carlo tree search (MCTS)* (Browne et al., 2012; Chaslot et al., 2008) such as used in *AlphaZero* to play chess, shogi and go (Silver et al., 2017). There are also approaches that combine model-free and model-based learning. For example by learning an explicit model of the world in a model-free manner and then applying model-based optimization to it such as *Imagination-Augmented Agents (I2A)* (Weber et al., 2017). Or by using *model-based priors (MBMF)* (Bansal et al., 2017; Wilson, Fern, and Tadepalli, 2014) or *model-based value expansion (MBVE)* (Feinberg et al., 2018) for model-free reinforcement learning.

To learn from human demonstrations, imitation learning can be used such as behavioral cloning or inverse RL.

In offline settings or settings where interaction is very costly *imitation learning* such as *behavioral cloning* can be used instead of reinforcement learning to solve optimal control problems using existing target trajectories (Bain and Sammut, 1995; Hussein et al., 2017; Ross, Gordon, and Bagnell, 2011). Another approach is *inverse reinforcement learning (IRL)* where the learning agent tries to infer the underlying reward function, based on an expert's behavior (Ng, Russell, et al., 2000; Ramachandran and Amir, 2007).

Hierarchical RL allows for learning policies at multiple levels of granularity.

Additionally to existing RL methodologies *hierarchical RL* is an approach where an MDP is formulated on multiple time scales or levels of detail. For example, there can be sub-policies (or *options*) that try to accomplish sub-goals set by a higher-level policy. This can enable the agent to represent high-level actions to reach a goal and the lower level motor commands needed to execute them (Barto and Mahadevan, 2003).

Environments and reward functions can enforce multi-task learning, and multiple RL agents that share an environment can learn collaborative or competitive behavior.

Lastly, there are different setups in which RL agents can learn. For example *multi-task* and *meta RL* aim to solve several different tasks and transfer knowledge between the different task domains (Vithayathil Varghese and Mahmoud, 2020; Yu et al., 2020). Reinforcement learning agents can also learn in a collaborative or competitive setting which is called *multiagent RL* (Busoniu, Babuska, and De Schutter, 2008). One specific subdomain of this is *self-play* where an agent learns by interacting with itself, for example, by playing chess matches against itself (Silver et al., 2017).

Overall the field of reinforcement learning is very active with many interesting and promising research directions. It has even been suggested that a simple reward signal in a complex enough environment is sufficient to create a wide array of intelligent behaviors indicating that RL may be enough to create an artificial general intelligence (Silver et al., 2021).

6

DRL AS A MODEL OF THE BRAIN?

"With these two-year-olds, as with scientist, finding the truth is more than a profession — it's a passion. And, as with scientists, that passion may sometimes make them sacrifice domestic happiness."

— The Scientist in the Crib (Gopnik, Meltzoff, and Kuhl, 1999)

Reinforcement learning has strong ties to models in behavioral psychology and some areas of neuroscience.

Reinforcement learning has originated out of the combination of two separate fields: The study of optimal control, usually not involving learning, and the study of learning by trial-and-error in behavioral psychology (Sutton and Barto, 2018). While being inspired by learning mechanisms in nature, RL has also provided a useful formal framework for research in that field, and its general principles have been successfully applied back to explain behavioral and neurological data. Today, RL-based models play an important role in the fields of psychology and neuroscience (Botvinick et al., 2020; Sutton and Barto, 2018).

The objective function in reinforcement learning of maximizing rewards seems to be a plausible mechanism to be implemented in animals as well.

Regarding the learning signal, reinforcement learning definitely seems more biologically plausible than supervised learning. Also regarding unsupervised learning, where the statistics of an unlabeled data set are extracted, reinforcement learning may be more realistic as the unsupervised extracted representations are not tuned or optimized for any behavior. The brain contains task and behavior optimized encodings that do not merely reflect input statistics (Freedman et al., 2001; Schoups et al., 2001; Sigala and Logothetis, 2002). Thus, reinforcement learning is an excellent middle ground between fully supervised learning and completely unsupervised extracting of statistics.

Deep RL combines the sensory processing capabilities of deep learning and the reward-based behavior optimization of RL.

Particularly deep reinforcement learning is an interesting parallel to the brain as it represents a complete sensorimotor system. The function approximation of deep learning enables the system to deal with complex, high-dimensional input. The reinforcement learning part enables the system to perform actions and optimize them for goal-directed and exploratory behavior. Thus, sensory and motor skills are learned end-to-end and can influence each other.

6.1 BIOLOGICAL PARALLELS AND DIFFERENCES

6.1.1 *Behavioral Psychology*

Operant and classical conditioning have a large overlap with the principles behind policy and value function optimization.

The idea of learning by trial-and-error originated from behavioral psychology and was first proposed by Edward Thorndike (1911) in his seminal work *Animal intelligence: Experimental studies*. In this, he states *"The Law of Effect is that: Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the*

animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond." (p. 244). This idea is very similar to the basic mechanism behind reinforcement learning and the definition of operant conditioning in behavioral psychology. In operant (and instrumental) conditioning, the consequences of behavior guide what is being learned, and the behavior determines the learning signal. Essentially, the learner is rewarded or punished for certain behaviors and, through this, learns to behave more optimally the next time (Skinner, 1938). Classical condition is a related concept, but here, the rewards and punishments are independent of the learners' actions, such that general stimulus associations are learned (Pavlov, 1927). One could roughly compare classical conditioning to learning a value function or a predictive model of the world and operand condition to optimizing a policy (Sutton and Barto, 2018).

The TD learning-inspired TD model is particularly good at explaining findings from classical conditioning experiments.

Several RL-based models have been proposed to explain results from classical conditioning experiments, such as the Rescorla-Wagner model (Rescorla and Wagner, 1972) and the TD model (Sutton and Barto, 1987). Especially the TD model, which uses the learning rule of TD-learning and adds a temporal component to the Rescorla-Wagner model, can explain many of the phenomena observed in classical conditioning experiments such as the interstimulus interval dependency, blocking, higher-order conditioning, the primacy effect, and serial compound conditioning (Sutton and Barto, 1987, 1990). Also properties of operant conditioning such as shaping (faster learning when gradually reinforcing steps towards the desired behavior) are shared with RL agents (Ng, Harada, and Russell, 1999; Skinner, 1951). This makes RL an interesting framework to model and explain behavior.

Latent learning in animals describes the acquisition of a world model without rewards which can assist in model-based behavior.

There are also parallels between model-based reinforcement learning and concepts in psychology, such as latent learning of cognitive maps and goal-directed behavior. The idea behind latent learning is that properties about the environment, such as the spatial layout of a room, are being learned even when there are no rewards. This 'passively' acquired information can then actively be utilized when a rewarding goal is introduced to produce efficient, goal-directed behavior (Blodgett, 1929). To accomplish latent learning, the agent must be able to learn not only stimulus-response associations like in operant conditioning but also stimulus-stimulus associations and thereby learn cognitive maps that can be repurposed in the presence of goals (Sutton and Barto, 2018; Tolman, 1948). This is similar to learning a model of a Markov decision process and then using this model in model-based RL.

In general, model-free and model-based RL can be roughly compared to habitual and goal-directed behavior, respectively, even though there can be considerable overlap (Sutton and Barto, 2018). For instance, model-free and model-based mechanisms can complement

each other in different habitual behaviors (Morris and Cushman, 2019) and goal-directed behaviors can gradually become habitual (Yin and Knowlton, 2006).

6.1.2 *Dopamin as a Learning Signal*

Dopamin functions as a learning signal for reward-based learning in the brain.

Additionally to the behavioral evidence for reinforcement learning, there is evidence from neuroscience, and reinforcement learning inspired models have been used to explain many experimental findings.

The neurotransmitter dopamine is seen as a central player for reward-based learning in the brain. Experiments show that stimulating dopaminergic neurons acts similar to a reward (sometimes even more rewarding than external rewards) and can be used to perform classical and instrumental conditioning (Olds and Milner, 1954; Saunders et al., 2018; Tsai et al., 2009)

Dopamin may transmit reward prediction errors to enable learning using similar principles as TD learning.

Additional experimental evidence suggests that dopamine may not actually act as a reward but instead as a prediction error between the expected future reward and the obtained reward, similar to the TD error (Montague, Dayan, and Sejnowski, 1996). This is called the *reward prediction error hypothesis* and is one of the prevailing theories of how reward-based learning occurs on the neuronal level through an RL-like algorithm, by learning a value estimate in the frontal cortex and the basal ganglia, instructed by dopaminergic neurons from the midbrain (Glimcher, 2011; Sutton and Barto, 2018).

This hypothesis explains many experimental findings of the phasic activity of dopaminergic neurons in a simple and elegant way (Colombo, 2014; Hollerman and Schultz, 1998; Romo and Schultz, 1990; Schultz, Dayan, and Montague, 1997). The neural activities do not just align with simple prediction errors like in the Rescorla-Wagner model but also incorporate precise stimulus-response timings as only predicted by the TD model (Ljungberg, Apicella, and Schultz, 1992; O'Doherty et al., 2003; Sutton and Barto, 2018). For instance, while there is an increase in firing rate when an unexpected reward is obtained (positive TD error), there is also a decrease in firing rate at the exact timing at which the agent expects a reward if it does not obtain one (negative TD error) (Schultz, Apicella, and Ljungberg, 1993; Schultz, Dayan, and Montague, 1997). Additionally, after training through conditioning, the dopaminergic neuron activity will disappear at the timing of the reward (if it was predicted and obtained) and shifts earlier in time to the appearance of the conditioned stimulus, which predicts the reward (Ljungberg, Apicella, and Schultz, 1992; Schultz, Dayan, and Montague, 1997). Furthermore, inducing prediction error activations artificially in the dopaminergic neurons of the midbrain suffices for strengthening cue-reward associations (Steinberg et al., 2013). Also negative prediction errors can be emulated by suppressing the activity of dopamine neurons (Chang et al., 2016).

There are some discrepancies between the TD model predictions and experimental recordings, for example, when the reward is presented earlier than expected (Hollerman and Schultz, 1998). These can, however, be mitigated by using more complex stimulus representations, modeling eligibility traces, or more complex statistical modeling of the temporal dynamics in the raw sensory input to obtain the stimulus representation of a partially observable state (Daw, Courville, and Touretzky, 2006; Ludvig, Sutton, and Kehoe, 2008; Sutton and Barto, 2018). Overall, predictions of the TD model match well with many recordings of neuronal activity in the brain.

Reward-related activity has been observed in sensory regions as early as V1 (Shuler and Bear, 2006). Experiments by Weglage et al. (2021) indicate that striatal projection neurons of all three main striatal output pathways in the mouse encode a stable, context-specific action space as well as value information and task progress. Furthermore, the action-value encodings are constant within a task but change between tasks (Weglage et al., 2021). Many of the experimental findings listed above suggest that reward and reward prediction errors play an essential role in learning in various brain regions.

6.1.3 *Meta RL and Hierarchical RL in the Brain*

Dopamin may serve more functions than the TD error in RL.

Additionally, dopamine seems to have a much broader functionality in the brain than just associative learning of reward values, and the classical RPE may play additional roles in the brain (Barter et al., 2015; Diederen and Fletcher, 2021). Alexander and Gershman (2021) suggest that the RPE signal is also used for more general representation learning by adapting parameters of the activation function of neurons. This can, for example, be achieved by changing the neuron's receptive field size and determining how wide of a range of patterns it responds to, demanding more precision in response to negative prediction errors and moving the center of receptive fields towards positive prediction errors. These mechanisms would help explain the role of the dopamine system in the representation of time and space, categorization behavior, working memory, abstract reasoning, and motor control (Alexander and Gershman, 2021).

In a new theory of reward-based learning proposed by Wang et al. (2018), phasic dopamine is used to train the prefrontal cortex as a kind of meta-learning system, able to solve many interrelated tasks with its own learned RL mechanism. This theory of learning how to learn matches with findings that motivated the previous RL-based theory of dopaminergic function but also incorporates some conflicting experimental findings (Wang et al., 2018).

The basic ideas of hierarchical reinforcement learning also seem to be supported by neurological data so far and help interpret some of the neural activations found in the brain (Botvinick, 2012). The brain might deal with the body's extremely high-dimensional action space by learning effector-specific values instead of values for every possible movement in space resulting from a combination of multi-

ple effectors. This evidence implies that a single, global scalar value for the prediction error may be too simplistic and that the brain can deal with separate reward and value functions (Gershman, Pesaran, and Daw, 2009). However, recordings in mice have found surprisingly uniform activation in dopamine neurons corresponding to a scaled prediction error, at least in very simplistic conditioning tasks (Eshel et al., 2016). Overall, there are still many outstanding questions on what the learning signal in the brain exactly represents and how it guides learning the complex perceptual, and behavioral problems brains need to solve.

6.1.4 *Actor-Critic in the Brain*

Evidence suggests a form of actor-critic learning to take place in the striatum.

Some evidence suggests that the brain implements an actor-critic learning procedure with the actor and critic located in the dorsal and ventral striatum, respectively (O'Doherty et al., 2004; Takahashi, Schoenbaum, and Niv, 2008). As explained in Sutton and Barto (2018), both the actor and the critic would receive state encodings from different cortical areas and use the TD error as a learning signal. The actor in the dorsal striatum would try to keep the TD error positive by adjusting the action probabilities encoded in its connections to the motor cortex (similar to instrumental conditioning). The critic in the ventral striatum would try to keep the TD error close to zero by making its value estimates as accurate as possible (similar to the TD model of classical conditioning).

Same as in the actor-critic setup shown in figure 8, the actor would not have direct access to the reward, and the TD error affects the dorsal and ventral striatum in different ways. This is evidenced in the brain by dopamine neurons having different effects on different target areas (Lammel, Lim, and Malenka, 2014). In the actor-critic model of the brain, the dopamine neurons in the ventral tegmental area (VTA) and the substantia nigra pars compacta (SNc), which are thought to transmit the TD error, project to both the ventral and dorsal striatum and affect their synaptic connections to the frontal cortex in different ways (Sutton and Barto, 2018; Takahashi, Schoenbaum, and Niv, 2008). Evidence from artificial agents shows that when training an actor using the TD error produced by a critic, very similar learning and behavior as observed in animals emerges (Suri and Schultz, 1999).

6.1.5 *Model-Free and Model-Based Processes in the Brain*

Model-based learning and planning in the brain is supported by the prefrontal cortex, the hippocampus, and possibly the dorsomedial striatum.

Besides the model-free, actor-critic type learning, the brain can also use some form of model-based learning. Especially the orbitofrontal cortex has been suggested to be involved in goal-directed behavior (Gremel and Costa, 2013; McDannald et al., 2011, 2012; Valentin, Dickinson, and O'Doherty, 2007) but also parts of the dorsal striatum seem to perform model-based processing. Specifically, the dorsomedial striatum may be crucial for goal-directed planning while the dorsolateral striatum takes on model-free functions like the actor in the

actor-critic setup (Gremel and Costa, 2013; Johnson, van der Meer, and Redish, 2007; Yin and Knowlton, 2006).

Additionally, the hippocampus's role in navigation indicates that it might represent state transitions of the environment that can then be used as a model to simulate possible future action-trajectories which are evaluated in the ventral striatum to select the best one (Johnson and Redish, 2007; Meer, Kurth-Nelson, and Redish, 2012; Ólafsdóttir et al., 2015; Pfeiffer and Foster, 2013; Stoianov et al., 2018). Even though it is pretty sure that model-free and model-based processes both take place in the mammalian brain, it is not clear how much overlap exists between them and if they are actually implemented in different neurological substrates (Doll, Simon, and Daw, 2012).

6.1.6 *Conclusion*

Overall the fields of reinforcement learning and the study of learning in the brain have and still do influence each other greatly (Botvinick et al., 2020; Dayan and Niv, 2008). Of course, an exact resemblance between RL, how it is today, and the brain will be unlikely, but the coevolution of the two fields makes both very interesting and promising avenues into the broad question of how (reward-based) learning works.

6.2 CURRENT CHALLENGES AND SOLUTION APPROACHES

> *"It was a human blind spot. We look at the world around us as a snapshot when it was really a movie, constantly changing."*
>
> — Prey (Crichton, 2002)

6.2.1 *Learning from Dynamically Changing Data (non-i.i.d.)*

Reinforcement learning takes place under non-i.i.d. conditions. When combined with deep learning, this can be problematic and make learning more unstable.

The design of a reinforcement learning environment is usually, by nature, non-i.i.d. This is because the agent actively samples the environment and the agent's policy determines the data distribution it receives for learning. Since the policy for data collection changes over time (in online-RL), the data distribution also changes. This can cause the learning to become extremely unstable, especially when working with deep neural networks. One way to alleviate the problem employed by Mnih et al. (2015) is to collect a buffer full of experiences and then update the neural network on random batches of experiences sampled from this buffer. This method is called *experience replay* and is inspired by activity patterns in the hippocampus during sleep that seem to replay recent experiences (Mnih et al., 2015; O'Neill et al., 2010; Skaggs and McNaughton, 1996). Nevertheless, training deep RL agents is still very unstable, and the performance between multiple runs, only differentiated by the random seed used to initialize the weights, can diverge strongly (Henderson et al., 2017; Irpan, 2018). An agent with a bad policy can easily perform worse than a random agent, and if this bad policy is also used for experience collection, it

can be challenging for the agent to recover since it will only receive bad quality data to learn from.

6.2.2 *Reproducibility and Hyperparameter*

Deep RL can be unstable, sensitive to hyperparameters and random initializations, and generally have a high variance.

Additionally, RL agents can be sensitive to the choice of hyperparameters, and the often long training times, combined with the already unstable training, can make them extremely difficult to tune (Ibarz et al., 2021). This divergence in performance between different hyperparameters or even just random seeds can make it difficult to study deep RL and gather reproducible results (Henderson et al., 2017). The fact that humans and other animals mostly function well in this world, despite high variance between individuals, suggests that more stable learning mechanisms must be employed.

6.2.3 *Exploration vs. Exploitation*

Finding the right balance between exploration and exploitation is a crucial problem that brains and RL agents both face.

Another fundamental problem of any real or simulated agent is the exploration-exploitation trade-off. It is the choice between exploring unknown parts of the environment in the hope of finding high-reward states and exploiting the rewarding states that are already known to the agent. In the RL framework, this can be seen as the amount of stochasticity or explicit novelty seeking in the behavioral policy. An agent needs to find some kind of middle ground between exploiting known reward states and exploring unknown states to find higher cumulative rewards. How to know the optimal solution to this trade-off in a given environment is an open question (Sutton and Barto, 2018).

Gopnik (2020) suggests that humans (and other animals) partially approach this problem in a way comparable to ML algorithms that change from a more exploring to a more exploiting approach over time (e.g., by lowering the temperature parameter in simulated annealing). During the protected period of childhood, little productive output has to be produced yet, and there is much space for a broad exploration through play and probabilistic hypothesis testing. Later in life, the solid foundations acquired in childhood can be exploited, and most work requires little exploration or revision of hypotheses (except that of certain professions in society such as scientists) (Gopnik, 2020). Out of balance explore-exploit behavior such as unreasonable risk-seeking or overly sticking to the familiar is part of many psychiatric disorders and can be influenced by different drugs (Addicott et al., 2017).

6.2.4 *Sample Efficiency*

Reinfocement learning is less sample efficient than supervised learning as the learning signal is less informative. Solution approaches include episodic- and meta-learning.

Even though most RL setups do not rely on a large amount of labeled data, they are still often sample inefficient and require millions of steps to learn a single task. This can also be framed as an inductive bias problem where a good inductive bias helps to solve a particular

class of problems very fast. On the other hand, weak inductive biases like those of ANNs can solve a wide range of problems but require many samples. Additionally, the reward in RL is not as informative as a label in supervised learning as it does not contain information on what the "correct" action or action sequence would have been. However, there are promising approaches to make deep RL more sample efficient such as episodic memory (using a memory of past solutions to similar problems to solve the current problem) and meta-learning (training a meta learner to learn how to learn a wide range of solutions fast) (Botvinick et al., 2019).

