Neurocomputational Mechanisms for Adaptive Self-Preservative Robot Behaviour

Dissertation

zur Erlangung des Doktorgrades
an der Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Informatik
der Universität Hamburg

eingereicht beim Fach-Promotionsausschuss Informatik von
Nicolás Ignacio Navarro Guerrero

Hamburg, 2016

Gutachter:	Prof. Dr. Stefan Wermter Dept. of Computer Science University of Hamburg, Germany
	Dr. Robert Lowe Interaction Lab, School of Informatics University of Skövde, Sweden Division of Cognition and Communication University of Gothenburg, Sweden
	Prof. Dr. Jianwei Zhang Dept. of Computer Science University of Hamburg, Germany
Vorsitzender der Prüfungskommission:	Prof. Dr. Frank Steinicke Dept. of Computer Science University of Hamburg, Germany
Tag der Disputation:	Tuesday 3rd May, 2016

Dedicado a mi hija Paz

Abstract

The field of neurocognitive robotics takes the processing mechanisms of the brain as inspiration and guidance: computer implementations of robot perception and action should be based on brain-like neural architectures and biologically plausible learning mechanisms. Unsupervised learning and reinforcement learning have led to good results on the emergence of internal sensory representations and intelligent reward-seeking behaviours, respectively. However, other aspects of animal behaviour are generally not considered, even though it has been argued that only a more comprehensive study of animal behaviour can lead to a deeper understanding of intelligent behaviour. This thesis does not attempt to provide a comprehensive model of animal behaviour, but rather tries to draw attention to the need for it by presenting the potential of neglected aspects of animal behaviour such as self-preservative behaviour.

Self-preservative behaviours are believed to impose the ground rules for more complex and motivated behaviour. Although many of these innate responses are hard-coded in the brain, they are not sufficient for the organisms' survival. They have to adapt, by learning, to new and unexpected situations within their lifetime and thereby be able to interact effectively with their environment. A key component on the lifetime adaptation is the formation of associations/memories between environmental predictors and relevant events, which mainly rely on punishment and reward learning.

We[1] postulate that a deeper understanding of innate and learned defensive mechanisms could also be helpful in developing future robot generations, making them more adaptable and robust. Therefore, in this thesis, we study and develop three neurocomputational self-preservative mechanisms in the context of humanoid service robots to demonstrate the potential and feasibility of including bio-inspired adaptive self-preservative mechanisms as part of real-world robotic systems. Our aim is to present possible ways in which robots can be endowed with such adaptive self-preservative mechanisms at different neurocognitive levels, going from abstract biological models to neurocomputational models.

The first experiment addresses the problem of search for an appetitive stimulus. Here a reinforcement learning (SARSA) algorithm was optimized to learn in a real-world scenario and manoeuvre a humanoid robot towards a charging station.

[1] Throughout the thesis the 'scientific' we is used instead of the personal "I", even if personal opinions and ideas are expressed.

The second experiment focuses on the role of punishment and nociceptive sensory input in motor learning. Both types of feedback play an important role in driving attention, and modulating decision making and action. However, they have not been thoroughly studied in computational models. Here, we compared the effect of both types of feedback on an Actor-Critic learning algorithm (CACLA).

Finally, in our last experiment, we studied the role of noxious stimuli in the formation of anticipatory behaviour. This experiment is based on Pavlovian and instrumental conditioning and how environmental cues can be used to anticipate negative outcomes. A hybrid approach using an echo state network (ESN) and a dopamine modulated Pavlovian conditioning model was used to anticipate nociceptive sensory input based on auditory cues.

In all three experiments we showed how often neglected, self-preservative mechanisms could solve meaningful artificial intelligence problems while providing the basis for new neuro-inspired computational processes. In particular, we showed how bio-inspired sensorimotor signals associated with nociception and pain can be exploited for learning beyond triggering reactive behaviours. We also developed novel extensions to the learning algorithms used.

Zusammenfassung

Im Gebiet der neurokognitiven Robotik werden die Verarbeitungsmechanismen des Gehirns als Inspiration und Leitlinie verwendet. Inspiriert von diesen Mechanismen des Gehirns sollten Computerimplementationen der Roboterwahrnehmung und -aktion auf neuronalen Architekturen und biologisch plausiblen Lernmechanismen basieren. Die Verwendung von Unsupervised- und Reinforcement-Learning hat zu guten Ergebnissen in der Bildung interner sensorischer Repräsentationen und intelligentem, durch Belohnung gesteuertem Verhalten geführt. Allerdings werden andere Aspekte im Verhalten von Tieren in der Regel nicht berücksichtigt, obwohl oft argumentiert wird, dass nur eine umfassendere Untersuchung des Verhaltens von Tieren zu einem tieferen Verständnis von intelligentem Verhalten führen kann, wie es in dieser Arbeit diskutiert wird.

Selbsterhaltung ist ein Beispiel für eine solche bisher vernachlässigte, aber ursprüngliche und wesentliche Fähigkeit eines jeden Organismus. Es wird vermutet, dass der Selbsterhaltungstrieb Grundregeln für komplexeres und motiviertes Verhalten setzt. Obwohl viele dieser angeborenen Reaktionen fest im Gehirn codiert sind, sind sie nicht ausreichend, um das Überleben des Organismus zu sichern. Er muss sich durch Lernen an neue und unerwartete Situationen in seinem Leben anpassen und ist nur so in der Lage, effektiv mit seiner Umwelt zu interagieren. Eine Schlüsselkomponente für die lebenslange Anpassung ist die Bildung von Assoziationen beziehungsweise Erinnerungen von Umweltprädiktoren und relevanten Ereignissen, welche vor allem auf Lernen durch Bestrafung und Belohnung angewiesen sind.

Wir setzen voraus, dass ein tieferes Verständnis der angeborenen und erlernten Schutzmechanismen auch hilfreich bei der Entwicklung künftiger Robotergenerationen sein könnte, um diese Roboter anpassungsfähig und robust zu machen. Daher untersuchen und entwickeln wir in dieser Arbeit drei neuroinformatische Selbsterhaltungsmechanismen im Kontext humanoider Serviceroboter und zeigen das Potential und die Durchführbarkeit der Integration von bio-inspirierten adaptiven Selbsterhaltungsmechanismen als Teil realer Robotersysteme auf. Unser Ziel ist es, mögliche Ansätze zu präsentieren, durch die Roboter auf verschiedenen neurokognitiven Ebenen mit adaptiven Selbsterhaltungsmechanismen ausgestattet werden können, angefangen mit abstrakten biologischen Modellen bis hin zu implementierten neuroinformatischen Modellen.