Episodic RL can be compared to research that shows humans using exemplars of classes to categorize new instances of an object with similar features (alongside rule-based categorization) (Rouder and Ratcliff, 2006). In addition, there is some evidence that the hippocampus may play a role in retrieving episodic memories and, through this, (often unconsciously) influencing decisions in the prefrontal cortex using values associated with similar contexts (Bornstein and Norman, 2017; Weilbächer and Gluth, 2016; Wimmer and Shohamy, 2012). Overall, it seems to be a helpful learning mechanism, intertwined with model-based and model-free learning and particularly helpful with large, continuous state spaces that cannot be fully explored and contain long-range dependencies over time (Gershman and Daw, 2017).

That brains also perform meta-learning by learning-to-learn has been suggested both on the behavioral level (Harlow, 1949) and the neuronal level (Wang et al., 2018) and may be crucial to our ability to learn and generalize quickly and efficiently (Griffiths et al., 2019; Wimmer and Shohamy, 2012).

The combination of episodic RL and meta RL principles shows promising results, resembling decision making in human subjects (Ritter et al., 2018).

Other mechanisms may also aid sample-efficient learning. For instance, (Lampinen et al., 2021) show that introducing a hierarchical attention gated memory can help solve tasks that require the agent to retain information collected in the past. This includes object impermanence, finding hidden objects, rapid word learning, and generalizing previously learned knowledge to new tasks.

6.2.5 *Credit Assignment - Dealing with Delayed and Sparse Rewards*

RL agents have difficulty learning long-range dependencies and assigning credit from sparse, possibly delayed rewards.

Another problem that RL agents need to solve is credit assignment for delayed and sparse rewards. If an agent only receives a reward after achieving its goal, it can be difficult to assign credit or blame to the many actions leading up to it. Which action sequences were helpful to achieve the goal and which were useless or even harmful? When using discounted rewards, there is an implicit assumption that the most recent actions were the most influential on the actual outcome. However, when there are long-term dependencies and delayed rewards, it can take much time to learn these.

Especially sparse reward tasks where the reward function contains no information about the value of intermediate states (like distance to the goal) can take a very long time to learn. In the extreme case, if it is unlikely that a random policy ever reaches the goal, the task may never be solved at all. Similar to operant conditioning in animals, it can be helpful to reward approximations of the desired behavior, using *reward shaping* (Sutton and Barto, 2018).

The reward function determines what the agent will learn and may lead to undesired behavior if not carefully defined.

It can be extremely challenging to find the optimal reward function for an agent to learn the desired behavior. When the reward function is not optimal, the agent can quickly learn to exploit flaws in the reward function (also called *reward hacking*) such as repeatedly achieving a small intermediate goal instead of moving on to the desired final goal (Ibarz et al., 2021). As even early myths like that of King Midas show (who wished that everything he touched turned to gold, which the god Dionysus, similar to an RL agent but probably for different reasons, took quite literally by making him turn even his loved ones and his food into hard gold (Fry, 2017)), it can be complicated to express what we want. RL agents do not have common sense (as Midas expected from Dionysus) to interpret the developer's wishes and do anything that maximizes the specified reward function. In general, how the reward function is defined strongly influences what the agent will learn.

6.2.6 *From Single Reward Function to Multi-Task Learning*

Learning multiple tasks can be challenging but can also enable faster learning through knowledge transfer and acquiring more general representations.

In the case of reinforcement learning with multiple goals and only a global reward signal, the credit assignment problem poses another major difficulty. However, there are solutions proposed for figuring out which behaviors were responsible for solving sub-goals of an environment to learn independent behaviors for solving the different tasks (Rothkopf and Ballard, 2010).

Getting an agent to learn multiple tasks can be advantageous. For instance, RL agents trained on a single reward function are also vulnerable to adversarial attacks that exploit weaknesses in the learned policy (Huang et al., 2017). First results indicate that the more tasks an agent is trained on, the more robust it is against adversarial attacks on individual tasks (Mao et al., 2020). This suggests that part of the problem of adversarial vulnerability is the simplistic, single-task objective function being optimized in most setups.

Already training object detection on more classes leads to learning more general, meaningful representations instead of memorizing the classes, which helps learn new classes with fewer examples (Michaelis, Bethge, and Ecker, 2020). Learning on a wide variety of tasks and objective functions also leads to more general skills and zero-shot transfer capabilities to new tasks outside of the training distribution (Team et al., 2021). This work uses additional learning mechanisms such as propagating knowledge through generations of agents, but it is a beautiful example of how complex behavior and

general representations can originate from learning in a complex and varied environment.

Similar to meta-RL, where new tasks can be acquired quickly using knowledge about how previous tasks were solved, multi-task RL also aims to improve sample efficiency by sharing knowledge between simultaneously learned tasks (Yu et al., 2019). Additionally, the global reward signal can be complemented with attention mechanisms to make learning more efficient and to focus plasticity on early layers where credit assignment is difficult (Roelfsema and Ooyen, 2005).

It is not clear how complex a reward function needs to be to give rise to human cognition or what our objective function(s) look like.

Since the brain has to solve a wide variety of tasks over its lifespan, the reward functions may fluctuate with time and context and even between different brain areas. Therefore, using a more heterogeneous cost function that can be composed of unsupervised and supervised aspects at multiple time scales may be crucial for our general abilities (Marblestone, Wayne, and Körding, 2016). Marblestone, Wayne, and Körding (2016) suggests that *"there is not one mechanism of optimization but (potentially) many, not one cost function but a host of them, not one kind of a representation but a representation of whatever is useful, and not one homogeneous structure but a large number of them. All these elements are held together by the optimization of internally generated cost functions, which allows these systems to make good use of one another."* (p. 31). On the other hand, Silver et al. (2021) suggest that also a relatively simple reward signal may suffice to learn complex behavior if the environment is sufficiently complex and requires a variety of skills to maximize the reward function optimally.

6.2.7 *Open-Ended, Self-Supervised Learning*

Open-ended learning and play is important in animal learning and also provides benefits for learning in artificial agents.

Children and several other young animals across the animal kingdom, from monkeys to cats, rats, certain birds, fish and frogs, lizards, turtles, and all the way to some spiders, engage in open-ended, creative play (Burghardt, 2015) (adults too sometimes, but more rarely). In play, by definition, there is no specific task that needs to be accomplished. Instead, the individuals curiously interact with the environment and experiment, for instance, by setting themselves inconsequential goals such as kicking a ball of yarn across the room. This kind of exploratory play with no external reward or incentive is crucial in healthy child development and learning and has to be important enough to outweigh the cost of energy consumption and not actively contributing to society (Gopnik, 2016). It begs the question, whether internal motivations such as curiosity could also be a fundamental building block for artificial intelligence to help learn a general model of the world through self-guided exploration, which can then be applied to efficiently solve specific tasks once they arise.

Using internal rewards has the advantage of surpassing many of the problems mentioned above, such as specifying a proper external reward function that leads to learning the desired behavior. Also the time it takes to learn a new task could be improved if task-unspecific information can be acquired beforehand without any external learn-

ing signal. Additionally, learning can happen in a rich environment with a plethora of different objects without requiring detailed human labeling. Rich, varied learning data may be critical to some of the problems of ANNs, such as generalization and few-shot learning (Michaelis, Bethge, and Ecker, 2020; Team et al., 2021). Finally, internal rewards such as curiosity can be used as incentives to increase exploration and escape local minima in the loss function. Here the exploration is not just some added stochasticity in the policy but rather a directed exploration into novel states.

There are many approaches to self-supervised learning. However, I will limit my focus on learning through prediction and using artificial curiosity as internal rewards.

6.2.8 *Learning through Prediction*

Prediction signals are found in many parts of the brain and may aid in efficient information processing and transmission.

As elaborated in Liu et al. (2021), predictive coding is interpreted in different ways in the literature. One idea is that prediction is used to efficiently encode information by discarding all the predictable parts of the input and only propagating the surprising elements. Another approach is to encode the most predictive information for future states. The latter view centers on encoding the most behaviorally relevant information, such as information that allows the agent to estimate the movement of objects. Both aspects may play a role in the information processing of the brain (Liu et al., 2021). Hawkins and Ahmad (2016) suggest that prediction may not just be happening by communication between neurons but also within a neuron's dendrites. Here, being in a predictive state increases the likelihood of information being passed onward from one neuron to another by creating an action potential. Ali et al. (2021) demonstrate that predictive coding may be a general principle of efficient, hierarchical information processing that does not need to be explicitly wired into a neural network. They show that when optimizing energy efficiency, properties of predictive coding emerge by themselves in recurrent neural networks learning on image sequences (Ali et al., 2021). Overall, prediction is a common theme in efficient information processing and is most likely implemented in the brain in some way or another.

A popular efficient-coding-through-prediction framework in neuroscience proposed by Rao and Ballard (1999) hypothesizes that sensory input to the brain is encoded as predictions and prediction errors through a hierarchical structure with feed-forward and feed-back connections. The idea is that the top-down connections of one area constantly make predictions about the input. The difference between the prediction on the lowest level of processing and the actual sensory input is the prediction error. The prediction error will then be the input (bottom-up) to the next prediction module. This means that only unpredicted elements are being propagated forward in the processing hierarchy. This way of encoding information is incredibly efficient, and when applied to visual input, it produces response properties similar to those found in the visual cortex (Rao and Ballard, 1999).

There is evidence for predictive coding in the primate brain as early as the retina's ganglion cells where predictive motion information is passed onward, and unpredictive information is discarded (Liu et al., 2021). Activations in V1 of mice are predictive of the subsequent observations and increase when the prediction is violated (Fiser et al., 2016). Additionally, the predictions are not just dependent on the current visual input but also the location of the mouse and are learned through experience (Fiser et al., 2016).

The idea of learning models of the world by predicting future inputs and then comparing the predictions to the actual input is interesting as it requires no direct supervision. It has been adopted in machine learning research in various domains, leading to performance increases and replicating phenomena and representations found in the brain (from single-neuron response properties to perceptual motion illusions) (Lotter, Kreiman, and Cox, 2016; Lotter, Kreiman, and Cox, 2020; Oord, Li, and Vinyals, 2018; Wen et al., 2018). For instance, experiments by Singer et al. (2018) demonstrate that optimizing a network to predict its successive inputs leads to representations that resemble those found in V1 and A1 for visual and auditory input, respectively. Furthermore, models with a higher prediction performance also had more similar representations to the brain (Singer et al., 2018). Furthermore, predictive models for language have been shown to match neural representations better than static embeddings and seem biologically plausible (Goldstein et al., 2021).

Predictive models can also be used to plan into the future and aid goal-directed behavior.

Learning to predict dynamic changes over time can help learn appearance invariant representations that can generalize to new objects and help disentangle objects and backgrounds from each other. Moreover, if the predictions are action-conditioned, they can additionally serve for goal-directed planning (Finn, Goodfellow, and Levine, 2016).

The free energy principle is a formal framework describing a system's desire to minimize surprise by improving its predictions or matching its behavior to them.

In the theory of *active inference* the idea of predictive coding and minimizing the prediction error is generalized to motor outputs (Friston et al., 2016). In this framework, the prediction error is described as *surprise*, approximated by *variational free energy*, and needs to be minimized, either by updating the internal model of the world or by acting to match the current predictions (Schwartenbeck et al., 2013). The idea that biological systems minimize surprise by learning hierarchical, generative models of the world that try to make accurate predictions about future states is described in the *free energy principle* (Friston, Kilner, and Harrison, 2006; Friston, 2010). Minimizing long-term surprise requires actively seeking information to improve the current internal model, which leads to curious behavior and exploration (Schwartenbeck et al., 2013).

Friston et al. (2017) demonstrate that using active inference with deep hierarchical models can model eye movements and effects known from biological systems such as mismatch negativity and P300 appear. When using the free energy principle under a hierarchical generative model for visual navigation in a robot, aspects such as path integration and localization naturally emerge (Çatal et al., 2021). However, the free energy principle has some constraining assumptions which

make its applications to complex systems as we find in nature infeasible (Aguilera et al., 2021; Paolo, Thompson, and Beer, 2021). Additional criticisms include it being unfalsifiable and lacking explanatory power (Colombo and Wright, 2021; Gershman, 2019). Nevertheless, it is an interesting, influential, and unifying framework for understanding action and perception.

6.2.9 *Curiosity as Intrinsic Motivation*

Curiosity can serve as a strong intrinsic motivation in humans, animals, and artificial agents.

An exciting example of learning from intrinsic rewards is curiosity. Here, the agent receives a reward when it encounters novel states that are unknown to its internal model of the world. The novelty of a state can be assessed in different ways. For instance, the agent may be learning a predictive model of the world and judge the novelty of a state by how good its predictions are (well-known states have better predictions than unknown states). The idea is to encourage actions that minimize future uncertainty and to sample information that will improve the internal model to make better future predictions (Schmidhuber, 2010; Schmidhuber, 1991).

Curiosity may be essential for efficiently learning a model without external supervision.

Curiosity may play an essential role in the latent learning of cognitive maps observed in animals, where the animal acquires general information about an environment without a specific task which it can then later use for goal-directed behavior (Wang and Hayden, 2021).

It is already long known that animals all across the animal kingdom engage in curious behavior, exploration, and play in a way that cannot simply be explained by the pursuit of external rewards or avoidance of punishment (White, 1959). In fact, animals and humans even accept moderate amounts of pain or forgo an immediate reward in order to explore and obtain new information (Hsee and Ruan, 2016; Nissen, 1930). Already in the 1920s, Dashiell noted that *"one who has handled white rats knows well enough that when these animals are returned to a renovated nest box they pay little or no attention to food placed there even though unfed for twenty-four hours, but give themselves up for a while to explorations over and through their new bedding."* (Dashiell, 1925, p.208) The exploratory drive in rats is so strong that they even cross an electrified area to reach novel parts of the environment, despite the absence of any other drives such as hunger, socializing, or exercise (Nissen, 1930).

Already the ancient Greek myth of Pandora's box demonstrates how humans go to great lengths to acquire new information and even risk adverse outcomes to resolve uncertainty and collect new information. For example, when presented with some pens of which some emit an electric shock when pressed, people who do not know which of the pens are rigged tend to press them more often than people who do, even though there is no possible positive outcome associated with pressing the pens (Hsee and Ruan, 2016). Of course, this curious drive only applies up to a certain point. *"One does not know exactly how it feels putting one's hand into the meat grinder. However, one does not want to know."* (Schmidhuber, 1991, p.4). Learning usually

happens from a combination of extrinsic and intrinsic motivations. We can balance simultaneously seeking new information and trying to avoid painful events.

Intrinsic motivations such as curiosity may play a crucial role in a variety of aspects of human cognition. *"Behavior [that cannot be successfully conceptualized in terms of primary drives] includes visual exploration, grasping, crawling and walking, attention and perception, language and thinking, exploring novel objects and places, manipulating the surroundings, and producing effective changes in the environment. The thesis is then proposed that all of these behaviors have a common biological significance: they all form part of the process whereby the animal or child learns to interact effectively with his environment. (...) Further, it is maintained that competence cannot be fully acquired simply through behavior instigated by drives. It receives substantial contributions from activities which, though playful and exploratory in character, at the same time show direction, selectivity, and persistence in interacting with the environment. Such activities in the ultimate service of competence must therefore be conceived to be motivated in their own right."* (White, 1959, p.329).

Curiosity in deep RL agents can lead to learning complex behaviors without external rewards.

Pathak et al. (2017) recently demonstrated the effectiveness of curiosity in artificial agents. In the study, the agents learn two dynamics models of the environment: a forward and an inverse model. They then use the error of the two models as the reward for the agent's policy. The forward model predicts the next state based on the current state and the current action. The inverse model learns to infer the action that was performed between the last state and the current state. The inverse model's objective function ensures that the feature space in which the predictions are made contains action-relevant information. Therefore, the agent is not infinitely fascinated by random noise or other unpredictable elements such as leaves blowing in the wind, fire, or ocean waves (even though to me they are still pretty captivating to look at).

Learning the dynamics models is a self-supervised process where the collected experiences are the inputs and labels for learning, and the model tries to minimize the prediction error. Learning the policy is done using reinforcement learning, maximizing the reward obtained from the prediction error. This reward can be sufficient to learn all sorts of video games without any external reward from the environment (Burda et al., 2019a). The agent needs to learn how to progress in the game to satisfy its curiosity since the starting states of the game quickly become very predictable and boring. This method is also used in Clay et al. (2021b) in the research section for learning a model of the world without external rewards.

There are several different ways to express novelty which come with different perks and problems.

One problem with this approach is that if action conditioned stochasticity appears, the agent can get stuck in these stages. This is also called the noisy-TV problem, as it can be tested by adding a TV that shows a random image whenever the agent presses the remote button. Since the change of the image is conditioned on the agent's action but the image that is shown is inherently unpredictable, this will get the agent to be stuck in front of the TV screen (Burda et al., 2019a).

Pathak, Gandhi, and Gupta (2019) offer a solution to this by letting the agent learn an ensemble of forward dynamics models. They then take the disagreement between the models as a reward instead of their prediction errors. The idea is that when the agent is in a novel state, there will be a high variance between the model predictions, but when the agent is in a familiar state, all models will predict very similar things. This setup is able to deal with the noisy TV problem because, after some learning, all models will predict an average of the stochastic next states and therefore have little disagreement even if the prediction error is still relatively high.

Alternatively, novelty can be determined independently of the current action and next state. This also avoids the noisy TV problem and can be easier and faster to learn. Burda et al. (2019b) do this by using a random feature network to encode the observations. The internal model then learns to model the features of the current state by trying to match the random feature network. In familiar states, this will work better than in unfamiliar ones, providing an alternative measure of novelty that is not based on environment transitions.

An attractive technical advantage of these techniques is that the reward function is now part of the agent's model and not the environment anymore. Therefore, errors can be propagated directly through the dynamics models to the policy, making the need for more sample inefficient reinforcement learning obsolete. The agent can directly make use of its knowledge about the environment dynamics to adapt the policy in a way that novel states are reached, which speeds up learning significantly (Pathak, Gandhi, and Gupta, 2019).

Learning through curious exploration helps learn a general world model that can be used for more efficient learning of specific tasks.

Combining a curiosity objective with model-based reinforcement learning can leverage planning to seek out novel states efficiently. The learned world models can then be used to learn many motor control tasks in a zero- or few-shot manner (Sekar et al., 2020).

Representations learned in a curious way need no explicit supervision and can therefore encompass more general, task-unspecific knowledge about the properties of the environment. These representations that are learned through interaction are better suited to learn specific tasks with very few labeled examples or supervised episodes than representations that were acquired without interaction (Clay et al., 2021b; Du, Gan, and Isola, 2021). Nevertheless, first experiments show that children and curious artificial agents still exhibit different exploration behavior. Artificial curiosity formulations have a hard time performing active, efficient, information-seeking experiments and instead revisit randomly found novel states until they become boring (Kosoy et al., 2020).

Goal-directed learning can make exploration more directed and knowledge easier transferrable to new tasks. Goals do not have to be set externally but can also be self-defined.

Exploration in children is not just random or novelty-seeking but can also be goal-directed and purposeful (Meder et al., 2021). In free play, the child can set goals that do not fulfill any extrinsic purpose. For example, a child might try to jump on one leg from one side of the room to the other for no other reason than to figure out if it can. Or it may spontaneously sort blocks by color or try to stack them as high as possible without expecting a reward for doing so (and even risking the higher likelihood of the little brother kicking them over).

Additionally, exploration can be directed at obtaining new information. For instance, when unsure about how a particular toy works, pre-schoolers often design and execute clever experiments that test different possible hypotheses (Cook, Goodman, and Schulz, 2011).

In reinforcement learning, it can also be advantageous to learn in a goal-directed way without specifying explicit goals for the agent. When the agent learns to accomplish an explicit goal, it only starts learning once it actually reaches the goal, which can, depending on the goal, be very unlikely with a random initial policy. Instead of manually applying reward shaping, Andrychowicz et al. (2017) propose a method called *hindsight experience replay* to learn from all the trials where the agent did not reach the goal. Here, the agent treats the outcome of an action sequence as if it was the actual goal (even though it most likely was not) and updates its policy accordingly. This means that even though the agent did not intend to end up in this specific state, it can remember how to get there if this state will ever be a goal state in the future. This type of learning can be crucial for learning complex behaviors in sparse reward environments and for learning more general goal-conditioned policies (Andrychowicz et al., 2017).