Das erste Experiment behandelt das Problem der energetischen Autonomie. Wir trainierten einen Roboter darauf, Belohnung durch appetitive Stimuli anzustreben.

Es wurde ein Reinforcement Learning Algorithmus (SARSA) implementiert und weiterentwickelt, der in einem realen Szenario lernen und einen humanoiden Roboter zu einer Ladestation manövrieren soll.

Das zweite Experiment konzentriert sich auf die Rolle der Bestrafung und nozizeptiver Stimuli beim Erlernen motorischer Aktionen. Diese Arten von Feedback spielen eine wichtige Rolle bei der Steuerung von Aufmerksamkeit und der Modularisierung der Entscheidungsfindung. Sie wurden jedoch noch nicht vollständig in Computermodellen untersucht. Wir vergleichen die Wirkung dieser Arten von Feedback auf einen Actor-Critic-basierten Lernalgorithmus (CACLA).

Im letzten Experiment untersuchen wir die Rolle noxischer Stimuli in der Bildung antizipierenden Verhaltens. Dieses Experiment basiert auf Pawlowscher und instrumenteller Konditionierung und untersucht, wie Umweltreize verwendet werden können, um negative Folgen zu antizipieren. Ein hybrider Ansatz unter der Verwendung eines Echo State Networks (ESN) und Dopamin-modulierender Pawlowschen Konditionierung wurde verwendet, um noxische sensorische Stimuli basierend auf auditorischen Reizen zu antizipieren.

In allen drei Versuchen haben wir gezeigt, wie bisher vernachlässigte Selbsterhaltungsmechanismen bedeutsame Probleme der künstlichen Intelligenz lösen können und gleichzeitig die Grundlage für neue neuroinspirierte Rechenprozesse liefern. Besonders haben wir gezeigt, wie biologisch inspirierte sensomotorische Signale, wie zum Beispiel Nozizeption und Schmerz, genutzt werden können, um Lernverfahren zu verbessern. Außerdem wurden in dieser Arbeit Erweiterungen zu den verwendeten Lernalgorithmen entwickelt.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

Self-preservative mechanisms play a fundamental role for both living and artificial agents. Although many animals' and humans' self-preservative responses are guided by reactive "hard-coded" behaviour, simple stimulus-response mappings alone are not sufficient to produce the protective behaviour needed in complex environments. Scientific evidence shows a tight relationship between self-preservative mechanisms such as bio-regulatory processes, self-protective and evaluative affective mechanisms and the development of intelligence (Arbib and Fellous, 2004; Ziemke and Lowe, 2009) in biological agents. The question arises as to how artificial systems can benefit from similar neural mechanisms for self-preservation. Artificial agents must also evaluate and adapt to new or unexpected situations during their lifetime by learning to bind new or neutral stimuli with evaluated responses, and possibly other learned behaviours. Several points of view have been discussed (Arbib and Fellous, 2004; Ziemke and Lowe, 2009; among others) from which it can be concluded that if we want artificial systems to act properly in highly dynamic environments and to co-exist with other autonomous systems and humans in a natural way, they will need similar adaptive regulatory and learning mechanisms that help them to be more efficient in changing surroundings and to produce safe and successful behaviours.

In recent years, there has been a growing interest in the implications and development of safety systems and standards for human-robot interaction, which has mainly focused on human safety (Harper and Virk, 2010; Murphy and Woods, 2009), neglecting the fact that robot self-protection also leads to human safety. Today's cognitive robots have been endowed with numerous sensors and actuators but, surprisingly, most of them do not incorporate sophisticated mechanisms of self-protection, if any at all, both for human protection or robot self-protection. So far, the most common approach has been to physically separate the robot's workspace in time and space from the human's workspace to ensure safe operation in industrial applications. However, the inexorable need for robots to coexist with humans in order to tackle a broad and steadily growing number of tasks inevitably requires a different paradigm.

Initial attempts focus mainly on limiting the robot's size, weight and power[1], and increasing the robot's compliance[2]. In the literature many examples of self-protective robots can be found, but they are still reduced to very simple reactive systems. For instance, it is easy to find various implementations of collision avoidance (Cellier et al., 1995; Yan et al., 2012, 2013) which can be seen from a self-protective point of view and also for human safety. Other more sophisticated examples try to mimic whole body reactions to reduces damage due to fall and self-collision (Ha and Liu, 2015; Ruiz-del-Solar et al., 2010; Shimizu et al., 2012). Here the robot uses proprioceptive information to adapt its posture and thus reduce damage when falling. Although these measures are important for safety they are engineered for pre-defined situations and based on reflexes with little or no adaptation nor learning. Unfortunately, these solutions are not enough to ensure human, infrastructure, or robot safety both inside and outside the laboratory.

In contrast to the above-mentioned research, the experiments discussed in this thesis – detailed in Chapters 5 to 7 – aim to demonstrate the potential and feasibility of including biologically-inspired adaptive self-preservative mechanisms as part of real-world robotic systems. Three experiments motivated at different neurocognitive levels, going from abstract biological models to neurocomputational models, show possible ways in which robots can be endowed with such adaptive self-preservative mechanisms.

1.1. Aim and Objectives

While adaptive self-preservative behaviours are relevant in a wide variety of domains, the experiments shown here are focused and constrained to the representative domain of a domestic service robot (KSERA, Yan et al., 2012, 2013). In this project, brain microcircuits involved in learning self-preservative mechanisms, in particular, self-preservative mechanisms for a cognitive robot architecture based on biological principles known to be present in the basal ganglia and amygdala are developed and examined. The following research questions and hypotheses are addressed:

- Could self-preservation be driven by appetitive behaviours in addition to avoidance behaviours?
- What role could punishment and nociception take in robot learning, besides triggering reactive behaviours?

[1] Nao, Paro, Roomba
[2] Baxter, Mekabot, REEM-C

- What does a computational model need in order to enable aversive conditioning in a cognitive robot framework?

On a more detailed level, objectives associated with each particular research question will be given later in their respective chapters.

Threat detection is relevant to a wide variety of domains. Apart from self-preservative robots, which lead to safer human-robot interaction, there is also potential for related threat applications in other domains, such as monitoring of patients in hospitals or the elderly in their own homes. As mentioned earlier in this chapter, there is related research on the detection of threat and self-protection (e.g. Ruiz-del-Solar et al., 2010; Shimizu et al., 2012; Yan et al., 2012, 2013; among others), however, the approach presented here departs from an approach using solely reactive and non-adaptive behaviours and moves towards adaptive ones.

1.2. Contribution of the Work

The focus of this work is on unsupervised and semi-supervised (i.e. human guided) learning mechanisms that allow the acquisition of adaptive self-preservative skills. Advanced pre- and post-processing techniques for handling sensory inputs, though important robotic features, were not developed here. The main contributions of this work are as follows:

- Firstly, a new real-world learning algorithm based on SARSA and supervised reinforcement learning has been developed. This algorithm successfully applies a novel Gaussian distributed activation pattern of the state input vector for generalization.
- Secondly, it was shown that nociceptive input signals, as input to a neural network, can safely be used alongside rewarding feedback and can even aid motor skill learning. Furthermore, nociceptive input signals are more effective than punishment and can even correct the detrimental effects produced by punishment.
- Finally, a novel architecture for the acquisition of long-lasting fear memories was developed. In the presented simulations, this architecture can trigger anticipatory responses. It shows stimulus generalization and discrimination which is consistent with animal and human studies at a functional level.

1.3. Research Methodology

The work produced for this thesis is biologically-inspired though only from a functional point of view. No attempt is made to produce a detailed biological model of any kind.

This thesis is broadly concerned with various aspects of self-preservative behaviour, from which we mainly focus on aspects associated with adaptive self-protection but also with adaptive appetitive behaviour. Throughout the thesis we often refer to different aspects of adaptive self-preservative mechanisms. Although, adaptation in a broader sense may encompass both adaptation by learning, observable at the level of the individual, and also evolutionary adaptation, notable only at the species-level, here, we primarily refer to lifetime adaptations through learning.

To focus this research, the representative domain of a domestic service robot was chosen as initial scenario. Although this is true the goal was not to deliver plug-and-play products but to explore the benefits of adaptive self-preservative mechanisms under this context.

The development and evaluation of the different biologically-inspired robot architectures, besides offering to capture relevant constraints of nature's robust and outstanding solutions for living agents, offers the possibility to bridge neural and behavioural levels. Through computational modelling a better understanding of learning mechanisms and neural information processing underlying reinforcement learning, conditioning dynamics and affective reactions was attempted. Other advantages of combining biologically-inspired architectures targeting a real robot can be listed as follows:

- Firstly, a robot provides useful constraints on both sensory stimuli and actions.
- Secondly, by constraining the system to work in a real environment, the adaptive processes that must take place are clarified.
- Thirdly, the robot's embodiment further constrains specific behaviours and sensory processing that might take place.
- Finally, the actions demonstrated in neuropsychological experimental scenarios can very well be tested and evaluated by a mobile robot in a similar experimental set-up.

The project is structured into three experiments. The first experiment develops and evaluates the seeking of appetitive stimuli. This is included as a subset of self-preservative behaviours to emphasize the importance of achieving an adequate balance between avoidance and seeking behaviours. The second experiment,

explores the effects of punishment and nociceptive signals on learning. This experiment focuses on autonomous inverse kinematic learning, and converts abstract and external feedback into perceptual input thus exploiting the robot's embodiment. The third and final experiment addresses the acquisition of affective reactive behaviours. This experiment focuses on cue-dependent fear conditioning and how sensory cues can be used for the development of artificial self-preservative systems.

1.4. Structure of the Thesis

This thesis is structured into 7 chapters. A great amount of the content has been published as papers or distributed as internal reports and reworked for this thesis. The initial chapters place this thesis within the field of cognitive robotics. They provide a non-exhaustive overview of the broad field touched on by this thesis.

The present chapter, Chapter 1, introduces the concepts that influenced the design of the neural architectures and experiments, and also defines the scope and objectives of this thesis.

Chapter 2 presents the neurobiological foundations of adaptive self-preservative mechanisms. These include the basic mechanisms of reinforcement learning and conditioning learning focusing on a functional, rather than physiological perspective. Additionally, an overview on the state of the art on artificial self-protective mechanisms is provided in Chapter 3.

Design and results of the computational studies, including simulations and real-world robot experiments, are described in Chapter 5, 6 and 7. The experiments show different settings where robots can benefit from learned self-preservative mechanisms. In Chapter 5, addressing the first experiment, a humanoid robot learns to navigate towards a target region to recharge or to grasp an object. In Chapter 6 the effects on learning of nociceptive input signals in comparison to punishment is studied. Finally, Chapter 7 introduces our neural computational model of the amygdala, including a functional description, physiological mappings and results from simulations. The model was designed and tested to process real auditory signals and to mimic neurobiological behaviours.

In Chapter 8 resumes the outcomes of the individual experiments. It also discusses their contribution to the research field of embodied cognition and presents suggestions for future work.

Finally, Appendix 4 gives an overview of the previous versions of the different methodologies used in this research. This includes Hebbian learning, multilayer perceptron, back-propagation algorithms, reinforcement learning, eligibility trace, and echo state networks (ESN). A detailed view on the developed modifications is

presented alongside the corresponding experiment in their corresponding chapters.

Chapter 2
Neural Self-Preservative Circuitry in Mammals

Self-preservative mechanisms such as eating, drinking, thermoregulatory, aggressive, and sexual behaviours belong to the essential capabilities of any organism (LeDoux, 2012; Sternson, 2013). As shown by Macnab and Koshland (1972), even single-cell organisms are able to detect and respond to both harmful and appetitive stimuli. Because these responses are critical for survival, organisms need to quickly assess the biological significance of the stimuli and respond. Hence, most of the responses are hard-coded and involuntarily *triggered*. This suggests that such evaluation is based on simple sensory cues processed by subcortical areas. Evidence from both animal and human studies supports this idea. For instance, rats can recognize predators only by their distinctive odour (LeDoux, 2012) and humans can process subcortically visually threatening stimuli (Canteras, 2002) and even certain facial cues (M. H. Johnson, 2005). However, it may also exist a 'slower' but more accurate evaluation via a cortical route, which may rectify or enhance the responses activated by the fast and coarse subcortical evaluation mechanisms (Canteras, 2002; Resnik et al., 2011). In this work, the subcortical aspects of survival, nocifensive behaviours and evaluation of biologically significant stimuli will be considered. Nonetheless, as hard-coded responses are not enough to cope with an ever changing environment, this thesis focuses primarily on the subcortical mechanisms that permit adaptation of innate self-preservative behaviours during the organism's lifetime, i.e. learning.