With a goal-conditioned policy, goal-setting can also serve as intrinsic motivation. Agents with a compositional representation of goals (e.g., in the form of simple language) can imagine new goals by combining building blocks of known goals, and achieving these self-set goals can then be the agent's intrinsic reward (Colas et al., 2020b). By setting its own goals, the agent can create a learning curriculum that maximizes learning progress by focussing on goals that are just challenging enough for its current skill level (Colas et al., 2019).

Combining intrinsic motivation with goal-conditioned reinforcement learning is part of the new field of *developmental machine learning* with many exciting and promising avenues to get artificial agents to learn more like children (Colas et al., 2020a).

6.2.10 *Conclusion*

Understanding the way learning works is tackled from many different angles with some fruitful and exciting overlaps between them.

Using deep reinforcement learning to learn with a complete sensorimotor loop and to embed perception in an action-oriented context is a promising step towards more natural learning in ANNs. There are still many problems with this approach and several promising research directions working to solve these. However, there is already a lot of interesting overlap and exchange between several disciplines under the framework of reinforcement learning, such as neuroscience, psychology, and developmental robotics.

Learning in social contexts is an essential aspect of human development but will not be addressed here.

One specific aspect of learning that I did not explicitly mention yet and will not look at in the research section is learning in social contexts. This includes, for example, observational learning (often formalized as imitation learning) as well as learning in collaborative and competitive setups. We are social animals and learn a lot through interaction with others (Gopnik, 2016). Social learning and language allow us as a species to pass knowledge down through the gener-

ations and successively build on it. Passing information from one mind to another is a highly complex process, including hallmarks of social intelligence such as theory of mind and considering other's mental states when collecting evidence or teaching others (Gweon, 2021). Even subtle social cues such as eye contact and gaze direction of others play a role in human decision-making (Belkaid et al., 2021). Research in multi-agent reinforcement learning has shown complex behaviors, and even artificial languages to emerge but also comes with its own set of additional problems (Gronauer and Diepold, 2021). Although this field of social learning in a multi-agent environment is fascinating and undoubtedly also contributes to the understanding of human intelligence, it is out of the scope of this dissertation.

In this dissertation, I focus on the effect and the importance of active learning through weakly supervised interaction.

In the following research section, I will look at how learning through embodied interaction impacts what is being learned. First in humans and then in machines, contrasting classical ANN training to training ANNs using deep reinforcement learning. Lastly, I also incorporate artificial curiosity to encourage novelty-seeking exploration and test the trained embodied agents on a fast-mapping task.

Actively exploring the world seems to be crucial for learning in brains and may also be for learning in machines. In the words of Robert W. White: *"Man's huge cortical association areas might have been a suicidal piece of specialization if they had come without a steady, persistent inclination toward interacting with the environment."* (White, 1959, p.330)

Part II

RESEARCH

7

LEARNING OF SPATIAL PROPERTIES OF A LARGE-SCALE VIRTUAL CITY WITH AN INTERACTIVE MAP

Sabine U König, Viviane Clay, Debora Nolte, Laura Duesberg, Nicolas Kuske, and Peter König (2019). "Learning of spatial properties of a large-scale virtual city with an interactive map." In: *Frontiers in human neuroscience* 13, p. 240. DOI: 10.3389/fnhum.2019.00240

8

EYE TRACKING IN VIRTUAL REALITY

Viviane Clay, Peter König, and Sabine König (2019). "Eye tracking in virtual reality." In: *Journal of Eye Movement Research* 12.1. ISSN: 19958692. DOI: 10.16910/jemr.12.1.3.

9

EMBODIED SPATIAL KNOWLEDGE ACQUISITION IN IMMERSIVE VIRTUAL REALITY: COMPARISON TO MAP EXPLORATION

Sabine U. König, Ashima Keshava, Viviane Clay, Kirsten Rittershofer, Nicolas Kuske, and Peter König (2021). "Embodied Spatial Knowledge Acquisition in Immersive Virtual Reality: Comparison to Map Exploration." In: *Frontiers in Virtual Reality* 2, p. 4. ISSN: 2673-4192. DOI: 10.3389/frvir.2021.625548.

10

LEARNING SPARSE AND MEANINGFUL REPRESENTATIONS THROUGH EMBODIMENT

Viviane Clay, Peter König, Kai-Uwe Kühnberger, and Gordon Pipa (2021a). "Learning sparse and meaningful representations through embodiment." In: *Neural Networks* 134, pp. 23–41. DOI: 10.1016/j.neunet.2020.11.004.

11

FAST CONCEPT MAPPING: THE EMERGENCE OF HUMAN ABILITIES IN ARTIFICIAL NEURAL NETWORKS WHEN LEARNING EMBODIED AND SELF-SUPERVISED

This chapter is currently only published as a pre-print. The manuscript has been submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence on February 1st, 2021 and is currently under review there.

Viviane Clay, Peter König, Gordon Pipa, and Kai-Uwe Kühnberger (2021b). "Fast Concept Mapping: The Emergence of Human Abilities in Artificial Neural Networks when Learning Embodied and Self-Supervised." In: *arXiv preprint arXiv:2102.02153*

Part III

DISCUSSION

12

DISCUSSION

"Your scientists were so preoccupied with whether or not they could, they didn't stop to think if they should."

— Ian Malcolm - Jurassic Parc (*Jurassic Parc* 1993)

In the research presented here, we look at the effect of embodiment on learning, first in humans and then in machines.

In the research presented here, we first looked at learning in humans, and the effect of embodiment on the kind of spatial representations acquired. For this, we devised and tested a new methodology to let participants learn in a large-scale virtual city and to measure their body and eye movements inside virtual reality (Clay, König, and König, 2019). We then used this method for a controlled experiment contrasting embodied spatial navigation to learning from an interactive map. The results of this experiment are presented in two papers and demonstrate that learning in the embodied setup indeed leads to different encodings of spatial knowledge (König et al., 2019; König et al., 2021).

Next, we tested whether this effect also transfers to artificial neural networks. To test this, we trained an embodied agent in a sparse reward, vision-based, 3D maze navigation task. We compared the learned representations to those acquired from training on images from the same environment but without embodiment. This included one self-supervised neural network, learning to reconstruct the input, and one fully supervised network, classifying objects in the images. All three networks were identical apart from the task they needed to solve and, therefore, the output dimensionality and objective function. The results confirmed again that learning in an embodied and interactive way leads to learning significantly different representations, which are sparser and action-oriented (Clay et al., 2021a).

Representing the world in an action-oriented way makes sense. There is no evolutionary benefit in representing the world as realistically as possible. Instead, we need to represent the world in a way that helps us act upon it most optimally. If action is disregarded in models of the brain (even models of sensory processing), this aspect is left out. Being able to act well within one's ecological niche is probably the most substantial evolutionary pressure on all aspects of information processing in the brain.

In Clay et al. (2021b), we took a closer look at whether the learned representations may enable capabilities that are difficult to attain for the non-embodied ANNs. We show that the representations of the embodied agent can be used to perform fast-mapping and attain an above chance object detection performance with just one labeled example, further increasing with every additional example. This is significantly better than the autoencoder representations and on-par with the classifier, whose representation was optimized on several million labeled examples.

The exact mechanism underlying the emerging sparsity is still unclear.

A contributing factor to this astonishing performance could be that the agent learns very sparse encodings of the visual input, using much fewer neurons to encode one image than the other two networks. However, even though we performed many control experiments to exclude other explanations besides the difference in the embodiment, the exact cause for this is still unclear (Clay et al., 2021a). We tried to keep all other variables constant to isolate the effect of embodiment. However, some aspects necessarily come with it, such as learning from non-i.i.d data. It may be that the sparsity originates from a detailed aspect of the mechanics of reinforcement learning or from the uninformative, sometimes contradicting, learning signal and not from the embodiment itself. Nevertheless, noisy, uninformative learning signals and a dynamic data distribution would also be aspects of human learning and still support the claim that making the learning setup more natural leads to learning different representations and skills.

Sensorimotor learning can happen without any external supervision by using intrinsic motivation such as curiosity.

Finally, in Clay et al. (2021b) we demonstrated that these sparse and meaningful representations can be learned without any external supervision. For this, we used curiosity as intrinsic motivation as described in Pathak et al. (2017). We, therefore, show that a useful encoding of high dimensional visual input can be learned end-to-end embedded in a sensorimotor system, simply through curious interaction with the world.

Overall, the publications presented here show that learning environment matters, and particularly that embodied, sensorimotor learning leads to learning different representations of the environment than the classical fully-supervised learning setup on static images.

The studies realism could be further increased by including multiagent interactions and using a more interactive, dynamic, life-long learning environment.

Interesting further investigations could include looking at the role of language and social interaction in learning. For instance, does it help an agent to learn names for objects to conceptualize the world better and represent it more efficiently? Does action shape the representation of communicable concepts? Moreover, how does communication with other agents shape their representations of the world?

Additionally, one aspect that was quite unnatural in the last two studies is that the agent learns in an episodic environment. This means that the environment resets, and the agent starts from the start state again after every few thousand steps. This is in strong contrast to the kind of life-long learning that takes place in brains. Also, even though the environment we used involved some random variations in looks and floor plan, it remains relatively stable and constant. The real world is constantly changing, and so are the bodies of children who explore it, such that there can be much more dynamic change in the data distribution over a lifetime. To deal with a complex, ever-changing world, additional mechanisms such as model-based and meta reinforcement learning may be required (Clavera et al., 2018).

Furthermore, our studies focused on learning navigation behavior. Of course, there are many other important behaviors such as interaction with objects and tool use that require a different range of cognitive functions and motor abilities. The RL agent in our studies learned with a limited set of discrete actions in a world that contained few ob-

jects that could be manipulated. Learning behaviors for the complex, continuous action spaces that humans have is extremely challenging and may require hierarchical representations of action routines or subgoals to be solved efficiently (Botvinick, 2012). An agent learning in a complex environment with more interactable objects would likely learn different, more object-centric representations than an agent primarily optimized for efficient navigation.

All these aspects of biological plausibility and levels of implementing artificial intelligence listed in the introduction beg the question: Which are the most critical aspects that should be focused on?

The level of computation is, of course, not the only relevant level to understand information processing.

Even though in this thesis I make a case for the computational aspects, such as the environment and objective function, I would not say that those are the only essential elements. I have tried to argue that already changing the learning setup to more interactive and open-ended learning can lead to vastly different representations of the world (Clay et al., 2021a,b) and some research suggest that remarkable capabilities can emerge from such a setup (Team et al., 2021). However, making the learning environment more natural can only help up to a certain point. To achieve human-like multi-task and transfer learning capabilities that use a robust understanding of the world to generalize quickly and efficiently, adjustments on the level of representation, algorithm, and hardware implementation that support better learning of causality and compositionality with stronger inductive biases for such tasks will most likely be required.

Alternatively, evolutionary methods could be used instead of human-engineered solutions to find good inductive biases for learning in a complex world. Of course, we can try to copy the solutions that billions of years of evolution came up with. However, since we still lack a comprehensive understanding of many of them, it may be a suitable alternative to simulate evolution instead. Khadka and Tumer (2018) show that the combination of inheriting and optimizing good starting conditions across generations through evolution and learning through experience within an individual's lifetime using reinforcement learning can be better than either method on its own.

There are still many open questions on all levels of information processing that can be investigated in the brain and in artificial agents.

Even within the computational level, many remaining open questions make it difficult to move towards more "brain-like" learning in ANNs. For instance, what is the brain's objective function? How much of the learning in brains is supervised, semi-supervised and unsupervised? Also, more broadly across the three levels, big questions remain open. For example, how much knowledge and skill is already hardwired present at birth and optimized through evolution? What is the role of sleep? How does attention influence cognition and learning? And, is attention an essential feature or just a biologically necessary efficiency constraint? These and many more open questions are all fascinating steps towards a better understanding of cognition and learning. Answers to them will further both the fields of neuroscience and psychology as well as artificial intelligence.

The road to human-like intelligence is long, and, as table 1 shows, there are still many differences to current biologically motivated solutions. Artificial neural networks are a helpful tool and an exciting

step towards a model of the brain, but this goes only up to a certain point, as many of their pitfalls show. Moving further towards creating an intelligence similar to ours will most likely involve looking closer at all of the levels of information processing in the brain. Adjusting the computational tasks that need to be solved, for instance, by using new benchmarks, can guide the search for better implementations on the lower levels.

Building more capable AI systems can help us better understand our own intelligence.

In the end, this line of research is not just aiming at building more powerful machines but also at understanding some of the many open questions about the workings of our brain. Studying this AI framework can help us understand human learning and cognition better (Richards et al., 2019). If we can build intelligent machines that can achieve similar tasks as we can, we will be one step closer to understanding our own intelligence. Finding answers to the big mysteries of who we are and how we learn and think the way we do is my main motivation to study this. Whether it is the right path, only time can tell.

13

CONCLUSION

"Nothing is written."

— T.E. Lawrence - Lawrence of Arabia (*Lawrence of Arabia* 1962)

Learning environment and task are always reflected in the learned knowledge of a model.

Throughout the past years, one principle has emerged over and over in my research: What a model learns is a reflection of the task and environment in which it learns. This seems intuitive by itself, but the implications of it can be far-reaching. It means that if we want to understand cognition and our brains, we need to understand the tasks that need to be solved. Moreover, if we want our artificial models to learn brain-like representations, we need to let them learn tasks that brains need to solve.

Disregarding dynamic, sensorimotor, and weakly supervised learning in AI systems leaves out important aspects of our cognition.

This insight is not reflected in many of the current machine learning benchmarks. The collections of static, labeled images for computer vision benchmarks are vastly different from the dynamically sampled visual inputs to a brain. Without interaction, AI cannot actively sample the world or observe how physical objects behave, which means they lack crucial aspects of our cognition. Understanding the brain apart from the environment it is situated in may be impossible.

Throughout this dissertation, I make the case that brains are sensorimotor learning systems, and if we want to model them, our models should be as well. Learning through interaction and embodiment affects what is being learned. I presented research that shows that making learning more natural can improve robustness, generalization, efficiency, and performance. Overall, the effect that tasks have on learning, skills, and the representations of reality can not be discounted and are crucial in any model of learning and cognition.

Part IV

APPENDIX

A

DATA FROM NEURAL NETWORK TRAINING IN THE OBSTACLE TOWER ENVIRONMENT TO INVESTIGATE EMBODIED, WEAKLY SUPERVISED LEARNING

The following is the description of a published dataset on the Mendeley Data repository. The dataset won the *Mendeley FAIRest Dataset Award*.

Viviane Clay (2020). "Data from Neural Network Training in the Obstacle Tower Environment to Investigate Embodied, Weakly Supervised Learning." In: *Mendeley Data*. DOI: 10.17632/ZDH4D5WS2Z.2

DESCRIPTION

This repository presents data collected to investigate the role of embodiment and supervision in learning. This is done inside a simulated 3D maze world with a navigation task using mainly visual input in the form of RGB images. The main contribution of this data repository is to provide a network model trained in this environment with weak supervision and a closed loop between action and perception. Additionally, control networks are provided which were trained with varying degrees of supervision and embodiment. In the corresponding paper [1] the representations of these networks are compared based on sparsity measures and well as content of the encodings and the possibility to extract semantic labels. For the training of the control conditions several new data sets were created which are also included here. They contain a collection of images from the simulated world with corresponding semantic labels (hand labeled). Overall, they provide a good basis for further analysis and a more in-depth investigation of representation learning and the effect of embodiment and supervision on representations.

STEPS TO REPRODUCE

Data was generated through a 3D simulation of a maze environment called Obstacle Tower. The data of interest are the trained neural network weights and the networks activations corresponding with different input frames. Three main networks were trained. A reinforcement learning agent which trained through interaction with the simulated environment, an autoencoder trained to reconstruct images collected by the agent and a classifier, trained to classify objects in the images. Exact training and testing conditions, hyperparameter and network structure are provided in the corresponding paper.

For the training of the reinforcement learning agent the Unity ml-agents toolkit PPO implementation is used with small modifications for extra data collection and control experiments. The code we used can be found here: https://github.com/vkakerbeck/ml-agents-dev. Model checkpoint files are saved for different points in training but mostly the final version of the network is analysed in the corresponding paper [1]. The autoencoder and classifier are trained using Python with TensorFlow and Keras. The corresponding code can be found here: https://github.com/vkakerbeck/Learning-World-Representations/tree/master/DataAnalysis. The data also contains activations in the hidden layer of the network corresponding to 4000 test images for all three networks. Code for this can be found in the same GitHub repository. The datasets used for training the autoencoder and classifier were created by collecting observations in the Obstacle Tower environment using the trained agent. These observations were then labelled automatically, and the labels were cross checked by hand.

A Description of the individual files is included in the data folder (Description.txt). Due to storage constraints not all model checkpoint

files used to create figure 6 of the paper could be uploaded. However, feel free to contact me (vkakerbeck[at]uos.de) if you are intrested in these detailed checkpoint files of the controll runs and I will make them available to you.

[1] Clay, V., König, P., Kühnberger, K.-U. & Pipa, G. Learning sparse and meaningful representations through embodiment. Neural Networks (2020). http://www.sciencedirect.com/science/article/pii/S0893608020303890.

DATASET METRICS

(Recorded on October 15th 2021)
Views: 379
Downloads: 29
Tweets: 1
Categories: Artificial Neural Networks, Reinforcement Learning, Cognitive Science, Cognitive Representation
Version 1 Published: 10 August 2020
Version 2 Published: 5 December 2020

B

A QUANTITATIVE ANALYSIS OF THE TAXONOMY OF ARTISTIC STYLES

This publication did non thematically fit into the dissertation and is therefor only included for completeness in the appendix.

Viviane Clay, Johannes Schrumpf, Yannick Tessenow, Helmut Leder, Ulrich Ansorge, and Peter König (2020). "A quantitative analysis of the taxonomy of artistic styles." In: *Journal of Eye Movement Research* 13.2. DOI: 10.16910/jemr.13.2.5

BIBLIOGRAPHY

Achiam, Joshua (2018). *Spinning Up in Deep Reinforcement Learning*. URL: https://spinningup.openai.com/ (visited on 08/04/2021).

Addicott, M A, J M Pearson, M M Sweitzer, D L Barack, and M L Platt (2017). "A Primer on Foraging and the Explore/Exploit Trade-Off for Psychiatry Research." In: *Neuropsychopharmacology* 42.10, pp. 1931–1939. ISSN: 1740-634X. DOI: 10.1038/npp.2017.108.

Adolph, Karen E. (2000). "Specificity of Learning: Why Infants Fall Over a Veritable Cliff." In: *Psychological Science* 11.4. PMID: 11273387, pp. 290–295. DOI: 10.1111/1467-9280.00258.

Adolph, Karen E., Marion A. Eppler, and Eleanor J. Gibson (1993). "Crawling versus Walking Infants' Perception of Affordances for Locomotion over Sloping Surfaces." In: *Child Development* 64.4, pp. 1158–1174. DOI: https://doi.org/10.1111/j.1467-8624.1993.tb04193.x.

Aguilera, Miguel, Beren Millidge, Alexander Tschantz, and Christopher L Buckley (2021). "How Particular is the Physics of the Free Energy Principle?" In: *Arxiv*. arXiv: 2105.11203v1.

Ahmad, Subutai and Luiz Scheinkman (2019). "How can we be so dense? The benefits of using highly sparse representations." In: *arXiv preprint arXiv:1903.11257*.

Ahnert, Lieselotte (2014). *Theorien in der Entwicklungspsychologie*. Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-34805-1.

Akbas, Emre and Miguel P. Eckstein (2017). "Object detection through search with a foveated visual system." In: *PLOS Computational Biology* 13.10, pp. 1–28. DOI: 10.1371/journal.pcbi.1005743.

Alexander, William H. and Samuel J. Gershman (2021). *Representation learning with reward prediction errors*. arXiv: 2108.12402 [q-bio.NC].

Ali, Abdullahi, Nasir Ahmad, Elgar de Groot, Marcel A. J. van Gerven, and Tim C. Kietzmann (2021). "Predictive coding is a consequence of energy efficiency in recurrent neural networks." In: *bioRxiv*. DOI: 10.1101/2021.02.16.430904.