Both innate responses and learned adaptations are task- and species-specific, nonetheless vertebrates have developed common core mechanisms to regulate self-preservative behaviours (Krichmar, 2008; LeDoux, 2012). Examples of such mechanisms are homeostatic maintenance, thermoregulation, defence against predators and reproduction. Although many aspects of these mechanisms are innate and regulated subconsciously, they can be associated with neutral stimuli or modulated via specialized and sophisticated neural circuits.

The reason for such specialized subcortical circuits is to facilitate the recognition of crucial environmental stimuli and quickly elicit appropriate behavioural responses.

Furthermore, the nature of pressing survival needs may likely demand different behavioural responses. For instance, behavioural responses linked to environmental cues indicative of a predator or other sources of harm vary largely from those elicited under environmental cues linked to potential food sources or a potential mate.

At the core of one of these specialized subcortical circuits is the amygdala. The amygdala is responsible for the association of neutral sensory cues with stimuli of innate significance in a process known as conditioning, see Section 2.5. Via conditioning, neutral stimuli with high predictive power of the occurrence of biologically significant stimuli acquire motivational relevance. As a result, animals can modulate onset, duration, intensity and maybe other aspects of innate responses (LeDoux, 2012).

The external sensory information used to judge the biological significance of an event can be categorized into three types based on the kind of responses or survival circuits that are *activated*. For each of these categories an innate and a learned variant can be identified (LeDoux, 2012):

'Drives' or 'triggers' activate specific survival circuits. *Innate or unconditioned stimuli* elicit innate responses such as when the *smell* of food initiates salivation. On the other hand, *learned or conditioned stimuli* elicit innate responses only after being associated with an innate stimulus via Pavlovian conditioning such as when a *tone* paired with the delivery of food initiates salivation.

Incentives or motivational stimuli modulate instrumental goal-directed behaviour. This is a two-stage process, firstly the agent is in a searching or anticipatory state and secondly, when the goal is reached, an innate consummatory response is performed. *Innate or primary incentives* elicit approach or avoidance behaviour towards the stimuli such as the *presentation of food* to a hungry animal. *Learned or secondary incentives* guide approach or avoidance behaviour based on experience such as in a lever pressing for food scenario when a *tone* signals the availability of food.

Reinforcer stimuli support the learning of Pavlovian or instrumental associations by changing the probability of an instrumental response being executed. *Innate or primary reinforcers* induce the formation of associations between neutral stimuli and unconditioned stimuli through Pavlovian and instrumental conditioning such as *tasty food* or *shelter*. *Learned or secondary reinforcers* induce the formation of associations between neutral stimuli and unconditioned stimuli through Pavlovian conditioning or association with other goal-directed responses through instrumental conditioning such as *money*.

So far, the directly observable responses from survival circuits have been mentioned, namely innate behavioural responses and instrumental goal-directed be-

haviours. However, the activation of survival circuits has an invisible and broader impact on cognitive processes such as general arousal and attention (LeDoux, 2012), see Figure 2.1. This is achieved by releasing specific neurotransmitters such as dopamine (DA), norepinephrine (NE), serotonin (5-HT) and acetylcholine (ACh) and peptides, e.g. orexins (ORX). The neurotransmitters have a quick and generalized modulatory effect on the brain, particularly with respect to reward anticipation (DA), novelty and saliency (NE), stress and threatening stimuli (5-HT), attention (ACh), and sleep cycles and energy expenditure (ORX) (Krichmar, 2008).

Survival circuits also modulate behaviour and cognitive processes by releasing hormones into the circulatory system such as cortisol, epinephrine and norepinephrine. The effects of these hormones are considerably slower than those attributed to neurotransmitters. It is believed that hormones help to sustain survival states for a longer time (LeDoux, 2012). Their extensive impact on survival circuits, via behavioural responses and more importantly via neuromodulatory systems, suggests that they may constitute the foundation for motivated behaviour and ultimately cognition in higher organisms (Krichmar, 2008; Sternson, 2013).

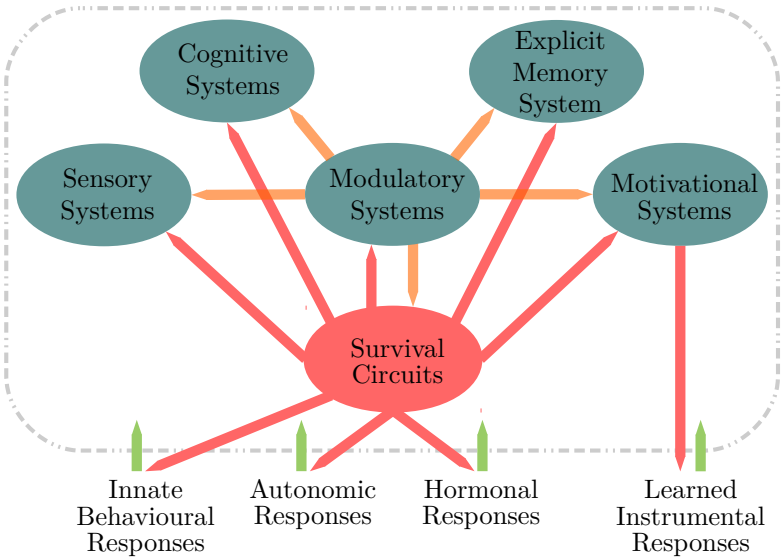


Figure 2.1 – Overview of the interactions of survival circuits with cognitive processes and behavioural responses. Survival circuits have a broad impact on many systems on the brain and bodily responses (red arrows). Sensory feedback (green arrows) provides necessary information to guide behaviour and form memories. All other systems (yellow arrows) are affected by the survival circuit, biological relevant stimuli are prioritized. Adapted from LeDoux (2012).

2.1. Pain System

Pain denotes a complex psychological and neurophysiological mechanism used to protect the body from injury (Westlund and Sluka, 2013). The experience of pain involves most of the central nervous system (CNS) and its effects can have long-lasting behavioural repercussions; evidence for it are the reports of pain in the absence of any physiological cause such as chronic pain (Staahl and Drewes, 2004).

Pain is defined by the International Association for the Study of Pain (IASP)[1] as "an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage" (Bonica, 1979, p. 250).

Under the definition of the IASP, pain is always subjective and it is defined in terms of an experience, therefore, solely stimulation of nociceptors or nociceptive pathways is not considered as pain. This is in line with the fact that the same noxious stimulus is perceived differently under different circumstances or different internal states such as anxiety and expectation (Brooks and Tracey, 2005; McGrath, 1994).

The pain system is responsible for processing pain signals in humans and other mammals. It consists of specialized receptors called nociceptors, several nociceptive pathways and brain structures responsible for processing and modulating diverse responses called *nocifensive behaviours* such as somatic and autonomic responses, endocrine changes, affective responses and memory (Westlund and Willis, 2012).

Pain may be perceived as having different intensities and associated with a variety of sensations, depending on the tissue affected by the noxious stimuli and the type of nociceptor activated. The pain perceived at the area of injured tissue is referred to as *primary hyperalgesia*, whereas the pain perceived at adjacent areas is known as *secondary hyperalgesia* (Staahl and Drewes, 2004). Pain perception can also be affected by the temporal and spatial separation of stimuli. When a stimulus is applied with low frequencies ($< 0.\bar{3}\,\mathrm{Hz}$) the pain intensity is not affected. However, if the same stimulus is applied at higher frequencies the pain perception will increase by a process known as *temporal summation* or *central integration*. Similarly, if a large area is stimulated, a greater intensity of pain will be perceived in a process called *spatial summation*. Pain may also be perceived far from the original stimulus which is known as *referred pain* (Staahl and Drewes, 2004). With respect to the temporal evolution of pain perception, the acute, sharp and pricking pain perceived immediately after the painful stimulus is known as *first pain* or *fast pain*, whereas the diffuse and weaker pain perceived after the first pain is known as *second pain* or *slow pain* (Staahl and Drewes, 2004).

[1] `https://www.iasp-pain.org/`

Pain Pathways

Pain or nociceptive pathways refer to neural circuits involved in the transmission of information related to noxious stimuli and behavioural responses, which include pain sensations, inflammation, reflexive withdrawal, scratching, endocrine changes, motivated and affective responses and learning (Gebhart and Schmidt, 2013, p. 2186).

The pain pathways start at the nociceptor level; nociceptors are specialized peripheral nerve endings that transduce damaging or potentially damaging stimuli into neural information. Nociceptors are found in most organs covering skin, muscle, joints, viscera, dental pulp, and dura. There are two main types of nociceptors, i.e. mechanical ($A\delta$) and polymodal nociceptors (C). As the name indicates, mechanical nociceptors respond to mechanical stimuli and polymodal receptors to different noxious stimuli such as mechanical, thermal, and chemical (Westlund and Willis, 2012). Mechanical nociceptors ($A\delta$) have myelinated axons and can transmit information at speeds ranging from 4 to 35 m/s. On the contrary, polymodal nociceptors (C) have unmyelinated axons resulting in a conduction speed slower than 2.5 m/s (Westlund and Willis, 2012).

The activation of nociceptors may trigger a number of responses named nocifensive responses or behaviours which range from autonomic reflexes to complex conscious pain responses (Westlund and Willis, 2012). Nocifensive and defensive behaviours are hierarchically organized in a series of nested and increasingly complex control loops which involve at least the periaqueductal gray, hypothalamus, stria terminalis, amygdala and ultimately the cortex (Blessing and Benarroch, 2012; Canteras, 2002).

Autonomic reflexes can be considered as part of the primary group of nocifensive responses, e.g. inflammation, activation of the immune system, endocrine changes, vocalizations and motor reflexes. For instance, responses related to noxious cutaneous stimuli, called *nociceptive withdrawal reflexes* or *nociceptive flexor withdrawal reflexes*, are designed to prevent or reduce tissue damage by eliciting fast motor responses (Gebhart and Schmidt, 2013, p. 2226), see Figure 2.2. On the other hand, autonomic reflexes associated with visceral systems are more general, e.g. changes in heart rate, blood pressure and respiration up to complex behavioural responses such as scratching. The particular type of responses associated with visceral noxious stimuli are also known as pseudo-affective responses, because they do resemble affective responses associated with painful stimuli but they are not able to prevent damage or eliminate the threat (Gebhart and Schmidt, 2013, p. 2277).

On the other hand, nocifensive responses such as avoidance, motivated and affective behaviour, memory formation and learning, are elicited by complex neural pathways. For example, nocifensive behaviours related to muscle and joint pain

are characterized by a decrease of force and joint use, and a decrease of the mechanical withdrawal threshold (Gebhart and Schmidt, 2013, pp. 2284-2289). The mechanisms involved in complex nocifensive behaviours rely on a sophisticated network of sensorimotor pathways.

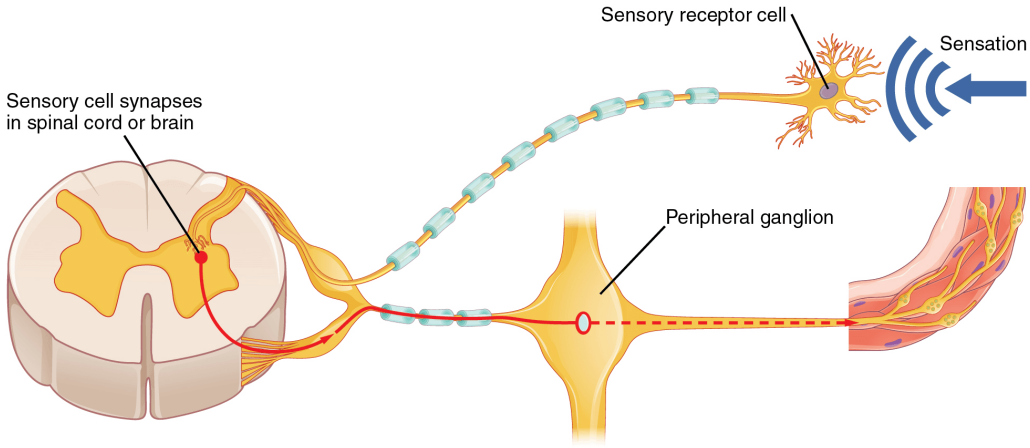


Figure 2.2 – Schematic representation of reflex circuits. Reflex circuits are integrated within the central nervous system; nonetheless, it is also possible that the sensor neuron alone triggers a reflex. By OpenStax College (2013, p. 635)[2].

There are a number of ascending pathways including the spinocervical, spinoreticular, spinomesencephalic, spinoparabrachial, spinohypothalamic and spinolimbic tracts. In humans, the spinothalamic tract (STT) is considered as the major ascending pathway of noxious stimuli. The spinothalamic tract is somatotopically and functionally organized contributing to the localization of pain on a body representation (Westlund and Willis, 2012).