Andrychowicz, Marcin, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba (2017). "Hindsight Experience Replay." In: *CoRR* abs/1707.01495. arXiv: 1707.01495. URL: http://arxiv.org/abs/1707.01495.

Ansorge, Ulrich, Markus Kiefer, Shah Khalid, Sylvia Grassl, and Peter König (2010). "Testing the theory of embodied cognition with subliminal words." In: *Cognition* 116.3, pp. 303–320. ISSN: 0010-0277. DOI: https://doi.org/10.1016/j.cognition.2010.05.010.

Ariga, Atsunori, Yuki Yamada, and Yusuke Yamani (2016). "Early Visual Perception Potentiated by Object Affordances: Evidence From a Temporal Order Judgment Task." In: *i-Perception* 7.5. PMID: 27698991, p. 2041669516666550. DOI: 10.1177/2041669516666550.

Attinger, Alexander, Bo Wang, and Georg B Keller (2017). "Visuomotor Coupling Shapes the Functional Development of Mouse Visual Cortex." In: *Cell* 169.7, 1291–1302.e14. ISSN: 0092-8674. DOI: 10.1016/j.cell.2017.05.023.

Aytekin, Murat, Cynthia F Moss, and Jonathan Z Simon (2008). "A Sensorimotor Approach to Sound Localization." In: *Neural Computation* 20.3, pp. 603–635. ISSN: 0899-7667. DOI: 10.1162/neco.2007.12-05-094.

Azevedo, Frederico A.C., Ludmila R.B. Carvalho, Lea T. Grinberg, José Marcelo Farfel, Renata E.L. Ferretti, Renata E.P. Leite, Wilson Jacob Filho, Roberto Lent, and Suzana Herculano-Houzel (2009). "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain." In: *Journal of Comparative Neurology* 513.5, pp. 532–541. DOI: https://doi.org/10.1002/cne.21974.

Bright Eyes, Conor Oberst (2020). "Dance and Sing." In: *Down in the Weeds, Where the World once was*.

Bain, Michael and C. Sammut (1995). "A Framework for Behavioural Cloning." In: *Machine Intelligence 15*.

Baker, Nicholas, Hongjing Lu, Gennady Erlikhman, and Philip J. Kellman (2018). "Deep convolutional networks do not classify based on global object shape." In: *PLOS Computational Biology* 14.12, pp. 1–43. DOI: 10.1371/journal.pcbi.1006613.

Ballard, Dana H. (1989). "Reference Frames for Animate Vision." In: *Proceedings of the 11th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI89. Morgan Kaufmann Publishers Inc., 1635–1641.

Baldwin, Dare A. (1991). "Infants' Contribution to the Achievement of Joint Reference." In: *Child Development* 62.5, pp. 875–890. DOI: https://doi.org/10.1111/j.1467-8624.1991.tb01577.x.

Ballard, Dana H. (1991). "Animate vision." In: *Artificial Intelligence* 48.1, pp. 57–86. ISSN: 0004-3702. DOI: https://doi.org/10.1016/0004-3702(91)90080-4.

Ballard, Dana H. and Christopher M. Brown (1992). "Principles of animate vision." In: *CVGIP: Image Understanding* 56.1. Purposive, Qualitative, Active Vision, pp. 3–21. ISSN: 1049-9660. DOI: https://doi.org/10.1016/1049-9660(92)90081-D.

Baldwin, Dare A (1993). "Early referential understanding: Infants' ability to recognize referential acts for what they are." In: *Developmental Psychology* 29.5, pp. 832–843. ISSN: 1939-0599(Electronic),0012-1649(Print). DOI: 10.1037/0012-1649.29.5.832.

Bambach, Sven, David Crandall, Linda Smith, and Chen Yu (2018). "Toddler-Inspired Visual Object Learning." In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., pp. 1201–1210.

Bansal, Somil, Roberto Calandra, Sergey Levine, and Claire J. Tomlin (2017). "MBMF: Model-Based Priors for Model-Free Reinforcement Learning." In: *CoRR* abs/1709.03153. arXiv: 1709.03153.

Banino, Andrea et al. (2018). "Vector-based navigation using grid-like representations in artificial agents." In: *Nature* 557, pp. 429–433. DOI: 10.1038/s41586-018-0102-6.

Barto, Andrew G and Sridhar Mahadevan (2003). "Recent advances in hierarchical reinforcement learning." In: *Discrete event dynamic systems* 13.1, pp. 41–77.

Barter, Joseph W., Suellen Li, Dongye Lu, Ryan A. Bartholomew, Mark A. Rossi, Charles T. Shoemaker, Daniel Salas-Meza, Erin Gaidis, and Henry H. Yin (2015). "Beyond reward prediction errors: the role of dopamine in movement kinematics." In: *Frontiers in Integrative Neuroscience* 9, p. 39. ISSN: 1662-5145. DOI: 10.3389/fnint.2015.00039.

Bartunov, Sergey, Adam Santoro, Blake A. Richards, Luke Marris, Geoffrey E. Hinton, and Timothy P. Lillicrap (2018). "Assessing the Scalability of Biologically-Motivated Deep Learning Algorithms and Architectures." In: *NeurIPS*, pp. 9390–9400. URL: http://papers.nips.cc/paper/8148-assessing-the-scalability-of-biologically-motivated-deep-learning-algorithms-and-architectures.

Battaglia, Peter W., Jessica B. Hamrick, and Joshua B. Tenenbaum (2013). "Simulation as an engine of physical scene understanding." In: *Proceedings of the National Academy of Sciences* 110.45, pp. 18327–18332. ISSN: 0027-8424. DOI: 10.1073/pnas.1306572110.

Baxter, Jonathan (2000). "A model of inductive bias learning." In: *Journal of artificial intelligence research* 12, pp. 149–198.

Beauchamp, Michael, Laurent Petit, Timothy Ellmore, John Ingeholm, and James Haxby (2001). "A Parametric fMRI Study of Overt and Covert Shifts of Visuospatial Attention." In: *NeuroImage* 14, pp. 310–21. DOI: 10.1006/nimg.2001.0788.

Belkaid, Marwen, Kyveli Kompatsiari, Davide De Tommaso, Ingrid Zablith, and Agnieszka Wykowska (2021). "Mutual gaze with a robot affects human neural activity and delays decision-making processes." In: *Science Robotics* 6.58, eabc5044. DOI: 10.1126/scirobotics.abc5044.

Bellman, Richard (1966). "Dynamic programming." In: *Science* 153.3731, pp. 34–37.

Bengio, Yoshua, Jérôme Louradour, Ronan Collobert, and Jason Weston (2009). "Curriculum learning." In: *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48.

Bengio, Yoshua, Dong-Hyun Lee, J. Bornschein, and Zhouhan Lin (2015). "Towards Biologically Plausible Deep Learning." In: *ArXiv* abs/1502.04156.

Beniaguev, David, Idan Segev, and Michael London (2021). "Single cortical neurons as deep artificial neural networks." In: *Neuron*. ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2021.07.002.

Bermejo, Fernando, Mercedes X. Hüg, and Ezequiel A. Di Paolo (2020). "Rediscovering Richard Held: Activity and Passivity in Perceptual Learning." In: *Frontiers in Psychology* 11, p. 844. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2020.00844.

Betz, Torsten, Tim C Kietzmann, Niklas Wilming, and Peter König (2010). "Investigating task-dependent top-down effects on overt visual attention." In: *Journal of Vision* 10.3, p. 15. ISSN: 1534-7362. DOI: 10.1167/10.3.15.

Bhalla, Mukul and Dennis R Proffitt (1999). *Visual–motor recalibration in geographical slant perception.* DOI: 10.1037/0096-1523.25.4.1076.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer. ISBN: 9780387310732.

Bishop, J. Mark and Andrew O. Martin (2014). "Contemporary Sensorimotor Theory: A Brief Introduction." In: *Contemporary Sensorimotor Theory*. Ed. by John Mark Bishop and Andrew Owen Martin. Cham: Springer International Publishing, pp. 1–22. ISBN: 978-3-319-05107-9.

Bliss, T. V.P. and T. Lømo (1973). "Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path." In: *The Journal of Physiology* 232.2, pp. 331–356. ISSN: 14697793. DOI: 10.1113/jphysiol.1973.sp010273.

Blodgett, H C (1929). "The effect of the introduction of reward upon the maze performance of rats." In: *University of California Publications in Psychology* 4, pp. 113–134.

Bornstein, Aaron M and Kenneth A Norman (2017). "Reinstated episodic context guides sampling-based decisions for reward." In: *Nature Neuroscience* 20.7, pp. 997–1003. ISSN: 1546-1726. DOI: 10.1038/nn.4573.

Botvinick, Matthew Michael (2012). "Hierarchical reinforcement learning and decision making." In: *Current Opinion in Neurobiology* 22.6. Decision making, pp. 956–962. ISSN: 0959-4388. DOI: https://doi.org/10.1016/j.conb.2012.05.008.

Botvinick, Matthew, Sam Ritter, Jane X Wang, Zeb Kurth-Nelson, Charles Blundell, and Demis Hassabis (2019). "Reinforcement Learning, Fast and Slow." In: *Trends in Cognitive Sciences* 23.5, pp. 408–422. ISSN: 1364-6613. DOI: 10.1016/j.tics.2019.02.006.

Botvinick, Matthew, Jane X. Wang, Will Dabney, Kevin J. Miller, and Zeb Kurth-Nelson (2020). "Deep Reinforcement Learning and Its Neuroscientific Implications." In: *Neuron* 107.4, pp. 603–616. ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2020.06.014.

Brendel, Wieland and Matthias Bethge (2019). "Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SkfMWhAqYQ.

Briganti, Alicia M. and Leslie B. Cohen (2011). "Examining the role of social cues in early word learning." In: *Infant Behavior and Development* 34.1, pp. 211–214. ISSN: 0163-6383. DOI: https://doi.org/10.1016/j.infbeh.2010.12.012.

Browne, Cameron B, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton (2012). "A

survey of monte carlo tree search methods." In: *IEEE Transactions on Computational Intelligence and AI in games* 4.1, pp. 1–43.

Brown, Tom B, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. (2020). "Language models are few-shot learners." In: *arXiv preprint arXiv:2005.14165*.

Bukowski, C. (1966). *The Genius of the Crowd*. 7 Flowers Press.

Burghardt, Gordon (2015). "Play in fishes, frogs and reptiles." In: *Current Biology* 25. DOI: 10.1016/j.cub.2014.10.027.

Burda, Yuri, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros (2019a). "Large-Scale Study of Curiosity-Driven Learning." In: *ICLR*.

Burda, Yuri, Harrison Edwards, Amos Storkey, and Oleg Klimov (2019b). "Exploration by random network distillation." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=H1lJJnR5Ym.

Busoniu, Lucian, Robert Babuska, and Bart De Schutter (2008). "A Comprehensive Survey of Multiagent Reinforcement Learning." In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38.2, pp. 156–172. DOI: 10.1109/TSMCC.2007.913919.

Cadena, Santiago A., George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, and Alexander S. Ecker (2019). "Deep convolutional models improve predictions of macaque V1 responses to natural images." In: *PLOS Computational Biology* 15.4. Ed. by Wolfgang Einhäuser, e1006897. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006897.

Carey, Susan and E. Bartlett (1978). "Acquiring a Single New Word." In: *Proceedings of the Stanford Child Language Conference* 15, pp. 17–29.

Caruana, Richard A. (1993). "Multitask Learning: A Knowledge-Based Source of Inductive Bias." In: *Machine Learning Proceedings 1993*. San Francisco (CA): Morgan Kaufmann, pp. 41–48. ISBN: 978-1-55860-307-3. DOI: https://doi.org/10.1016/B978-1-55860-307-3.50012-5.

Cembrowski, Mark S and Nelson Spruston (2019). "Heterogeneity within classical cell types is the rule: lessons from hippocampal pyramidal neurons." In: *Nature Reviews Neuroscience* 20.4, pp. 193–204. ISSN: 1471-0048. DOI: 10.1038/s41583-019-0125-5.

Chaslot, Guillaume, Sander Bakkes, Istvan Szita, and Pieter Spronck (2008). "Monte-Carlo Tree Search: A New Framework for Game AI." In: *AIIDE* 8, pp. 216–217.

Chang, Chun Yun, Guillem R Esber, Yasmin Marrero-Garcia, Hau-Jie Yau, Antonello Bonci, and Geoffrey Schoenbaum (2016). "Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors." In: *Nature Neuroscience* 19.1, pp. 111–116. ISSN: 1546-1726. DOI: 10.1038/nn.4191.

Chen, Lili, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mor-

datch (2021). *Decision Transformer: Reinforcement Learning via Sequence Modeling*. Tech. rep. arXiv: 2106.01345v1.

Cho, Tatsuya, Kentaro Katahira, Kazuo Okanoya, and Masato Okada (2011). "Node perturbation learning without noiseless baseline." In: *Neural Networks* 24.3, pp. 267–272. ISSN: 08936080. DOI: 10.1016/j.neunet.2010.12.001.

Clark, Andy (2017). "Embodied, Situated, and Distributed Cognition." In: *A Companion to Cognitive Science*. John Wiley & Sons, Ltd. Chap. 39, pp. 506–517. ISBN: 9781405164535. DOI: https://doi.org/10.1002/9781405164535.ch39.

Clavera, Ignasi, Anusha Nagabandi, Ronald S. Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn (2018). "Learning to Adapt: Meta-Learning for Model-Based Control." In: *CoRR* abs/1803.11347. arXiv: 1803.11347. URL: http://arxiv.org/abs/1803.11347.

Clay, Viviane, Peter König, and Sabine König (2019). "Eye tracking in virtual reality." In: *Journal of Eye Movement Research* 12.1. ISSN: 19958692. DOI: 10.16910/jemr.12.1.3.

Clay, Viviane (2020). "Data from Neural Network Training in the Obstacle Tower Environment to Investigate Embodied, Weakly Supervised Learning." In: *Mendeley Data*. DOI: 10.17632/ZDH4D5WS2Z.2.

Clay, Viviane, Johannes Schrumpf, Yannick Tessenow, Helmut Leder, Ulrich Ansorge, and Peter König (2020). "A quantitative analysis of the taxonomy of artistic styles." In: *Journal of Eye Movement Research* 13.2. DOI: 10.16910/jemr.13.2.5.

Clay, Viviane, Peter König, Kai-Uwe Kühnberger, and Gordon Pipa (2021a). "Learning sparse and meaningful representations through embodiment." In: *Neural Networks* 134, pp. 23–41. DOI: 10.1016/j.neunet.2020.11.004.

Clay, Viviane, Peter König, Gordon Pipa, and Kai-Uwe Kühnberger (2021b). "Fast Concept Mapping: The Emergence of Human Abilities in Artificial Neural Networks when Learning Embodied and Self-Supervised." In: *arXiv preprint arXiv:2102.02153*.

Clark, Andy (1997). *Being There: Putting Mind, Body, and World Together Again*.

Coelho, Paulo (1998). *The Alchemist: a Fable about Following Your Dreams*. HarperColins.

Colombo, Matteo (2014). "Deep and beautiful. The reward prediction error hypothesis of dopamine." In: *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 45, pp. 57–67. ISSN: 1369-8486. DOI: https://doi.org/10.1016/j.shpsc.2013.10.006.

Colas, Cédric, Pierre Fournier, Mohamed Chetouani, Olivier Sigaud, and Pierre-Yves Oudeyer (2019). "CURIOUS: Intrinsically Motivated Modular Multi-Goal Reinforcement Learning." In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 1331–1340. URL: https://proceedings.mlr.press/v97/colas19a.html.

Colas, Cédric, Tristan Karch, Olivier Sigaud, and Pierre-Yves Oudeyer (2020a). "Intrinsically Motivated Goal-Conditioned Reinforcement Learning: a Short Survey." In: *CoRR* abs/2012.09830. arXiv: 2012.09830. URL: https://arxiv.org/abs/2012.09830.

Colas, Cédric, Tristan Karch, Nicolas Lair, Jean-Michel Dussoux, Clément Moulin-Frier, Peter F. Dominey, and Pierre-Yves Oudeyer (2020b). "Language as a Cognitive Tool to Imagine Goals in Curiosity Driven Exploration." In: *NeurIPS*. URL: https://proceedings.neurips.cc/paper/2020/hash/274e6fcf4a583de4a81c6376f17673e7-Abstract.html.

Colombo, Matteo and Cory Wright (2021). "First principles in the life sciences: the free-energy principle, organicism, and mechanism." In: *Synthese* 198.14, pp. 3463–3488. ISSN: 1573-0964. DOI: 10.1007/s11229-018-01932-w.

Cook, Claire, Noah D. Goodman, and Laura E. Schulz (2011). "Where science starts: Spontaneous experiments in preschoolers' exploratory play." In: *Cognition* 120.3. Probabilistic models of cognitive development, pp. 341–349. ISSN: 0010-0277. DOI: https://doi.org/10.1016/j.cognition.2011.03.003.

Cox, David Daniel and Thomas Dean (2014). "Neural Networks and Neuroscience-Inspired Computer Vision." In: *Current Biology* 24.18, R921–R929. ISSN: 0960-9822. DOI: 10.1016/j.cub.2014.08.026.

Craighero, Laila and Giacomo Rizzolatti (2005). "CHAPTER 31 - The Premotor Theory of Attention." In: *Neurobiology of Attention*. Ed. by Laurent Itti, Geraint Rees, and John K. Tsotsos. Burlington: Academic Press, pp. 181–186. ISBN: 978-0-12-375731-9. DOI: https://doi.org/10.1016/B978-012375731-9/50035-5.

Crapse, Trinity B and Marc A Sommer (2008). "Corollary discharge across the animal kingdom." In: *Nature Reviews Neuroscience* 9.8, pp. 587–600. ISSN: 1471-0048. DOI: https://doi.org/10.1038/nrn2457.

Craighero, Laila, Luciano Fadiga, Giacomo Rizzolatti, and Carlo Umiltà (1999). *Action for perception: A motor-visual attentional effect*. US. DOI: 10.1037/0096-1523.25.6.1673.

Crichton, M. (2002). *Prey*. HarperCollins. ISBN: 9780066214122.

Crick, Francis (1989). "The recent excitement about neural networks." In: *Nature* 337.6203, pp. 129–132. ISSN: 00280836. DOI: 10.1038/337129a0.

Cybenko, George (1989). "Approximation by superpositions of a sigmoidal function." In: *Mathematics of control, signals and systems* 2.4, pp. 303–314.

Dapello, Joel, Tiago Marques, Martin Schrimpf, Franziska Geiger, David D. Cox, and James J. DiCarlo (2020). "Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations." In: *NeurIPS*. URL: https://proceedings.neurips.cc/paper/2020/hash/98b17f068d5d9b7668e19fb8ae470841-Abstract.html.

Dashiell, J F (1925). "A quantitative demonstration of animal drive." In: *Journal of Comparative Psychology* 5.3, pp. 205–208. ISSN: 0093-4127(Print). DOI: 10.1037/h0071833.

Daucé, Emmanuel (2018). "Active Fovea-Based Vision Through Computationally-Effective Model-Based Prediction." In: *Frontiers in Neurorobotics* 12, p. 76. ISSN: 1662-5218. DOI: 10.3389/fnbot.2018.00076.

Daw, Nathaniel D, Aaron C Courville, and David S Touretzky (2006). "Representation and Timing in Theories of the Dopamine System." In: *Neural Computation* 18.7, pp. 1637–1677. ISSN: 0899-7667. DOI: 10.1162/neco.2006.18.7.1637.

Dayan, Eran, Antonino Casile, Nava Levit-Binnun, Martin A. Giese, Talma Hendler, and Tamar Flash (2007). "Neural representations of kinematic laws of motion: Evidence for action-perception coupling." In: *Proceedings of the National Academy of Sciences* 104.51, pp. 20582–20587. ISSN: 0027-8424. DOI: 10.1073/pnas.0710033104.

Dayan, Peter and Yael Niv (2008). "Reinforcement learning: The Good, The Bad and The Ugly." In: *Current Opinion in Neurobiology* 18.2. Cognitive neuroscience, pp. 185–196. ISSN: 0959-4388. DOI: https://doi.org/10.1016/j.conb.2008.08.003.

Delange, Matthias, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars (2021). "A continual learning survey: Defying forgetting in classification tasks." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1. DOI: 10.1109/TPAMI.2021.3057446.

Diederen, Kelly M. J. and Paul C. Fletcher (2021). "Dopamine, Prediction Error and Beyond." In: *The Neuroscientist* 27.1. PMID: 32338128, pp. 30–46. DOI: 10.1177/1073858420907591.

Dijkstra, N, P Zeidman, S Ondobaka, M A J van Gerven, and K Friston (2017). "Distinct Top-down and Bottom-up Brain Connectivity During Visual Perception and Imagery." In: *Scientific Reports* 7.1, p. 5677. ISSN: 2045-2322. DOI: 10.1038/s41598-017-05888-8.

Doll, Bradley B, Dylan A Simon, and Nathaniel D Daw (2012). "The ubiquity of model-based reinforcement learning." In: *Current Opinion in Neurobiology* 22.6. Decision making, pp. 1075–1081. ISSN: 0959-4388. DOI: https://doi.org/10.1016/j.conb.2012.08.003.

Dong, Hao, Zihan Ding, and Shanghang Zhang (2020). *Deep Reinforcement Learning: Fundamentals, Research and Applications*. Springer Nature.

Du, Yilun, Chuang Gan, and Phillip Isola (2021). "Curious Representation Learning for Embodied Intelligence." In: *Arxiv*. arXiv: 2105.01060v1.

Duan, Jiafei, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan (2021). "A Survey of Embodied AI: From Simulators to Research Tasks." In: *CoRR* abs/2103.04918. arXiv: 2103.04918. URL: https://arxiv.org/abs/2103.04918.

Duchi, John, Elad Hazan, and Yoram Singer (2011). "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization." In: *Journal of Machine Learning Research* 12.61, pp. 2121–2159. URL: http://jmlr.org/papers/v12/duchi11a.html.

Engel, Andreas K., Alexander Maye, Martin Kurthen, and Peter König (2013). "Where's the action? The pragmatic turn in cognitive science." In: *Trends in Cognitive Sciences* 17, pp. 202–209.

Eshel, Neir, Ju Tian, Michael Bukwich, and Naoshige Uchida (2016). "Dopamine neurons share common response function for reward prediction error." In: *Nature Neuroscience* 19.3, pp. 479–486. ISSN: 1546-1726. DOI: 10.1038/nn.4239.

Etzel, Joset A., Valeria Gazzola, and Christian Keysers (2008). "Testing Simulation Theory with Cross-Modal Multivariate Classification of fMRI Data." In: *PLoS ONE* 3.11. Ed. by Bernhard Baune, e3690. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0003690.

Fagioli, Sabrina, Bernhard Hommel, and Ricarda Ines Schubotz (2007). "Intentional control of attention: action planning primes action-related stimulus dimensions." In: *Psychological Research* 71.1, pp. 22–29. ISSN: 1430-2772. DOI: 10.1007/s00426-005-0033-3.

Feinberg, Vladimir, Alvin Wan, Ion Stoica, Michael I. Jordan, Joseph E. Gonzalez, and Sergey Levine (2018). "Model-Based Value Estimation for Efficient Model-Free Reinforcement Learning." In: *CoRR* abs/1803.00101. arXiv: 1803.00101. URL: http://arxiv.org/abs/1803.00101.

Felleman, D J and D C Van Essen (1991). "Distributed hierarchical processing in the primate cerebral cortex." In: *Cerebral cortex (New York, N.Y. : 1991)* 1.1, pp. 1–47. ISSN: 1047-3211 (Print). DOI: 10.1093/cercor/1.1.1-a.

Fifty, Christopher, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn (2021). *Efficiently Identifying Task Groupings for Multi-Task Learning*. arXiv: 2109.04617 [cs.LG].

Finn, Chelsea, Ian Goodfellow, and Sergey Levine (2016). "Unsupervised Learning for Physical Interaction through Video Prediction." In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2016/file/d9d4f495e875a2e075a1a4a6e1b9770f-Paper.pdf.

Fiser, Aris, David Mahringer, Hassana K Oyibo, Anders V Petersen, Marcus Leinweber, and Georg B Keller (2016). "Experience-dependent spatial expectations in mouse visual cortex." In: *Nature neuroscience* 19.12, pp. 1658–1664.

Foglia, Lucia and Robert A. Wilson (2013). "Embodied cognition." In: *WIREs Cognitive Science* 4.3, pp. 319–325. DOI: https://doi.org/10.1002/wcs.1226.

Freedman, D. J., M. Riesenhuber, T. Poggio, and E. K. Miller (2001). "Categorical representation of visual stimuli in the primate prefrontal cortex." In: *Science* 291.5502, pp. 312–316. ISSN: 00368075. DOI: 10.1126/science.291.5502.312.

French, Robert M. (1999). "Catastrophic forgetting in connectionist networks." In: *Trends in Cognitive Sciences* 3.4, pp. 128–135. ISSN: 1364-6613. DOI: https://doi.org/10.1016/S1364-6613(99)01294-2.

Friston, Karl, James Kilner, and Lee Harrison (2006). "A free energy principle for the brain." In: *Journal of Physiology Paris* 100.1-3, pp. 70–87. ISSN: 09284257. DOI: 10.1016/j.jphysparis.2006.10.001.

Friston, Karl (2010). "The free-energy principle: A unified brain theory?" In: *Nature Reviews Neuroscience* 11.2, pp. 127–138. ISSN: 1471003X. DOI: 10.1038/nrn2787.

Friston, Karl, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, John OĎoherty, and Giovanni Pezzulo (2016). "Active inference and learning." In: *Neuroscience and Biobehavioral Reviews* 68, pp. 862–879. ISSN: 0149-7634. DOI: https://doi.org/10.1016/j.neubiorev.2016.06.022.

Friston, Karl J., Richard Rosch, Thomas Parr, Cathy Price, and Howard Bowman (2017). *Deep temporal models and active inference*. DOI: 10.1016/j.neubiorev.2017.04.009.

Fry, S. (2017). *Mythos: The Greek Myths Retold*. Stephen Fry's Greek Myths. Penguin Books Limited. ISBN: 9781405934169.

Fujimoto, Scott, Herke Hoof, and David Meger (2018). "Addressing function approximation error in actor-critic methods." In: *International Conference on Machine Learning*. PMLR, pp. 1587–1596.

Fukushima, Kunihiko (1980). "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position." In: *Biological Cybernetics* 36.4, pp. 193–202. ISSN: 1432-0770. DOI: 10.1007/BF00344251.

Geirhos, R., C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann (2018). "Generalisation in humans and deep neural networks." In: *Advances in Neural Information Processing Systems 31*. URL: https://arxiv.org/abs/1808.08750.

Geirhos, R., P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel (2019). "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." In: URL: https://openreview.net/forum?id=Bygh9j09KX.

Geirhos, R., K. Narayanappa, B. Mitzkus, M. Bethge, F. A. Wichmann, and W. Brendel (2020). "On the surprising similarities between supervised and self-supervised models." In: URL: https://arxiv.org/pdf/2010.08377.pdf.

Gershman, Samuel J., Bijan Pesaran, and Nathaniel D. Daw (2009). "Human Reinforcement Learning Subdivides Structured Action Spaces by Learning Effector-Specific Values." In: *Journal of Neuroscience* 29.43, pp. 13524–13531. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.2469-09.2009.

Gershman, Samuel J. and Nathaniel D. Daw (2017). "Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework." In: *Annual Review of Psychology* 68.1. PMID: 27618944, pp. 101–128. DOI: 10.1146/annurev-psych-122414-033625.

Gerven, Marcel van (2017). "Computational Foundations of Natural Intelligence." In: *Frontiers in Computational Neuroscience* 11, p. 112. ISSN: 1662-5188. DOI: 10.3389/fncom.2017.00112.

Gershman, Samuel J. (2019). *What does the free energy principle tell us about the brain?* arXiv: 1901.07945.

Gerstner, Wulfram, Andreas K. Kreiter, Henry Markram, and Andreas V. M. Herz (1997). "Neural codes: Firing rates and beyond."

In: *Proceedings of the National Academy of Sciences* 94.24, pp. 12740–12741. ISSN: 0027-8424. DOI: 10.1073/pnas.94.24.12740.

Gibson, James J (1962). "Observations on active touch." In: *Psychological Review* 69.6, pp. 477–491. ISSN: 1939-1471(Electronic),0033-295X(Print). DOI: 10.1037/h0046962.

— (1979). *The Ecological Approach to Visual Perception*.

Gilbert, Charles D. and Wu Li (2013). *Top-down influences on visual processing*. DOI: 10.1038/nrn3476.

Glimcher, Paul W. (2011). "Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis." In: *Proceedings of the National Academy of Sciences* 108.Supplement 3, pp. 15647–15654. ISSN: 0027-8424. DOI: 10.1073/pnas.1014269108.

Goldinger, Stephen D, Megan H Papesh, Anthony S Barnhart, Whitney A Hansen, and Michael C Hout (2016). "The poverty of embodied cognition." In: *Psychonomic Bulletin & Review* 23.4, pp. 959–978. ISSN: 1531-5320. DOI: 10.3758/s13423-015-0860-1.

Goldstein, Ariel et al. (2021). "Thinking ahead: spontaneous prediction in context as a keystone of language in humans and machines." In: *bioRxiv*. DOI: 10.1101/2020.12.02.403477.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Gopnik, Alison and Laura Schulz (2004). "Mechanisms of theory formation in young children." In: *Trends in cognitive sciences* 8.8, pp. 371–377. ISSN: 1364-6613 (Print). DOI: 10.1016/j.tics.2004.06.005.

Gopnik, Alison and Henry M Wellman (2012). "Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory." In: *Psychological bulletin* 138.6, pp. 1085–1108. ISSN: 1939-1455 (Electronic). DOI: 10.1037/a0028044.

Gopnik, A. (2016). *The Gardener and the Carpenter: What the New Science of Child Development Tells Us About the Relationship Between Parents and Children*. Farrar, Straus and Giroux. ISBN: 9781429944335.

Gopnik, Alison (2020). "Childhood as a solution to explore–exploit tensions." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 375.1803, p. 20190502. DOI: 10.1098/rstb.2019.0502.

Gopnik, Alison, Andrew N Meltzoff, and Patricia K Kuhl (1999). *The scientist in the crib: Minds, brains, and how children learn.* New York, NY, US: William Morrow & Co, pp. xv, 279–xv, 279. ISBN: 0-688-15988-5 (Hardcover).

Graziano, Michael S, Gregory S Yap, and Charles G Gross (1994). "Coding of visual space by premotor neurons." In: *Science* 266.5187, pp. 1054–1057.

Gremel, Christina M and Rui M Costa (2013). "Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions." In: *Nature Communications* 4.1, p. 2264. ISSN: 2041-1723. DOI: 10.1038/ncomms3264.

Griffiths, Thomas L, Frederick Callaway, Michael B Chang, Erin Grant, Paul M Krueger, and Falk Lieder (2019). "Doing more with less: meta-reasoning and meta-learning in humans and machines." In: *Current Opinion in Behavioral Sciences* 29. Artificial Intelligence,

pp. 24–30. ISSN: 2352-1546. DOI: https://doi.org/10.1016/j.cobeha.2019.01.005.

Griffiths, Thomas L. (2020). "Understanding Human Intelligence through Human Limitations." In: *Trends in Cognitive Sciences* 24.11, pp. 873–883. ISSN: 1364-6613. DOI: https://doi.org/10.1016/j.tics.2020.09.001.

Gronauer, Sven and Klaus Diepold (2021). "Multi-agent deep reinforcement learning: a survey." In: *Artificial Intelligence Review*. ISSN: 1573-7462. DOI: 10.1007/s10462-021-09996-w.

Guerguiev, Jordan, Timothy P. Lillicrap, and Blake A. Richards (2017). "Towards deep learning with segregated dendrites." In: *eLife* 6. ISSN: 2050084X. DOI: 10.7554/eLife.22901. arXiv: 1610.00161.

Guerguiev, Jordan, Konrad P. Körding, and Blake A. Richards (2019). "Spike-based causal inference for weight alignment." In: arXiv: 1910.01689. URL: http://arxiv.org/abs/1910.01689.

Gweon, Hyowon (2021). "Inferential social learning: cognitive foundations of human social learning and teaching." In: *Trends in Cognitive Sciences* 25.10, pp. 896–910. ISSN: 1364-6613. DOI: https://doi.org/10.1016/j.tics.2021.07.008.

Haarnoja, Tuomas, Aurick Zhou, Pieter Abbeel, and Sergey Levine (2018). "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor." In: *International conference on machine learning*. PMLR, pp. 1861–1870.

Hamrick, Jessica B, Peter W Battaglia, Thomas L Griffiths, and Joshua B Tenenbaum (2016). "Inferring mass in complex scenes by mental simulation." In: *Cognition* 157, pp. 61–76. DOI: 10.1016/j.cognition.2016.08.012.

Harlow, Harry F (1949). *The formation of learning sets.* US. DOI: 10.1037/h0062474.

Haslinger, Robert, Gordon Pipa, Bruss Lima, Wolf Singer, Emery N. Brown, and Sergio Neuenschwander (2012). "Context Matters: The Illusive Simplicity of Macaque V1 Receptive Fields." In: *PLOS ONE* 7.7, pp. 1–17. DOI: 10.1371/journal.pone.0039699.

Hasson, Uri, Samuel A. Nastase, and Ariel Goldstein (2020). "Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks." In: *Neuron* 105.3, pp. 416–434. ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2019.12.002.

Hawkins, Jeff and Subutai Ahmad (2016). "Why Neurons Have Thousands of Synapses, a Theory of Sequence Memory in Neocortex." In: *Frontiers in Neural Circuits* 10, p. 23. ISSN: 1662-5110. DOI: 10.3389/fncir.2016.00023.

Hawkins, Jeff, Subutai Ahmad, and Yuwei Cui (2017). "A Theory of How Columns in the Neocortex Enable Learning the Structure of the World." In: *Frontiers in Neural Circuits* 11, p. 81. ISSN: 1662-5110. DOI: 10.3389/fncir.2017.00081.

Held, Richard and Alan Hein (1963). *Movement-Produced Stimulation in the Development of Visually Guided Behavior*. Tech. rep. 5, pp. 872–876. DOI: https://doi.org/10.1037/h0040546.

Held, Richard and Jerold Rekosh (1963). "Motor-sensory feedback and the geometry of visual space." In: *Science* 141.3582, pp. 722–723.

Henderson, Peter, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger (2017). "Deep Reinforcement Learning that Matters." In: *CoRR* abs/1709.06560. arXiv: 1709.06560. URL: http://arxiv.org/abs/1709.06560.

Hendrycks, Dan, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song (2019). "Natural Adversarial Examples." In: *CoRR* abs/1907.07174. arXiv: 1907.07174. URL: http://arxiv.org/abs/1907.07174.

Hermann, Katherine L, Ting Chen, and Simon Kornblith (2019). "The Origins and Prevalence of Texture Bias in Convolutional Neural Networks." In: *arXiv preprint arXiv:1911.09071*.

Hessel, Matteo, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Daniel Horgan, Bilal Piot, Mohammad Gheshlaghi Azar, and David Silver (2017). "Rainbow: Combining Improvements in Deep Reinforcement Learning." In: *CoRR* abs/1710.02298. arXiv: 1710.02298. URL: http://arxiv.org/abs/1710.02298.

Hill, Felix, Olivier Tieleman, Tamara von Glehn, Nathaniel Wong, Hamza Merzic, and Stephen Clark (2021). "Grounded Language Learning Fast and Slow." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=wpSWuz_hyqA.

Hinton, Geoffrey E and Ruslan R Salakhutdinov (2006). "Reducing the dimensionality of data with neural networks." In: *science* 313.5786, pp. 504–507.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory." In: *Neural computation* 9.8, pp. 1735–1780.

Hole, Kjell Jørgen and Subutai Ahmad (2021). "A thousand brains: toward biologically constrained AI." In: *SN Applied Sciences* 3.8, p. 743. ISSN: 2523-3971. DOI: 10.1007/s42452-021-04715-0.

Hollerman, Jeffrey R and Wolfram Schultz (1998). "Dopamine neurons report an error in the temporal prediction of reward during learning." In: *Nature Neuroscience* 1.4, pp. 304–309. ISSN: 1546-1726. DOI: 10.1038/1124.

Hornik, Kurt (1991). "Approximation capabilities of multilayer feedforward networks." In: *Neural Networks* 4.2, pp. 251–257. ISSN: 0893-6080. DOI: https://doi.org/10.1016/0893-6080(91)90009-T.

Hsee, Christopher K. and Bowen Ruan (2016). "The Pandora Effect: The Power and Peril of Curiosity." In: *Psychological Science* 27.5. PMID: 27000178, pp. 659–666. DOI: 10.1177/0956797616631733.

Hsu, Kyle, Sergey Levine, and Chelsea Finn (2018). "Unsupervised Learning via Meta-Learning." In: *CoRR* abs/1810.02334. URL: http://arxiv.org/abs/1810.02334.

Huang, Sandy H., Nicolas Papernot, Ian J. Goodfellow, Yan Duan, and Pieter Abbeel (2017). "Adversarial Attacks on Neural Net-

work Policies." In: *CoRR* abs/1702.02284. arXiv: 1702.02284. URL: http://arxiv.org/abs/1702.02284.

Hubel, D. H. and T. N. Wiesel (1959). "Receptive fields of single neurones in the cat's striate cortex." In: *The Journal of Physiology* 148.3, pp. 574–591. ISSN: 00223751. DOI: 10.1113/jphysiol.1959.sp006308.

— (1968). "Receptive fields and functional architecture of monkey striate cortex." In: *The Journal of Physiology* 195.1, pp. 215–243. ISSN: 00223751. DOI: 10.1113/jphysiol.1968.sp008455.

Hupé, J M, A C James, B R Payne, S G Lomber, P Girard, and J Bullier (1998). "Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons." In: *Nature* 394.6695, pp. 784–787. ISSN: 0028-0836 (Print). DOI: 10.1038/29537.

Hussein, Ahmed, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne (2017). "Imitation Learning: A Survey of Learning Methods." In: *ACM Comput. Surv.* 50.2. ISSN: 0360-0300. DOI: 10.1145/3054912.

Ibarz, Julian, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine (2021). "How to train your robot with deep reinforcement learning: lessons we have learned." In: *The International Journal of Robotics Research* 40.4-5, pp. 698–721. DOI: 10.1177/0278364920987859.

Iida, Fumiya, Rolf Pfeifer, Luc Steels, and Yasuo Kuniyoshi (2004). *Embodied Artificial Intelligence: International Seminar, Dagstuhl Castle, Germany, July 7-11, 2003, Revised Selected Papers*. Ed. by Fumiya Iida, Rolf Pfeifer, Luc Steels, and Yasuo Kuniyoshi. Vol. 3139. Springer. DOI: 10.1007/B99075.

Irpan, Alex (2018). *Deep Reinforcement Learning Doesn't Work Yet*. https://www.alexirpan.com/2018/02/14/rl-hard.html.

Jabri, M. and B. Flower (1992). "Weight perturbation: an optimal architecture and learning technique for analog VLSI feedforward and recurrent multilayer networks." In: *IEEE Transactions on Neural Networks* 3.1, pp. 154–157. DOI: 10.1109/72.105429.

Jafari, Matiar, Tyson Aflalo, Srinivas Chivukula, Spencer Sterling Kellis, Michelle Armenta Salas, Sumner Lee Norman, Kelsie Pejsa, Charles Yu Liu, and Richard Alan Andersen (2020). "The human primary somatosensory cortex encodes imagined movement in the absence of sensory information." In: *Communications Biology* 3.1, p. 757. ISSN: 2399-3642. DOI: 10.1038/s42003-020-01484-1.

Jaramillo-Avila, Uziel and Sean R. Anderson (2019). "Foveated image processing for faster object detection and recognition in embedded systems using deep convolutional neural networks." In: *CoRR* abs/1908.09000. arXiv: 1908.09000. URL: http://arxiv.org/abs/1908.09000.

Johnson, Adam and A. David Redish (2007). "Neural Ensembles in CA3 Transiently Encode Paths Forward of the Animal at a Decision Point." In: *Journal of Neuroscience* 27.45, pp. 12176–12189. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.3761-07.2007.