Spinocervical, spinoreticular, and spinomesencephalic tracts modulate the perception of noxious stimuli. They also contribute to arousal, attention and regulation of autonomic and physiological responses to noxious stimuli, such as increases in respiration and heart rate and control of states of consciousness (Westlund and Willis, 2012).

Spinoparabrachial, spinohypothalamic, and spinolimbic tracts provide direct and indirect projections to the limbic system, and consequently they are involved in the modulation of autonomic and endocrine functions, localization of noxious stimuli, as well as autonomic control, motivational and affective responses, and learning and memory (Westlund and Willis, 2012).

There are also several cortical areas that respond to noxious stimuli. For instance, the somatosensory areas, particularly SI and SII, seem to be involved in the sensory discrimination of acute pain, the anterior cingulate gyrus is linked to the modulation

[2]Under Creative Commons Attribution License (CC-BY 3.0).

of affective responses to painful stimuli and the insula can be associated with the formation of pain memories (Westlund and Willis, 2012).

Noxious stimuli also activate descending pathways that decrease or accentuate pain perception. Evidence suggests that the activation of complex circuits at a supraspinal level inhibits nociceptive information travelling to higher areas in the brain, known as *the endogenous analgesia system*, while the descending projections of higher brain circuits facilitate nociception. The stimulation of brainstem circuitry involved in the inhibitory circuits can be used as pain relief in humans, a phenomenon named *stimulation-produced analgesia* which involves the periaqueductal gray, the nucleus raphe magnus, the ventrolateral medulla, and the dorsolateral pons (Westlund and Willis, 2012). The activation of opioid receptors involving both serotonergic and non-serotonergic neurons can also produce analgesia. Under stressful conditions pain modulation is achieved by the activity of serotonergic neurons while non-serotonergic neuronal activity modulates pain in unstressful conditions (Hornung, 2012).

2.2. Brainstem and Diencephalon

2.2.1. The Brainstem

The brainstem is located above the spinal cord and it connects the spinal cord to the brain. Adjacent to the brainstem is the cerebellum and together with the brainstem, they form the hindbrain (Purves et al., 2012, p. 720). The brainstem is divided into three main regions: medulla, pons and midbrain, see Figure 2.3. Additional to its unquestionable role as an information relay due to its key location, the brainstem is involved in a large number of regulatory functions including those of the vestibular, visual, auditory, and somatosensory system, motor cranial nerves and autonomic functions. The regulatory influence of the brainstem on most of these functions is the result of the collective action of multiple nuclei (Paxinos et al., 2012).

There are three characteristic functions of the brainstem:

- Firstly, the brainstem is essential for the coordination of autonomic regulatory functions such as visceral, cardiovascular, respiratory and thermoregulatory functions (Blessing and Benarroch, 2012).
- Secondly, it innervates the head and neck areas providing low and middle order complexity control, i.e. it supports sensory-motor reflex arcs in the head and neck, as well as more sophisticated muscle coordination like those involved in swallowing (Blessing and Benarroch, 2012).

- Finally, it is associated with the regulation of states of consciousness and sleep thanks to its dense projections to and from the forebrain (Purves et al., 2012, p. 722).

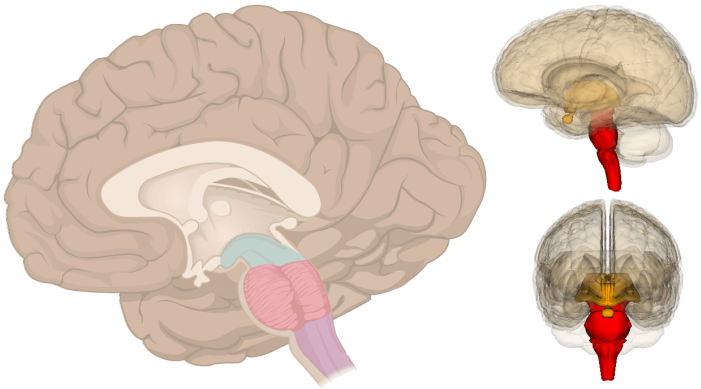


Figure 2.3 – The brainstem consists of the medulla (purple, first from the bottom), pons (pink, second from the bottom) and midbrain (cyan, third from the bottom). Trivially modified, original by Life Science Databases (LSDB)[3], Mitsuhashi et al. (2009)[4] and OpenStax College (2013, p. 536)[5].

The medulla or medulla oblongata is located immediately above the spinal cord and extends up to the lower border of the pons. The medulla relays somatosensory and proprioceptive information (Paxinos et al., 2012; Wilkinson, 1992a). It is also involved in the regulation of autonomic reflexes and of the parasympathetic and sympathetic divisions of the autonomic nervous systems. For instance, the baroreceptor-vasomotor and cardiomotor reflexes, thermoregulatory and respiratory functions, salivation, swallowing, gastric and intestinal function and vomiting are all functions regulated by the medulla (Blessing and Benarroch, 2012).

Above the medulla is found the pons. Both the pons and the medulla regulate cardiovascular and respiratory systems, among other vital functions (OpenStax College, 2013). The main components of the pons are the tegmentum, the cranial nerve nuclei, the locus coeruleus and the raphe nuclei. The pons interconnects the vestibular and oculogyric nuclei and thus it is associated with the coordination of head, eyes and neck movements. It is involved in visceral, proprioceptive, olfactory and auditory pathways. Projections from the pons to the limbic system are also noticeable. The locus coeruleus stands out by the release of noradrenaline (norepinephrine) and its extensive projections throughout the brain.

The raphe nuclei is composed of a mixed population of neurons. Interestingly, most of the serotonergic neurons (5-HT) in the brain are located within the

[3] `http://lifesciencedb.jp/`
[4] Under Creative Commons Attribution-Share Alike 2.1 Japan License (CC-BY-SA-2.1).
[5] Under Creative Commons Attribution License (CC-BY 3.0).

raphe nuclei (Hornung, 2012). Descending projections from the raphe nuclei are involved in the inhibition of pain transmission at the primary synapses level. Furthermore, the analgesic effects of the activation of opioid receptors depend on serotonergic neurons. Ascending projections from the raphe are involved in the regulation of circadian cycles (sleep-wake cycle) (Hornung, 2012; Wilkinson, 1992a). The activity of serotonergic neurons is the highest whilst being waking and completely inactive during REM sleep (Hornung, 2012; Wilkinson, 1992a). Neural activity in the raphe dramatically increases above its baseline activity during complex and repetitive motor behaviour such as chewing and running, and also during induced hyperpnea (abnormal laboured breathing). Finally, serotonin has also been indirectly implicated in the modulation of a variety of affective and cognitive responses such as the regulation of aggressive behaviour and drug addiction (Hornung, 2012).