Johnson, Adam, Matthijs AA van der Meer, and A David Redish (2007). "Integrating hippocampus and striatum in decision-making."

In: *Current Opinion in Neurobiology* 17.6, pp. 692–697. ISSN: 0959-4388. DOI: https://doi.org/10.1016/j.conb.2008.01.003.

Jurassic Parc (1993).

Khaligh-Razavi, Seyed-Mahdi and Nikolaus Kriegeskorte (2014). "Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation." In: *PLoS Computational Biology* 10.11. Ed. by Jörn Diedrichsen, e1003915. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003915.

Kaminski, Juliane, J. Call, and J. Fischer (2004). "Word Learning in a Domestic Dog: Evidence for "Fast Mapping"." In: *Science* 304, pp. 1682 –1683.

Kaspar, Kai, Sabine Koenig, Jessika Schwandt, and Peter König (2014). "The experience of new sensorimotor contingencies by sensory augmentation." In: *Consciousness and Cognition* 28, 47–63. DOI: 10.1016/j.concog.2014.06.006.

Kaushik, Prakhar, Alex Gain, Adam Kortylewski, and Alan L. Yuille (2021). "Understanding Catastrophic Forgetting and Remembering in Continual Learning with Optimal Relevance Mapping." In: *CoRR* abs/2102.11343. arXiv: 2102.11343. URL: https://arxiv.org/abs/2102.11343.

Kawato, Mitsuo, Tomoe Kuroda, Hiroshi Imamizu, Eri Nakano, Satoru Miyauchi, and Toshinori Yoshioka (2003). "Internal forward models in the cerebellum: fMRI study on grip force and load force coupling." In: *Neural Control of Space Coding and Action Production*. Vol. 142. Progress in Brain Research. Elsevier, pp. 171–188. DOI: https://doi.org/10.1016/S0079-6123(03)42013-X.

Kawato, Mitsuo and Hiroaki Gomi (1992). "A computational model of four regions of the cerebellum based on feedback-error learning." In: *Biological Cybernetics* 68.2, pp. 95–103. ISSN: 1432-0770. DOI: 10.1007/BF00201431.

Kazantzakis, Nikos (1996). Simon and Schuster.

Kell, Alexander J.E., Daniel L.K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, and Josh H. McDermott (2018). "A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy." In: *Neuron* 98.3, 630–644.e16. ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2018.03.044.

Khadka, Shauharda and Kagan Tumer (2018). "Evolution-guided policy gradient in reinforcement learning." In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 1196–1208.

Kietzmann, Tim C, Courtney J Spoerer, Lynn K A Sörensen, Radoslaw M Cichy, Olaf Hauk, and Nikolaus Kriegeskorte (2019). "Recurrence is required to capture the representational dynamics of the human visual system." In: *Proceedings of the National Academy of Sciences of the United States of America* 116.43, pp. 21854–21863. ISSN: 1091-6490 (Electronic). DOI: 10.1073/pnas.1905544116.

Kieliba, Paulina, Danielle Clode, Roni O. Maimon-Mor, and Tamar R. Makin (2021). "Robotic hand augmentation drives changes in

neural body representation." In: *Science Robotics* 6.54. DOI: 10.1126/scirobotics.abd7935.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization." In: *arXiv preprint arXiv:1412.6980*.

Kirkpatrick, James et al. (2017). "Overcoming catastrophic forgetting in neural networks." In: *Proceedings of the National Academy of Sciences* 114.13, pp. 3521–3526. ISSN: 0027-8424. DOI: 10.1073/pnas.1611835114.

Kitazawa, Shigeru, Tatsuya Kimura, and Ping-Bo Yin (1998). "Cerebellar complex spikes encode both destinations and errors in arm movements." In: *Nature* 392.6675, pp. 494–497. ISSN: 1476-4687. DOI: 10.1038/33141.

Kloppenburg, Peter and Martin Paul Nawrot (2014). "Neural Coding: Sparse but On Time." In: *Current Biology* 24.19, R957–R959. ISSN: 0960-9822. DOI: https://doi.org/10.1016/j.cub.2014.08.041.

Koch, Christof and Gilles Laurent (1999). "Complexity and the Nervous System." In: *Science* 284, pp. 96–98. DOI: 10.1126/science.284.5411.96.

König, Sabine U. et al. (2016). "Learning New Sensorimotor Contingencies: Effects of Long-Term Use of Sensory Augmentation on the Brain and Conscious Perception." In: *PLOS ONE* 11.12. Ed. by Maurice Ptito, e0166647. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0166647.

König, Sabine U, Viviane Clay, Debora Nolte, Laura Duesberg, Nicolas Kuske, and Peter König (2019). "Learning of spatial properties of a large-scale virtual city with an interactive map." In: *Frontiers in human neuroscience* 13, p. 240. DOI: 10.3389/fnhum.2019.00240.

König, Sabine U, Caspar Goeke, Tobias Meilinger, and Peter König (2019). "Are allocentric spatial reference frames compatible with theories of Enactivism?" In: *Psychological Research* 83.3, pp. 498–513. ISSN: 1430-2772. DOI: 10.1007/s00426-017-0899-x.

Kosoy, Eliza, Jasmine Collins, David M. Chan, Jessica B. Hamrick, Sandy Huang, Alison Gopnik, and John F. Canny (2020). "Exploring Exploration: Comparing Children with RL Agents in Unified Environments." In: *CoRR* abs/2005.02880. URL: https://arxiv.org/abs/2005.02880.

Kretch, Kari S. and Karen E. Adolph (2013). "Cliff or Step? Posture-Specific Learning at the Edge of a Drop-Off." In: *Child Development* 84.1, pp. 226–240. DOI: https://doi.org/10.1111/j.1467-8624.2012.01842.x.

Kriegeskorte, Nikolaus (2015). "Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing." In: *Annual Review of Vision Science* 1.1. PMID: 28532370, pp. 417–446. DOI: 10.1146/annurev-vision-082114-035447.

Kurakin, Alexey, Ian J. Goodfellow, and Samy Bengio (2016). "Adversarial examples in the physical world." In: *CoRR* abs/1607.02533. arXiv: 1607.02533. URL: http://arxiv.org/abs/1607.02533.

Kurtz, Mark, Justin Kopinsky, Rati Gelashvili, Alexander Matveev, John Carr, Michael Goin, William Leiserson, Sage Moore, Nir Shavit, and Dan Alistarh (2020). "Inducing and Exploiting Acti-

vation Sparsity for Fast Inference on Deep Neural Networks." In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 5533–5543. URL: https://proceedings.mlr.press/v119/kurtz20a.html.

König, Sabine U., Ashima Keshava, Viviane Clay, Kirsten Rittershofer, Nicolas Kuske, and Peter König (2021). "Embodied Spatial Knowledge Acquisition in Immersive Virtual Reality: Comparison to Map Exploration." In: *Frontiers in Virtual Reality* 2, p. 4. ISSN: 2673-4192. DOI: 10.3389/frvir.2021.625548.

Lake, Brenden M, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman (2017). "Building machines that learn and think like people." In: *Behavioral and Brain Sciences* 40, e253. DOI: 10.1017/S0140525X16001837.

Lamme, Victor A F and Pieter R Roelfsema (2000). "The distinct modes of vision offered by feedforward and recurrent processing." In: *Trends in Neurosciences* 23.11, pp. 571–579. ISSN: 0166-2236. DOI: 10.1016/S0166-2236(00)01657-X.

Lammel, Stephan, Byung Kook Lim, and Robert C. Malenka (2014). "Reward and aversion in a heterogeneous midbrain dopamine system." In: *Neuropharmacology* 76. NIDA 40th Anniversary Issue, pp. 351–359. ISSN: 0028-3908. DOI: https://doi.org/10.1016/j.neuropharm.2013.03.019.

Lampinen, Andrew K, Stephanie C Y Chan, Andrea Banino, and Felix Hill (2021). "Towards mental time travel: a hierarchical memory for reinforcement learning agents." In: *Arxiv*. arXiv: 2105.14039v1.

Lansdell, Benjamin James and Konrad Paul Körding (2019). "Neural spiking for causal inference." In: *bioRxiv*. DOI: 10.1101/253351.

Landau, Barbara, Linda B Smith, and Susan S Jones (1988). "The importance of shape in early lexical learning." In: *Cognitive Development* 3.3, pp. 299–321. ISSN: 1879-226X(Electronic),0885-2014(Print). DOI: 10.1016/0885-2014(88)90014-7.

Lawrence of Arabia (1962).

Lecun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). *Deep learning*. DOI: 10.1038/nature14539.

Lengler, Johannes, Florian Jug, and Angelika Steger (2013). "Reliable Neuronal Systems: The Importance of Heterogeneity." In: *PLOS ONE* 8.12, pp. 1–10. DOI: 10.1371/journal.pone.0080694.

Leshno, Moshe, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken (1993). "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function." In: *Neural Networks* 6.6, pp. 861–867. ISSN: 0893-6080. DOI: https://doi.org/10.1016/S0893-6080(05)80131-5.

Leugering, Johannes, Pascal Nieters, and Gordon Pipa (2020). "Event-based pattern detection in active dendrites." In: *bioRxiv*. DOI: 10.1101/690792.

— (2021). "A Minimal Model of Neural Computation with Dendritic Plateau Potentials." In: *bioRxiv*. DOI: 10.1101/690792.

Lillicrap, T., Jonathan J. Hunt, A. Pritzel, N. Heess, T. Erez, Yuval Tassa, D. Silver, and Daan Wierstra (2016a). "Continuous control with deep reinforcement learning." In: *CoRR* abs/1509.02971.

Lillicrap, Timothy P, Daniel Cownden, Douglas B Tweed, and Colin J Akerman (2016b). "Random synaptic feedback weights support error backpropagation for deep learning." In: *Nature communications* 7.1, pp. 1–10.

Lillicrap, Timothy P., Adam Santoro, Luke Marris, Colin J. Akerman, and Geoffrey Hinton (2020). "Backpropagation and the brain." In: *Nature Reviews Neuroscience* 21.6, pp. 335–346. ISSN: 14710048. DOI: 10.1038/s41583-020-0277-3.

Liu, Belle, Arthur Hong, Fred Rieke, and Michael B Manookin (2021). "Predictive encoding of motion begins in the primate retina." In: *Nature Neuroscience* 24.9, pp. 1280–1291. ISSN: 1546-1726. DOI: 10.1038/s41593-021-00899-1.

Ljungberg, T., P. Apicella, and W. Schultz (1992). "Responses of monkey dopamine neurons during learning of behavioral reactions." In: *Journal of Neurophysiology* 67.1. PMID: 1552316, pp. 145–163. DOI: 10.1152/jn.1992.67.1.145.

Lobo, Lorena, Manuel Heras-Escribano, and David Travieso (2018). "The History and Philosophy of Ecological Psychology." In: *Frontiers in Psychology* 9, p. 2228. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2018.02228.

London, Michael and Michael Häusser (2005). "DENDRITIC COMPUTATION." In: *Annual Review of Neuroscience* 28.1. PMID: 16033324, pp. 503–532. DOI: 10.1146/annurev.neuro.28.061604.135703.

Lotter, William, Gabriel Kreiman, and David D. Cox (2016). "Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning." In: *CoRR* abs/1605.08104. URL: http://arxiv.org/abs/1605.08104.

Lotter, William, Gabriel Kreiman, and David Cox (2020). "A neural network trained for prediction mimics diverse features of biological neurons and perception." In: *Nature Machine Intelligence* 2.4, pp. 210–219. DOI: 10.1038/s42256-020-0170-9.

Ludvig, Elliot A, Richard S Sutton, and E James Kehoe (2008). "Stimulus Representation and the Timing of Reward-Prediction Errors in Models of the Dopamine System." In: *Neural Computation* 20.12, pp. 3034–3054. ISSN: 0899-7667. DOI: 10.1162/neco.2008.11-07-654.

Major, Guy, Matthew E. Larkum, and Jackie Schiller (2013). "Active Properties of Neocortical Pyramidal Neuron Dendrites." In: *Annual Review of Neuroscience* 36.1. PMID: 23841837, pp. 1–24. DOI: 10.1146/annurev-neuro-062111-150343.

Mao, Chengzhi, Amogh Gupta, Vikram Nitin, Baishakhi Ray, Shuran Song, Junfeng Yang, and Carl Vondrick (2020). "Multitask Learning Strengthens Adversarial Robustness." In: *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*. Vol. 12347. Lecture Notes in Computer Science. Springer, pp. 158–174. DOI: 10.1007/978-3-030-58536-5_10.

Marr, David (2010). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. The MIT Press. ISBN: 9780262289610. DOI: 10.7551/mitpress/9780262514620.001.0001.

Marblestone, Adam H., Greg Wayne, and Konrad P. Körding (2016). "Toward an Integration of Deep Learning and Neuroscience." In: *Frontiers in Computational Neuroscience* 10, p. 94. ISSN: 1662-5188. DOI: 10.3389/fncom.2016.00094.

Marcus, Gary (2018). "Deep Learning: A Critical Appraisal." In: *CoRR* abs/1801.00631. arXiv: 1801.00631. URL: http://arxiv.org/abs/1801.00631.

Marcus, Gary F. (1993). "Negative evidence in language acquisition." In: *Cognition* 46.1, pp. 53–85. ISSN: 0010-0277.

Markram, Henry, Joachim Lübke, Michael Frotscher, and Bert Sakmann (1997). "Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs." In: *Science* 275.5297, pp. 213–215. ISSN: 00368075. DOI: 10.1126/science.275.5297.213.

Marcus, G. F., S. Vijayan, S. Bandi Rao, and P. M. Vishton (1999). "Rule Learning by Seven-Month-Old Infants." In: *Science* 283.5398, pp. 77–80. DOI: 10.1126/science.283.5398.77.

Masse, Nicolas Y., Gregory D. Grant, and David J. Freedman (2018). "Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization." In: *Proceedings of the National Academy of Sciences* 115.44, E10467–E10475. ISSN: 0027-8424. DOI: 10.1073/pnas.1803839115.

Maye, Alexander and Andreas K. Engel (2011). "A discrete computational model of sensorimotor contingencies for object perception and control of behavior." In: *2011 IEEE International Conference on Robotics and Automation*, pp. 3810–3815. DOI: 10.1109/ICRA.2011.5979919.

Mazer, James A and Jack L Gallant (2003). "Goal-Related Activity in V4 during Free Viewing Visual Search: Evidence for a Ventral Stream Visual Salience Map." In: *Neuron* 40.6, pp. 1241–1250. ISSN: 0896-6273. DOI: https://doi.org/10.1016/S0896-6273(03)00764-5.

McDannald, Michael A., Federica Lucantonio, Kathryn A. Burke, Yael Niv, and Geoffrey Schoenbaum (2011). "Ventral Striatum and Orbitofrontal Cortex Are Both Required for Model-Based, But Not Model-Free, Reinforcement Learning." In: *Journal of Neuroscience* 31.7, pp. 2700–2705. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.5499-10.2011.

McDannald, Michael A., Yuji K. Takahashi, Nina Lopatina, Brad W. Pietras, Josh L. Jones, and Geoffrey Schoenbaum (2012). "Model-based learning and the contribution of the orbitofrontal cortex to the model-free world." In: *European Journal of Neuroscience* 35.7, pp. 991–996. DOI: https://doi.org/10.1111/j.1460-9568.2011.07982.x.

Meder, Björn, Charley M. Wu, Eric Schulz, and Azzurra Ruggeri (2021). "Development of directed and random exploration in chil-

dren." In: *Developmental Science* 24.4, e13095. DOI: https://doi.org/10.1111/desc.13095.

Meer, Matthijs van der, Zeb Kurth-Nelson, and A. David Redish (2012). "Information Processing in Decision-Making Systems." In: *The Neuroscientist* 18.4. PMID: 22492194, pp. 342–359. DOI: 10.1177/1073858411435128.

Mehrer, Johannes, Courtney J. Spoerer, Emer C. Jones, Nikolaus Kriegeskorte, and Tim C. Kietzmann (2021). "An ecologically motivated image dataset for deep learning yields better models of human vision." In: *Proceedings of the National Academy of Sciences* 118.8. ISSN: 0027-8424. DOI: 10.1073/pnas.2011417118.

Melchner, Laurie von, Sarah L Pallas, and Mriganka Sur (2000). "Visual behaviour mediated by retinal projections directed to the auditory pathway." In: *Nature* 404.6780, pp. 871–876. ISSN: 1476-4687. DOI: 10.1038/35009102.

Miall, R. C., D. J. Weir, D. M. Wolpert, and J. F. Stein (1993). "Is the Cerebellum a Smith Predictor?" In: *Journal of Motor Behavior* 25.3. PMID: 12581990, pp. 203–216. DOI: 10.1080/00222895.1993.9942050.

Michaelis, C., M. Bethge, and A. S. Ecker (2020). "Closing the Generalization Gap in One-Shot Object Detection." In: *ArXiv*. URL: https://arxiv.org/abs/2011.04267.

Mikaelian, Harutune and Richard Held (1964). "Two types of adaptation to an optically-rotated visual field." In: *The American Journal of Psychology* 77.2, pp. 257–263.

Mitchell, Tom M (1980). *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research ...

Mnih, Volodymyr et al. (2015). "Human-level control through deep reinforcement learning." In: *Nature* 518.7540, pp. 529–533. ISSN: 1476-4687. DOI: 10.1038/nature14236.

Mnih, Volodymyr, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu (2016). "Asynchronous Methods for Deep Reinforcement Learning." In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, pp. 1928–1937. URL: http://proceedings.mlr.press/v48/mniha16.html.

Montague, PR, P Dayan, and TJ Sejnowski (1996). "A framework for mesencephalic dopamine systems based on predictive Hebbian learning." In: *Journal of Neuroscience* 16.5, pp. 1936–1947. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.16-05-01936.1996.

Morris, Adam and Fiery Cushman (2019). "Model-Free RL or Action Sequences?" In: *Frontiers in Psychology* 10, p. 2892. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2019.02892.

Moran, J and R Desimone (1985). "Selective attention gates visual processing in the extrastriate cortex." In: *Science* 229.4715, pp. 782–784. ISSN: 0036-8075. DOI: 10.1126/science.4023713.

Moravec, Hans (1988). *Mind children: The future of robot and human intelligence*. Harvard University Press.

Mountcastle, V B (1997). "The columnar organization of the neocortex." In: *Brain* 120.4, pp. 701–722. ISSN: 0006-8950. DOI: 10.1093/brain/120.4.701.

Murphy, Robin, Esther Mondragón, and Victoria Murphy (2008). "Rule Learning by Rats." In: *Science (New York, N.Y.)* 319, pp. 1849–1851. DOI: 10.1126/science.1151564.

Nagel, Saskia K, Christine Carl, Tobias Kringe, Robert Märtin, and Peter König (2005). "Beyond sensory substitution—learning the sixth sense." In: *Journal of neural engineering* 2.4, R13.

Narvekar, Sanmit, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone (2020). "Curriculum learning for reinforcement learning domains: A framework and survey." In: *arXiv preprint arXiv:2003.04960*.

Nayebi, Aran, Javier Sagastuy-Brena, Daniel M. Bear, Kohitij Kar, Jonas Kubilius, Surya Ganguli, David Sussillo, James J. DiCarlo, and Daniel L. K. Yamins (2021). "Goal-Driven Recurrent Neural Network Models of the Ventral Visual Stream." In: *bioRxiv*. DOI: 10.1101/2021.02.17.431717.

Ng, Andrew Y, Stuart J Russell, et al. (2000). "Algorithms for inverse reinforcement learning." In: *Icml*. Vol. 1, p. 2.

Ng, Andrew Y, Daishi Harada, and Stuart Russell (1999). "Policy invariance under reward transformations: Theory and application to reward shaping." In: *Icml*. Vol. 99, pp. 278–287.

Niell, Cristopher M. and Michael P. Stryker (2010). "Modulation of Visual Responses by Behavioral State in Mouse Visual Cortex." In: *Neuron* 65.4, pp. 472–479. ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2010.01.033.

Nieves, Nicolas Perez and Dan F. M. Goodman (2021). "Sparse Spiking Gradient Descent." In: *CoRR* abs/2105.08810. arXiv: 2105.08810. URL: https://arxiv.org/abs/2105.08810.

Nissen, Henry W. (1930). "A Study of Exploratory Behavior in the White Rat by Means of the Obstruction Method." In: *The Pedagogical Seminary and Journal of Genetic Psychology* 37.3, pp. 361–376. DOI: 10.1080/08856559.1930.9944162.

Numenta (2021). *Sparsity Enables 100x Performance Acceleration in Deep Learning Networks A Technology Demonstration. Numenta Whitepaper.* Tech. rep. URL: https://numenta.com/.

O'Doherty, John P, Peter Dayan, Karl Friston, Hugo Critchley, and Raymond J Dolan (2003). "Temporal difference models and reward-related learning in the human brain." In: *Neuron* 38.2, pp. 329–37. ISSN: 0896-6273. DOI: 10.1016/s0896-6273(03)00169-7.

O'Doherty, John, Peter Dayan, Johannes Schultz, Ralf Deichmann, Karl Friston, and Raymond J. Dolan (2004). "Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning." In: *Science* 304.5669, pp. 452–454. ISSN: 0036-8075. DOI: 10.1126/science.1094285. URL: https://science.sciencemag.org/content/304/5669/452.

O'Mahony, N., Sean Campbell, Anderson Carvalho, L. Krpalkova, Gustavo Velasco Hernandez, Suman Harapanahalli, D. Riordan, and J. Walsh (2019). "One-Shot Learning for Custom Identification Tasks; A Review." In: *Procedia Manufacturing* 38. 29th International Conference on Flexible Automation and Intelligent Manufacturing (FAIM 2019), June 24-28, 2019, Limerick, Ireland, Beyond Industry 4.0: Industrial Advances, Engineering Education and Intelligent Manufacturing, pp. 186–193. ISSN: 2351-9789. DOI: https://doi.org/10.1016/j.promfg.2020.01.025.

O'Neill, Joseph, Barty Pleydell-Bouverie, David Dupret, and Jozsef Csicsvari (2010). "Play it again: reactivation of waking experience and memory." In: *Trends in Neurosciences* 33.5, pp. 220–229. ISSN: 0166-2236. DOI: https://doi.org/10.1016/j.tins.2010.01.006.

O'Regan, J. Kevin and Alva Noë (2001). "A sensorimotor account of vision and visual consciousness." In: *Behavioral and Brain Sciences* 24.5, 939–973. DOI: 10.1017/S0140525X01000115.

O'Regan, J. Kevin and Alva Noë (2001). "What it is like to see: A sensorimotor theory of perceptual experience." In: *Synthese* 129.1, pp. 79–103. ISSN: 1573-0964. DOI: 10.1023/A:1012699224677.

Ólafsdóttir, H Freyja, Caswell Barry, Aman B Saleem, Demis Hassabis, and Hugo J Spiers (2015). "Hippocampal place cells construct reward related sequences through unexplored space." In: *eLife* 4, e06063. ISSN: 2050-084X (Electronic). DOI: 10.7554/eLife.06063.

Olds, James and Peter Milner (1954). "Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain." In: *Journal of Comparative and Physiological Psychology* 47.6, pp. 419–427. DOI: 10.1037/h0058775.

Olshausen, Bruno A and David J Field (2004). "Sparse coding of sensory inputs." In: *Current Opinion in Neurobiology* 14.4, pp. 481–487. ISSN: 0959-4388. DOI: https://doi.org/10.1016/j.conb.2004.07.007.

— (2005). "How Close Are We to Understanding V1?" In: *Neural Computation* 17.8, pp. 1665–1699. ISSN: 0899-7667. DOI: 10.1162/0899766054026639.

Olshausen, B A and D J Field (1996). "Natural image statistics and efficient coding." In: *Network: Computation in Neural Systems* 7.2. PMID: 16754394, pp. 333–339. DOI: 10.1088/0954-898X_7_2_014.

Oord, Aäron van den, Yazhe Li, and Oriol Vinyals (2018). "Representation Learning with Contrastive Predictive Coding." In: *CoRR* abs/1807.03748. URL: http://arxiv.org/abs/1807.03748.

Ostry, David J., Mohammad Darainy, Andrew A. G. Mattar, Jeremy Wong, and Paul L. Gribble (2010). "Somatosensory Plasticity and Motor Learning." In: *Journal of Neuroscience* 30.15, pp. 5384–5393. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.4571-09.2010.

Perez-Nieves, Nicolas, Vincent Leung, Pier Dragotti, and Dan Goodman (2020). "Neural heterogeneity promotes robust learning." In: DOI: 10.1101/2020.12.18.423468.

Paolo, Ezequiel Di, Evan Thompson, and Randall D. Beer (2021). "Laying down a forking path: Incompatibilities between enaction and

the free energy principle." In: *PsyArXiv*. DOI: 10.31234/OSF.IO/D9V8F.

Parisi, German I., Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter (2019). "Continual lifelong learning with neural networks: A review." In: *Neural Networks* 113, pp. 54–71. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2019.01.012.

Pathak, Deepak, Pulkit Agrawal, Alexei Efros, and Trevor Darrell (2017). "Curiosity-Driven Exploration by Self-Supervised Prediction." In: pp. 488–489. DOI: 10.1109/CVPRW.2017.70.

Pathak, Deepak, Dhiraj Gandhi, and Abhinav Gupta (2019). "Self-Supervised Exploration via Disagreement." In: *ICML*.

Pavlov, I P (1927). *Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex.* Oxford, England: Oxford Univ. Press, pp. xv, 430–xv, 430.

Pfeiffer, Brad E and David J Foster (2013). "Hippocampal place-cell sequences depict future paths to remembered goals." In: *Nature* 497.7447, pp. 74–79. ISSN: 1476-4687. DOI: 10.1038/nature12112.

Pfeiffer, Michael and Thomas Pfeil (2018). "Deep Learning With Spiking Neurons: Opportunities and Challenges." In: *Frontiers in Neuroscience* 12, p. 774. ISSN: 1662-453X. DOI: 10.3389/fnins.2018.00774.

Piaget, J., H. Aebli, and B. Seiler (1992). *Das Erwachen der Intelligenz beim Kinde*. Klett-Cotta. ISBN: 3-423-15098-X.

Poirazi, Panayiota, Terrence Brannon, and Bartlett W. Mel (2003). "Pyramidal Neuron as Two-Layer Neural Network." In: *Neuron* 37.6, pp. 989–999. ISSN: 0896-6273. DOI: https://doi.org/10.1016/S0896-6273(03)00149-1.

Portelas, Rémy, Cédric Colas, Lilian Weng, Katja Hofmann, and Pierre-Yves Oudeyer (2020). "Automatic curriculum learning for deep rl: A short survey." In: *arXiv preprint arXiv:2003.04664*.

Pospisil, Dean A., Anitha Pasupathy, and Wyeth Bair (2018). "'Artiphysiology' reveals V4-like shape tuning in a deep network trained for image classification." In: *eLife* 7. ISSN: 2050084X. DOI: 10.7554/eLife.38242.

Pressfield, S. (2005). *The Virtues of War: A Novel of Alexander the Great*. Random House Publishing Group. ISBN: 9780553902006.

Prinz, W. (1990). "A Common Coding Approach to Perception and Action." In: *Relationships Between Perception and Action: Current Approaches*. Ed. by Odmar Neumann and Wolfgang Prinz. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 167–201. ISBN: 978-3-642-75348-0. DOI: 10.1007/978-3-642-75348-0_7.

Proffitt, Dennis R., Jeanine Stefanucci, Tom Banton, and William Epstein (2003). "The Role of Effort in Perceiving Distance." In: *Psychological Science* 14.2. PMID: 12661670, pp. 106–112. DOI: 10.1111/1467-9280.t01-1-01427.

Proclus (1992). *Proclus: A Commentary on the First Book of Euclid's Elements*. Trans. by G.R. Morrow. Princeton University Press, p. 57. ISBN: 9780691214672.

Pulvermüller, Friedemann and Luciano Fadiga (2010). "Active perception: sensorimotor circuits as a cortical basis for language." In: *Nature Reviews Neuroscience* 11.5, pp. 351–360. ISSN: 1471-0048. DOI: 10.1038/nrn2811. URL: https://doi.org/10.1038/nrn2811.

Pulvermüller, Friedemann, Rosario Tomasello, Malte R Henningsen-Schomers, and Thomas Wennekers (2021). "Biological constraints on neural network models of cognitive function." In: *Nature Reviews Neuroscience* 22.8, pp. 488–502. ISSN: 1471-0048. DOI: 10.1038/s41583-021-00473-5.

Qiu, Shilin, Qihe Liu, Shijie Zhou, and Chunjiang Wu (2019). "Review of Artificial Intelligence Adversarial Attack and Defense Technologies." In: *Applied Sciences* 9, p. 909. DOI: 10.3390/app9050909.

Quiroga, R Quian, G Kreiman, C Koch, and I Fried (2008). "Sparse but not Grandmother-cell coding in the medial temporal lobe." In: *Trends in Cognitive Sciences* 12.3, pp. 87–91. ISSN: 1364-6613. DOI: 10.1016/j.tics.2007.12.003.

Ramachandran, Deepak and Eyal Amir (2007). "Bayesian Inverse Reinforcement Learning." In: *IJCAI*. Vol. 7, pp. 2586–2591.

Rao, Rajesh P N and Dana H Ballard (1999). "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects." In: *Nature Neuroscience* 2.1, pp. 79–87. ISSN: 1546-1726. DOI: 10.1038/4580.

Rescorla, R and Allan Wagner (1972). "A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement." In: *Classical Conditioning: Current Research and Theory*.

Richards, Blake A and Timothy P Lillicrap (2019). "Dendritic solutions to the credit assignment problem." In: *Current Opinion in Neurobiology* 54. Neurobiology of Learning and Plasticity, pp. 28–36. ISSN: 0959-4388. DOI: https://doi.org/10.1016/j.conb.2018.08.003.

Richards, Blake A et al. (2019). "A deep learning framework for neuroscience." In: *Nature Neuroscience* 22.11, pp. 1761–1770. ISSN: 1546-1726. DOI: 10.1038/s41593-019-0520-2.

Ritter, S, JX Wang, Z Kurth-Nelson, and M Botvinick (2018). "Episodic Control as Meta-Reinforcement Learning." In: *bioRxiv*. DOI: 10.1101/360537.

Rizzolatti, Giacomo, Lucia Riggio, Isabella Dascola, and Carlo Umiltá (1987). "Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention." In: *Neuropsychologia* 25.1, Part 1, pp. 31–40. ISSN: 0028-3932. DOI: https://doi.org/10.1016/0028-3932(87)90041-8.

Roberts, Katherine L and Glyn W Humphreys (2011). "Action relations facilitate the identification of briefly-presented objects." In: *Attention, Perception, and Psychophysics* 73.2, pp. 597–612. ISSN: 1943-393X. DOI: 10.3758/s13414-010-0043-0.

Roelfsema, Pieter R and Arjen van Ooyen (2005). "Attention-gated reinforcement learning of internal representations for classification." In: *Neural computation* 17.10, pp. 2176–2214.

Romo, R. and W. Schultz (1990). "Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during

self-initiated arm movements." In: *Journal of Neurophysiology* 63.3. PMID: 2329363, pp. 592–606. DOI: 10.1152/jn.1990.63.3.592.

Ross, Stéphane, Geoffrey Gordon, and Drew Bagnell (2011). "A reduction of imitation learning and structured prediction to no-regret online learning." In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, pp. 627–635.

Rosenblatt, F. (1958). "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65 6, pp. 386–408.

Rothkopf, Constantin and Dana Ballard (2010). "Credit Assignment in Multiple Goal Embodied Visuomotor Behavior." In: *Frontiers in Psychology* 1, p. 173. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2010.00173.

Rouder, Jeffrey N. and Roger Ratcliff (2006). "Comparing Exemplar- and Rule-Based Theories of Categorization." In: *Current Directions in Psychological Science* 15.1, pp. 9–13. DOI: 10.1111/j.0963-7214.2006.00397.x.

Rowland, B., A. Maida, and Istvan S. N. Berkeley (2006). "Synaptic noise as a means of implementing weight-perturbation learning." In: *Connection Science* 18, pp. 69 –79.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986). "Learning representations by back-propagating errors." In: *Nature* 323.6088, pp. 533–536. ISSN: 1476-4687. DOI: 10.1038/323533a0.

Rupert, Rob (2019). *Embodied cognition*. DOI: 10.4324/9780415249126-V048-1.

Sacramento, João, Rui Ponte Costa, Yoshua Bengio, and Walter Senn (2018). "Dendritic cortical microcircuits approximate the backpropagation algorithm." In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc.

Salinas, E and T J Sejnowski (2001). "Gain modulation in the central nervous system: where behavior, neurophysiology, and computation meet." In: *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry* 7.5, pp. 430–440. ISSN: 1073-8584. DOI: 10.1177/107385840100700512.

Saunders, Benjamin T, Jocelyn M Richard, Elyssa B Margolis, and Patricia H Janak (2018). "Dopamine neurons create Pavlovian conditioned stimuli with circuit-defined motivational properties." In: *Nature neuroscience* 21.8, pp. 1072–1083. ISSN: 1546-1726. DOI: 10.1038/s41593-018-0191-4.

Scarinzi, Alfonsina (2020). "4Es Are Too Many: Why Enactive World-Making Does Not Need The Extended Mind Thesis (peer-reviewed - published)." In: I, pp. 237–254. DOI: 10.30687/Jolma/2723-9640/2020/02/005.

Schoups, Aniek, Rufin Vogels, Ning Qian, and Guy Orban (2001). "Practising orientation identification improves orientation coding in V1 neurons." In: *Nature* 412.6846, pp. 549–553. ISSN: 00280836. DOI: 10.1038/35087601. URL: www.nature.com.

Scheutz, Matthias (2003). *Computationalism: new directions*. MIT Press.

Schulz, Laura E, Alison Gopnik, and Clark Glymour (2007). "Preschool children learn about causal structure from conditional interventions." In: *Developmental science* 10.3, pp. 322–332. ISSN: 1363-755X (Print). DOI: 10.1111/j.1467-7687.2007.00587.x.

Schmidhuber, Jürgen (2010). "Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010)." In: *IEEE Transactions on Autonomous Mental Development* 2.3, pp. 230–247. DOI: 10.1109/TAMD.2010.2056368.

Schwartenbeck, Philipp, Thomas FitzGerald, Ray Dolan, and Karl Friston (2013). "Exploration, novelty, surprise, and free energy minimization." In: *Frontiers in Psychology* 4, p. 710. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2013.00710.

Schulman, John, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz (2015a). "Trust region policy optimization." In: *International conference on machine learning*. PMLR, pp. 1889–1897.

Schulman, John, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel (2015b). "High-dimensional continuous control using generalized advantage estimation." In: *arXiv preprint arXiv:1506.02438*.

Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov (2017). "Proximal policy optimization algorithms." In: *arXiv preprint arXiv:1707.06347*.

Schmidt, Maximilian, Rembrandt Bakker, Kelly Shen, Gleb Bezgin, Markus Diesmann, and Sacha Jennifer van Albada (2018). "A multi-scale layer-resolved spiking network model of resting-state dynamics in macaque visual cortical areas." In: *PLOS Computational Biology* 14.10, pp. 1–38. DOI: 10.1371/journal.pcbi.1006359.

Schott, Lukas, Jonas Rauber, Wieland Brendel, and Matthias Bethge (2018). "Robust Perception through Analysis by Synthesis." In: *CoRR* abs/1805.09190. arXiv: 1805.09190. URL: http://arxiv.org/abs/1805.09190.

Schrimpf, Martin et al. (2018). "Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?" In: *bioRxiv preprint*. URL: https://www.biorxiv.org/content/10.1101/407007v2.

Schrimpf, Martin, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo (2020). "Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence." In: *Neuron*. DOI: https://doi.org/10.1016/j.neuron.2020.07.040.

Schölkopf, Bernhard, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio (2021). "Towards Causal Representation Learning." In: *CoRR* abs/2102.11107. URL: https://arxiv.org/abs/2102.11107.

Schmidhuber, J. (1991). "A Possibility for Implementing Curiosity and Boredom in Model-Building Neural Controllers." In: *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*. Ed. by Jean-Arcady Meyer and Stewart W. Wilson, pp. 222–227.

Schultz, W, P Apicella, and T Ljungberg (1993). "Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task." In: *Journal of Neuroscience* 13.3, pp. 900–913. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.13-03-00900.1993.

Schultz, Wolfram, Peter Dayan, and P. Read Montague (1997). "A Neural Substrate of Prediction and Reward." In: *Science* 275.5306, pp. 1593–1599. ISSN: 0036-8075. DOI: 10.1126/science.275.5306.1593.

Sekar, Ramanan, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak (2020). "Planning to Explore via Self-Supervised World Models." In: *ICML*.

Sengupta, Abhronil, Yuting Ye, Robert Wang, Chiao Liu, and Kaushik Roy (2019). "Going Deeper in Spiking Neural Networks: VGG and Residual Architectures." In: *Frontiers in Neuroscience* 13, p. 95. ISSN: 1662-453X. DOI: 10.3389/fnins.2019.00095.

Sherman, S. Murray and R. W. Guillery (2011). "Distinct functions for direct and transthalamic corticocortical connections." In: *Journal of Neurophysiology* 106.3. PMID: 21676936, pp. 1068–1077. DOI: 10.1152/jn.00429.2011.

Sherman, S Murray and R W Guillery (2013). *Functional Connections of Cortical Areas: A New View from the Thalamus*. The MIT Press. ISBN: 9780262314992. DOI: 10.7551/mitpress/9780262019309.001.0001.

Shidara, M, K Kawano, H Gomi, and M Kawato (1993). "Inverse-dynamics model eye movement control by Purkinje cells in the cerebellum." In: *Nature* 365.6441, pp. 50–52. ISSN: 1476-4687. DOI: 10.1038/365050a0.

Shuler, Marshall G. and Mark F. Bear (2006). "Reward Timing in the Primary Visual Cortex." In: *Science* 311.5767, pp. 1606–1609. ISSN: 0036-8075. DOI: 10.1126/science.1123513.

Siegel, Markus, Konrad P Körding, and Peter König (2000). "Integrating Top-Down and Bottom-Up Sensory Processing by Somato-Dendritic Interactions." In: *Journal of Computational Neuroscience* 8.2, pp. 161–173. ISSN: 1573-6873. DOI: 10.1023/A:1008973215925.

Sigala, Natasha and Nikos K. Logothetis (2002). "Visual categorization shapes feature selectivity in the primate temporal cortex." In: *Nature* 415.6869, pp. 318–320. ISSN: 00280836. DOI: 10.1038/415318a. URL: www.nature.com.

Silver, David, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller (2014). "Deterministic Policy Gradient Algorithms." In: *31st International Conference on Machine Learning, ICML 2014* 1.

Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. (2017). "Mastering chess and shogi by self-play with a general reinforcement learning algorithm." In: *arXiv preprint arXiv:1712.01815*.

Silver, David, Satinder Singh, Doina Precup, and Richard S. Sutton (2021). "Reward is enough." In: *Artificial Intelligence* 299, p. 103535.

ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2021.103535.

Singer, Yosef, Yayoi Teramoto, Ben D.B. Willmore, Andrew J. King, Jan W.H. Schnupp, and Nicol S. Harper (2018). "Sensory cortex is optimised for prediction of future input." In: *eLife* 7. ISSN: 2050084X. DOI: 10.7554/eLife.31557.

Sinz, Fabian H., Xaq Pitkow, Jacob Reimer, Matthias Bethge, and Andreas S. Tolias (2019). "Engineering a Less Artificial Intelligence." In: *Neuron* 103.6, pp. 967–979. ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2019.08.034.

Skaggs, W E and B L McNaughton (1996). "Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience." In: *Science* 271.5257, pp. 1870–1873. ISSN: 0036-8075 (Print). DOI: 10.1126/science.271.5257.1870.

Skinner, B F (1938). *The behavior of organisms: an experimental analysis.* Oxford, England: Appleton-Century, p. 457.

— (1951). "How to Teach Animals." In: *Scientific American* 185.6, pp. 26–29. ISSN: 00368733, 19467087.