The midbrain or mesencephalon is the shortest region of the brainstem and connects the hindbrain to the rest of the brain. The midbrain consists of the crura cerebri or cerebral peduncle, the substantia nigra, the ventral tegmentum, among other nuclei (Wilkinson, 1992a). The midbrain is often considered part of the meso-limbic system for the profuse projections from the ventral tegmental area (VTA) to the nucleus accumbens, the amygdala, the septum and the ventral striatum (Halliday et al., 2012). The projections from the midbrain to the limbic system are predominantly dopaminergic, approx. $65 - 85\%$, and most of the remaining projections are GABAergic. Another important characteristic is the fact that most of the dopaminergic neurons, in rat brains as much as $75 - 90\%$, are concentrated in only three regions of the midbrain, namely the substantia nigra, the ventral tegmental area and the retrorubral field. Dopaminergic neurons in the midbrain project to different regions including the putamen, the limbic system, and various regions of the cortex. Dopamine is a key neurotransmitter involved in reinforcement learning, most typically is reward-driven learning (e.g. Schultz, 1998) but also it has been implicated in avoiding punishment (e.g. Boureau and Dayan, 2011), working memory and motivated behaviour, along with other motor and cognitive functions (Halliday et al., 2012).

The substantia nigra (SN) separates the crus cerebri from the tegmentum and projects to the corpus striatum. Both the crura cerebri and the substantia nigra tightly operate with the corpus striatum in the modulation of motor activity. The three regions in conjunction are often called *extrapyramidal nuclei* (Wilkinson, 1992a). The darker appearance of the substantia nigra is due to the release of neuromelanin during dopamine metabolism. Degeneration of dopaminergic neurons in the substantia nigra is associated with Parkinson's disease (Wilkinson, 1992a). The SN also contains GABAergic neurons which constitute the major information relay from the basal ganglia to the thalamus, colliculi, and tegmentum (Halliday et al., 2012).

The main subdivisions of the tegmentum are the ventral tegmental area (VTA) and the periaqueductal gray (PAG). The VTA is better known for the dopaminergic circuit formed with the limbic system and its fundamental implications in arousal, stress, motivation, drug addiction, memory retrieval and reinforcement dynamics. The VTA merges with the retrorubral field (RRF) and both are crucial for the mentioned dopaminergic system (Halliday et al., 2012). The VTA and RRF consist of a variety of loosely arranged cells. Approximately 50% of these cells are dopaminergic in the VTA and a lower portion in the RRF. Only 50% of the dopaminergic neurons in the VTA and RRF have the neuromelanin pigment found in SN neurons. Dopaminergic cells in the VTA and RRF are smaller and due to a reduced capacity to reuptake dopamine are also less vulnerable to neurodegeneration than their SN counterparts (Halliday et al., 2012). The VTA also has a dense population of cholinergic neurons that project to the limbic system (Geula and Mesulam, 2012; Wilkinson, 1992a). A reduced number of serotonergic neurons are also present in the VTA. Finally, there are GABAergic cells in the VTA, in some species up to $15 - 20\%$, which are believed to inhibit the activity of dopaminergic neurons within the VTA via collateral projections (Halliday et al., 2012). A variety of dense and reciprocal projections have been identified in the VTA, for instance downstream with the medulla (locus coeruleus and raphe nuclei) and upstream with several different structures such as hypothalamus, accumbens nucleus, amygdala and hippocampus. On the other hand, the RRF has a number of non-dopaminergic projections to the hippocampus and cortex which are believed to influence declarative and spatial memory (Halliday et al., 2012).

The periaqueductal gray (PAG) contributes to numerous behavioural responses including pain modulation, cardiovascular regulation, non-verbal vocalization and basic defensive behaviour (Carrive and Morgan, 2012; Paxinos et al., 2012). The PAG is the place of intersection of the ascending nociceptive pathway and the descending motor pathway from the limbic system, suggesting that the PAG enables cross modulation of both systems. In fact, most of the afferent projections to the PAG come from the medulla oblongata, the limbic system, and cortical areas, e.g. the hypothalamus, the amygdala (central and basolateral nucleus), the prefrontal cortex, the insular cortex, and the temporal cortex. Interestingly, the PAG does not project directly to the limbic system, but it has dense projections to the regions in the thalamus that project to other areas within the limbic system and the cortex, specifically the amygdala, the basal ganglia, and the prefrontal cortex (Carrive and Morgan, 2012). The PAG is heavily and reciprocally connected to the hypothalamus and together they are believed to be crucial in the control of autonomic and behavioural responses. It is also heavily and reciprocally connected to the medulla and it has other reciprocal projections to the pons, the locus coeruleus and the parabrachial nuclei. These projections within the brainstem's premotor centres suggest that the PAG is involved in the integration of basic behavioural responses (Carrive and Morgan, 2012).

As mentioned above, the PAG seems to enable limbic modulation of pain perception, however, the role of the PAG appears to be broader and it is thus included as part of the endogenous analgesia system. Stimulation of particular areas of the PAG causes stimulation-produced analgesia, although accompanied with severe secondary effects (Carrive and Morgan, 2012). Stimulation of the PAG can also elicit a range of passive and active defensive responses, including increase in blood pressure, heart rate and respiration, shift in blood flow from the viscera to hindlimb muscles, explosive running and jumping, vocalizations, among others. The PAG also participates in the control of non-verbal vocalizations and modulation of neocortical verbal expression in terms of loudness, pitch, intonation, and rhythm which are strongly influenced by the limbic system and affect (Carrive and Morgan, 2012).

2.2.2. The Diencephalon

The diencephalon is a structure located above the brainstem which conserves the name given during embryologic development, see Figure 2.4. Both afferent and efferent information to the brain travels through the diencephalon, with the single exception of the olfactory information which reaches the cortex via the olfactory bulb, then the glomeruli and finally the olfactory cortex (OpenStax College, 2013). The diencephalon is divided into the subthalamus, the hypothalamus, the thalamus, and the epithalamus, which are roughly organized from bottom to top. The subthalamus is a transitional zone between the midbrain and the diencephalon. It is also connected to the globus pallidus of the striatum. The hypothalamus is located just above the brainstem and it merges into the tegmentum. The thalamus is the largest component of the diencephalon. It is an egg-shaped mass of grey matter located slightly above the hypothalamus. The epithalamus is located above the thalamus connecting the diencephalon to the limbic system and basal ganglia. The epithalamus consists of the habenulae and pineal gland (Wilkinson, 1992b).