Smith, Daniel T. and Thomas Schenk (2012). "The Premotor theory of attention: Time to move on?" In: *Neuropsychologia* 50.6. Special Issue: Spatial Neglect and Attention, pp. 1104–1114. ISSN: 0028-3932. DOI: https://doi.org/10.1016/j.neuropsychologia.2012.01.025.

Smith, Linda B. and Lauren K. Slone (2017). "A Developmental Approach to Machine Learning?" In: *Frontiers in Psychology* 8, p. 2124. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2017.02124.

Sommerville, Jessica A and Jean Decety (2006). "Weaving the fabric of social interaction: Articulating developmental psychology and cognitive neuroscience in the domain of motor cognition." In: *Psychonomic Bulletin and Review* 13.2, pp. 179–200. ISSN: 1531-5320. DOI: 10.3758/BF03193831.

Soviany, Petru, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe (2021). "Curriculum Learning: A Survey." In: *CoRR* abs/2101.10382. arXiv: 2101.10382. URL: https://arxiv.org/abs/2101.10382.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: a simple way to prevent neural networks from overfitting." In: *The journal of machine learning research* 15.1, pp. 1929–1958.

Stein, Astrid von, Carl Chiang, and Peter König (2000). "Top-down processing mediated by interareal synchronization." In: *Proceedings of the National Academy of Sciences* 97.26, pp. 14748–14753. ISSN: 0027-8424. DOI: 10.1073/pnas.97.26.14748.

Steinberg, Elizabeth E, Ronald Keiflin, Josiah R Boivin, Ilana B Witten, Karl Deisseroth, and Patricia H Janak (2013). "A causal link between prediction errors, dopamine neurons and learning." In: *Nature Neuroscience* 16.7, pp. 966–973. ISSN: 1546-1726. DOI: 10.1038/nn.3413.

Stoianov, Ivilin, Cyriel Pennartz, Carien Lansink, and Giovanni Pezzulo (2018). "Model-based spatial navigation in the hippocampus-ventral striatum circuit: A computational analysis." In: *PLOS Com-*

putational Biology 14, e1006316. DOI: 10.1371/journal.pcbi.1006316.

Suri, R.E. and W. Schultz (1999). "A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task." In: *Neuroscience* 91.3, pp. 871–890. ISSN: 0306-4522. DOI: https://doi.org/10.1016/S0306-4522(98)00697-6.

Sutskever, Ilya, James Martens, George Dahl, and Geoffrey Hinton (2013). "On the importance of initialization and momentum in deep learning." In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Proceedings of Machine Learning Research 3. Atlanta, Georgia, USA: PMLR, pp. 1139–1147. URL: http://proceedings.mlr.press/v28/sutskever13.html.

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: An introduction*. MIT press. URL: http://incompleteideas.net/book/the-book-2nd.html.

— (1987). "A temporal-difference model of classical conditioning." In: *Proceedings of the ninth annual conference of the cognitive science society*. Seattle, WA, pp. 355–378.

— (1990). "Time-derivative models of Pavlovian reinforcement." In: *Learning and computational neuroscience: Foundations of adaptive networks.* Cambridge, MA, US: The MIT Press, pp. 497–537. ISBN: 0-262-07102-9 (Hardcover).

Suzuki, Mototaka, Dario Floreano, and Ezequiel A Di Paolo (2005). "The contribution of active body movement to visual development in evolutionary robots." In: *Neural Networks* 18.5-6, pp. 656–665.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus (2013). "Intriguing properties of neural networks." In: *arXiv preprint arXiv:1312.6199*.

Tye-Murray, Nancy, Brent P Spehar, Joel Myerson, Sandra Hale, and Mitchell S Sommers (2013). "Reading your own lips: Common-coding theory and visual speech perception." In: *Psychonomic Bulletin and Review* 20.1, pp. 115–119. ISSN: 1531-5320. DOI: 10.3758/s13423-012-0328-5.

Takahashi, Yuji, Geoffrey Schoenbaum, and Yael Niv (2008). "Silencing the critics: understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an Actor/-Critic model." In: *Frontiers in Neuroscience* 2, p. 14. ISSN: 1662-453X. DOI: 10.3389/neuro.01.014.2008.

Tang, Yong, Jens R. Nyengaard, Didima M.G. De Groot, and Hans Jørgen G. Gundersen (2001). "Total regional and global number of synapses in the human brain neocortex." In: *Synapse* 41.3, pp. 258–273. DOI: https://doi.org/10.1002/syn.1083.

Tanaka, Hirokazu, Takahiro Ishikawa, Jongho Lee, and Shinji Kakei (2020). "The Cerebro-Cerebellum as a Locus of Forward Model: A Review." In: *Frontiers in Systems Neuroscience* 14, p. 19. ISSN: 1662-5137. DOI: 10.3389/fnsys.2020.00019.

Team, Open Ended Learning et al. (2021). "Open-Ended Learning Leads to Generally Capable Agents." In: *CoRR* abs/2107.12808. arXiv: 2107.12808. URL: https://arxiv.org/abs/2107.12808.

Thorndike, Edward Lee (1911). *Animal intelligence: Experimental studies.* On cover: The animal behavior series. Lewiston, NY, US: Macmillan Press, pp. viii, 297–viii, 297. DOI: 10.5962/bhl.title.55072.

Tolman, Edward C (1948). "Cognitive maps in rats and men." In: *Psychological review* 55.4, p. 189.

Tsai, Hsing-Chen, Feng Zhang, Antoine Adamantidis, Garret D. Stuber, Antonello Bonci, Luis de Lecea, and Karl Deisseroth (2009). "Phasic Firing in Dopaminergic Neurons Is Sufficient for Behavioral Conditioning." In: *Science* 324.5930, pp. 1080–1084. ISSN: 0036-8075. DOI: 10.1126/science.1168878.

Turing, Alan M. (1950). "Computing Machinery and Intelligence." In: *Mind* LIX.236, pp. 433–460. ISSN: 0026-4423. DOI: 10.1093/mind/LIX.236.433.

Valentin, Vivian V., Anthony Dickinson, and John P. O'Doherty (2007). "Determining the Neural Substrates of Goal-Directed Learning in the Human Brain." In: *Journal of Neuroscience* 27.15, pp. 4019–4026. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.0564-07.2007.

Varela, Francisco J, Evan Thompson, and Eleanor Rosch (1991). *The embodied mind: Cognitive science and human experience.* Cambridge, MA, US: The MIT Press. ISBN: 9780262285476. DOI: https://doi.org/10.7551/mitpress/6730.001.0001.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need." In: *Advances in neural information processing systems*, pp. 5998–6008.

Vasudevan, Rama K, Maxim Ziatdinov, Lukas Vlcek, and Sergei V Kalinin (2021). "Off-the-shelf deep learning is not enough, and requires parsimony, Bayesianity, and causality." In: *npj Computational Materials* 7.1, p. 16. ISSN: 2057-3960. DOI: 10.1038/s41524-020-00487-0.

Vaughan, William (1988). "Formation of equivalence sets in pigeons." In: *Journal of Experimental Psychology: Animal Behavior Processes* 14.1, p. 36.

Veličković, Petar, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio (2017). "Graph attention networks." In: *arXiv preprint arXiv:1710.10903*.

Vithayathil Varghese, Nelson and Qusay H Mahmoud (2020). "A Survey of Multi-Task Deep Reinforcement Learning." In: *Electronics* 9.9. ISSN: 2079-9292. DOI: 10.3390/electronics9091363.

Vlach, Haley and Catherine Sandhofer (2012). "Fast Mapping Across Time: Memory Processes Support Children's Retention of Learned Words." In: *Frontiers in Psychology* 3, p. 46. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2012.00046.

Wang, Jane X., Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Demis Hassabis, and Matthew Botvinick (2018). "Prefrontal cortex as a meta-reinforcement learn-

ing system." In: *Nature Neuroscience* 21.6, pp. 860–868. ISSN: 15461726. DOI: 10.1038/s41593-018-0147-8.

Wang, Yaqing, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). "Generalizing from a Few Examples: A Survey on Few-Shot Learning." In: *ACM Comput. Surv.* 53.3. ISSN: 0360-0300. DOI: 10.1145/3386252.

Wang, Maya Zhe and Benjamin Y Hayden (2021). "Latent learning, cognitive maps, and curiosity." In: *Current Opinion in Behavioral Sciences* 38. Computational cognitive neuroscience, pp. 1–7. ISSN: 2352-1546. DOI: https://doi.org/10.1016/j.cobeha.2020.06.003.

Ward, Dave and Mog Stapleton (2012). "Es are good: Cognition as enacted, embodied, embedded, affective and extended." In: *Consciousness in Interaction: the Role of the Natural and Social Context in Shaping Consciousness*, pp. 89–104. DOI: 10.1075/aicr.86.06war.

Watanabe, Eiji, Akiyoshi Kitaoka, Kiwako Sakamoto, Masaki Yasugi, and Kenta Tanaka (2018). "Illusory Motion Reproduced by Deep Neural Networks Trained for Prediction." In: *Frontiers in Psychology* 9, p. 345. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2018.00345.

Weber, Theophane et al. (2017). "Imagination-Augmented Agents for Deep Reinforcement Learning." In: *CoRR* abs/1707.06203. URL: http://arxiv.org/abs/1707.06203.

Weglage, Moritz, Emil Wärnberg, Iakovos Lazaridis, Daniela Calvigioni, Ourania Tzortzi, and Konstantinos Meletis (2021). "Complete representation of action space and value in all dorsal striatal pathways." In: *Cell Reports* 36.4. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2021.109437.

Weisberg, Jill, Miranda van Turennout, and Alex Martin (2006). "A Neural System for Learning about Object Function." In: *Cerebral Cortex* 17.3, pp. 513–521. ISSN: 1047-3211. DOI: 10.1093/cercor/bhj176.

Weilbächer, Regina A and Sebastian Gluth (2016). "The Interplay of Hippocampus and Ventromedial Prefrontal Cortex in Memory-Based Decision Making." In: *Brain sciences* 7.1. ISSN: 2076-3425 (Print). DOI: 10.3390/brainsci7010004.

Wen, Haiguang, Kuan Han, Junxing Shi, Yizhen Zhang, Eugenio Culurciello, and Zhongming Liu (2018). "Deep Predictive Coding Network for Object Recognition." In: *CoRR* abs/1802.04762. URL: http://arxiv.org/abs/1802.04762.

Wexler, Mark (2003). "Voluntary head movement and allocentric perception of space." In: *Psychological Science* 14.4, pp. 340–346.

Wexler, Mark and Jeroen JA Van Boxtel (2005). "Depth perception by the active observer." In: *Trends in cognitive sciences* 9.9, pp. 431–438.

Wheeler, Diek W, Charise M White, Christopher L Rees, Alexander O Komendantov, David J Hamilton, and Giorgio A Ascoli (2015). "Hippocampome.org: a knowledge base of neuron types in the rodent hippocampus." In: *eLife* 4. Ed. by Frances K Skinner, e09960. ISSN: 2050-084X. DOI: 10.7554/eLife.09960.

Whittington, James CR and Rafal Bogacz (2017). "An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity." In: *Neural computation* 29.5, pp. 1229–1262.

White, Robert W (1959). "Motivation reconsidered: The concept of competence." In: *Psychological Review* 66.5, pp. 297–333. ISSN: 1939-1471(Electronic),0033-295X(Print). DOI: 10.1037/h0040934.

Wilson, Margaret (2002). "Six views of embodied cognition." In: *Psychonomic Bulletin & Review* 9.4, pp. 625–636. ISSN: 1531-5320. DOI: 10.3758/BF03196322. URL: https://doi.org/10.3758/BF03196322.

Wilson, Aaron, Alan Fern, and Prasad Tadepalli (2014). "Using Trajectory Data to Improve Bayesian Optimization for Reinforcement Learning." In: *Journal of Machine Learning Research* 15.8, pp. 253–282. URL: http://jmlr.org/papers/v15/wilson14a.html.

Wilson, Robert A., Lucia Foglia, Lawrence Shapiro, and Shannon Spaulding (2021). "Embodied Cognition." In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University.

Williams, Ronald J (1992). "Simple statistical gradient-following algorithms for connectionist reinforcement learning." In: *Machine learning* 8.3, pp. 229–256.

Wimmer, G. Elliott and Daphna Shohamy (2012). "Preference by Association: How Memory Mechanisms in the Hippocampus Bias Decisions." In: *Science* 338.6104, pp. 270–273. DOI: 10.1126/science.1223252.

Witt, Jessica K, Dennis R Proffitt, and William Epstein (2004). "Perceiving Distance: A Role of Effort and Intent." In: *Perception* 33.5. PMID: 15250663, pp. 577–590. DOI: 10.1068/p5090.

Witt, Jessica K and Dennis R Proffitt (2005). *See the ball, hit the ball: Apparent ball size is correlated with batting average.* DOI: 10.1111/j.1467-9280.2005.01640.x.

Witt, Jessica K, Dennis R Proffitt, and William Epstein (2005). *Tool Use Affects Perceived Distance, But Only When You Intend to Use It.* DOI: 10.1037/0096-1523.31.5.880.

Witt, Jessica K and Mila Sugovic (2010). "Performance and Ease Influence Perceived Speed." In: *Perception* 39.10. PMID: 21180356, pp. 1341–1353. DOI: 10.1068/p6699.

Witt, Jessica K. (2011). "Action's Effect on Perception." In: *Current Directions in Psychological Science* 20.3, pp. 201–206. DOI: 10.1177/0963721411408770.

Witt, Jessica K (2017). "Action potential influences spatial perception: Evidence for genuine top-down effects on perception." In: *Psychonomic Bulletin and Review* 24.4, pp. 999–1021. ISSN: 1531-5320. DOI: 10.3758/s13423-016-1184-5. URL: https://doi.org/10.3758/s13423-016-1184-5.

Wolpert, David H and William G Macready (1997). "No free lunch theorems for optimization." In: *IEEE transactions on evolutionary computation* 1.1, pp. 67–82.

Woźniak, Stanisław, Angeliki Pantazi, Thomas Bohnstingl, and Evangelos Eleftheriou (2020). "Deep learning incorporating biologi-

cally inspired neural dynamics and in-memory computing." In: *Nature Machine Intelligence* 2.6, pp. 325–336. ISSN: 2522-5839. DOI: 10.1038/s42256-020-0187-0.

Yamins, D. L. K., H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo (2014). "Performance-optimized hierarchical models predict neural responses in higher visual cortex." In: *Proceedings of the National Academy of Sciences* 111.23, pp. 8619–8624. ISSN: 0027-8424. DOI: 10.1073/pnas.1403112111.

Yamins, Daniel L.K. and James J. DiCarlo (2016). "Using goal-driven deep learning models to understand sensory cortex." In: *Nature Neuroscience* 19.3, pp. 356–365. ISSN: 15461726. DOI: 10.1038/nn.4244.

Yin, Henry H and Barbara J Knowlton (2006). "The role of the basal ganglia in habit formation." In: *Nature Reviews Neuroscience* 7.6, pp. 464–476. ISSN: 1471-0048. DOI: 10.1038/nrn1919.

Yu, Chen and Dana H. Ballard (2004). "On the Integration of Grounding Language and Learning Objects." In: *Proceedings of the 19th National Conference on Artifical Intelligence*. AAAI04. AAAI Press, 488–493. ISBN: 0262511835.

Yu, Tianhe, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine (2019). "Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning." In: *CoRR* abs/1910.10897. URL: http://arxiv.org/abs/1910.10897.

— (2020). "Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning." In: *Proceedings of the Conference on Robot Learning*. Ed. by Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura. Vol. 100. Proceedings of Machine Learning Research. PMLR, pp. 1094–1100. URL: http://proceedings.mlr.press/v100/yu20a.html.

Zador, Anthony M (2019). "A critique of pure learning and what artificial neural networks can learn from animal brains." In: *Nature communications* 10.1, pp. 1–7.

Zeiler, Matthew D. (2012). "ADADELTA: An Adaptive Learning Rate Method." In: *CoRR* abs/1212.5701. arXiv: 1212.5701. URL: http://arxiv.org/abs/1212.5701.

Zeiler, Matthew D. and Rob Fergus (2014). "Visualizing and Understanding Convolutional Networks." In: *Computer Vision – ECCV 2014*. Ed. by David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars. Cham: Springer International Publishing, pp. 818–833. ISBN: 978-3-319-10590-1.

Zhuang, Fuzhen, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He (2020). "A Comprehensive Survey on Transfer Learning." In: *Proceedings of the IEEE* PP, pp. 1–34. DOI: 10.1109/JPROC.2020.3004555.

Zylberberg, Joel, Jason Timothy Murphy, and Michael Robert DeWeese (2011). "A Sparse Coding Model with Synaptically Local Plasticity and Spiking Neurons Can Account for the Diverse Shapes of V1 Simple Cell Receptive Fields." In: *PLOS Computational Biology* 7, pp. 1–12. DOI: 10.1371/journal.pcbi.1002250.

de Haan, Bianca, Paul S. Morgan, and Chris Rorden (2008). "Covert orienting of attention and overt eye movements activate identical brain regions." In: *Brain Research* 1204, pp. 102–111. ISSN: 0006-8993. DOI: https://doi.org/10.1016/j.brainres.2008.01.105.

van Bergen, Ruben S and Nikolaus Kriegeskorte (2020). "Going in circles is the way forward: the role of recurrence in visual inference." In: *Current Opinion in Neurobiology* 65. Whole-brain interactions between neural circuits, pp. 176–193. ISSN: 0959-4388. DOI: https://doi.org/10.1016/j.conb.2020.11.009.

Çatal, Ozan, Tim Verbelen, Toon Van de Maele, Bart Dhoedt, and Adam Safron (2021). "Robot navigation as hierarchical active inference." In: *Neural Networks* 142, pp. 192–204. ISSN: 08936080. DOI: 10.1016/j.neunet.2021.05.010.

Erklärung über die Eigenständigkeit der erbrachten wissenschaftlichen Leistung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Bei der Auswahl und Auswertung folgenden Materials haben mir die nachstehend aufgeführten Personen in der jeweils beschriebenen Weise entgeltlich/ unentgeltlich geholfen.

1. Bei Veröffentlichung 1 und 3 bin ich nicht Erstautor denn die Studie wurden zu großen Teilen von Sabine König konzipiert, gemanaged und geschrieben. Ich habe die virtuelle Stadt gebaut, Eye-tracking integriert, Daten analysiert und das Manuskript überarbeitet. Zusätzlich haben Debora Nolte, Laura Duesberg, Nicolas Kuske, Ashima Keshava und Kirsten Rittershofer bei der Datenmessung, Interactive-Map Implementation, Datenanalyse und Manuskriptüberarbeitung geholfen. Genauere Attributionen sind in den jeweiligen „author contributions“ Kapiteln der zwei Paper. Peter König hat die Studien konzipiert, betreut, Ergebnisse diskutiert und an den Manuskripten geschrieben.

2. Veröffentlichung 2 wurde von Peter und Sabine König in regelmäßigen Treffen betreut, mitkonzipiert, diskutiert und schriftlich überarbeitet.

Veröffentlichungen 4 und 5 wurden in regelmäßigen Treffen von Gordon Pipa, Peter König und Kai-Uwe Kühnberger betreut und diskutiert. Auch das Manuskript wurde von allen mit überarbeitet.

3. Alle restlichen Kapitel der Dissertation (Einleitung incl. Unterkapitel, Diskussion und Fazit) wurden von mir allein verfasst. Der Inhalt wurde mit meinen Betreuern besprochen und Feedback wurde von mir mit eingebunden. Sprachliche Details wurden von Joshua Clay und mit Hilfe von Grammarly überarbeitet.

4. Veröffentlichung 6 im Appendix wurde zusammen mit Johannes Schrumpf und Yannick Tessenow verfasst. Die Datensammlung und Auswertung wurde von den drei Erstautoren in gleichem Umfang bewerkstelligt und von Peter Naeve und Falk Heuer unterstützt. Das Projekt wurde von Helmut Leder, Ulrich Ansorge und Peter König konzipiert und betreut.

Weitere Personen waren an der inhaltlichen materiellen Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater oder andere Personen) in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

....

(Ort. Datum) (Unterschrift)

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

https://bitbucket.org/amiede/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/