The subthalamus is an elongated biconvex structure located between the midbrain and diencephalon. It is implicated in the inhibition of the globus pallidus from the striatum. The globus pallidus projects to the thalamus and the substantia nigra and it is involved in control of voluntary movements (Wilkinson, 1992a).

The hypothalamus lies just above the midbrain and slightly below the thalamus. It is responsible for maintaining homeostasis balance by monitoring circulating hormones and metabolites and organizing physiological and behavioural responses accordingly (Sternson, 2013). It is also involved in the modulation of defensive mechanisms, the alimentary system and reproductive behaviours including finding a mate, building shelter and taking care of offspring (Wilkinson, 1992a). These mechanisms have been considered as early forms of motivated behaviour (Sternson,

2013). Furthermore, it can be said that the hypothalamus is essential for the survival of the individual and perpetuation of the species (Canteras, 2002).

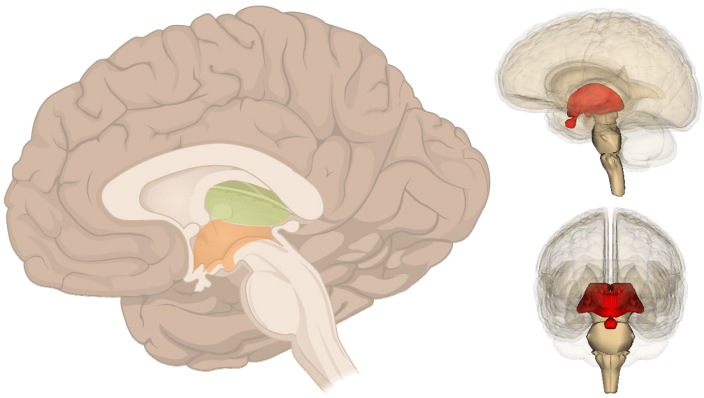


Figure 2.4 – The diencephalon located just above the brainstem. Its two main components, the hypothalamus (orange, left) and the thalamus (green, right). Trivially modified, original by Life Science Databases (LSDB)[6], Mitsuhashi et al. (2009)[7] and OpenStax College (2013, p. 535)[8].

The hypothalamus affects the endocrine systems through projections to the pituitary gland. It also influences both the sympathetic and parasympathetic divisions of the autonomic nervous system, specifically, influencing thermoregulation. For instance, states of hyperthermia causes vasodilation, sweating and lowered metabolism. Homologically, hypothermia causes vasoconstriction, shivering and increased thyroid activity (Wilkinson, 1992a). The hypothalamus is considered a *biological clock* for its involvement in circadian rhythm and other cyclical patterns; this role directly depends on environmental luminosity (Wilkinson, 1992a). In addition, the retinal afferents to the hypothalamus may also contribute to the strategic selection of defensive mechanisms, for instance, the decision between fleeing or freezing may depend on the levels of visibility, e.g. under conditions of low visibility a noisy escape may reveal the current position. By contrary, freezing reduces noise generation and favours mimicry and camouflage (Canteras, 2002). Maintenance of energetic and nutritional levels depends on the activity of specialized hormone- and nutrient-sensing neurons in the hypothalamus. These neurons respond to slowly varying levels of hormones and metabolites that signal energetic and nutritional demands and subsequently elicit eating behaviour. The mammillary body located in the medial part of the hypothalamus appears to be involved in spatial working memory and navigation. It has been argued that the mammillary body in association with the ventral tegmental nucleus is involved in the selection of appropriate defensive strategies (Canteras, 2002).

[6] `http://lifesciencedb.jp/`
[7] Under Creative Commons Attribution-Share Alike 2.1 Japan License (CC-BY-SA-2.1).
[8] Under Creative Commons Attribution License (CC-BY 3.0).

Both bodily and behavioural responses attributed to affective states regulated by the limbic system and prefrontal cortex are expressed via the hypothalamus, e.g. blushing, sweating, increase in blood pressure and pulse rate, defensive and sexual behaviour (Wilkinson, 1992a). Afferent inputs from the limbic system originate predominantly from the basolateral complex of the amygdala, the lateral septal nucleus and the bed nuclei of the stria terminalis (BNST), but it also receives input from the prefrontal cortex. In line with the hierarchical organization of defensive behaviour mentioned earlier, the expressions of affective and defensive responses elicited by the hypothalamus are coordinated and goal-directed and not explosive as the ones elicited by the PAG (Canteras, 2002). Stimulation of particular regions in the hypothalamus elicits a wide range of complex goal-directed drives such as intense eating, drinking, aggressive, and sexual behaviours as well as a number of somatomotor and autonomic responses resembling innate defensive responses. Similarly, lesions in the hypothalamus can prevent the expression of defensive responses. Hence, the hypothalamus is not simply a mediator of the limbic system and the cortex but it plays an active role in the generation of motivated behaviour (Canteras, 2002; Sternson, 2013).

The thalamus is the largest part of the diencephalon followed by the hypothalamus. It is an egg-shaped mass of grey matter next to the striatum. The thalamus is reciprocally connected with the cortex and it receives information regarding all sensory modalities including somatosensory, auditory and visual information though not olfactory information which is processed independently (Wilkinson, 1992a). Contrary to earlier beliefs, the thalamus does not simply relay information to the cortex and limbic system, but it also has specialized circuits to coarsely integrate and process such information. Furthermore, the thalamus and not the amygdala has been associated with the detection of biologically significant stimuli, e.g. visual and auditory ones (Pessoa and Adolphs, 2010; Weinberger, 2011). The thalamus can be subdivided into motor and sensory functional areas (Wilkinson, 1992a).

The motor thalamus consists of the ventrolateral nucleus (VL) and the ventral anterior nucleus (VA) which receive dense subcortical projections, i.e. mainly from the basal ganglia and the cerebellum (Mai and Forutan, 2012; Wilkinson, 1992a). The VL seems to contain topographical representations of different body parts and appears to be involved in the coordination of complex motor behaviours such as balance and fine motor skills. The VA is part of various recursive and parallel motor control circuits formed primarily between the associative cortex, the limbic system, the oculomotor, and other sensorimotor systems. Consequently, the VA is associated with the initiation, organization and control of voluntary movements (Mai and Forutan, 2012). Projections from the substantia nigra and globus pallidus modulate (activates or suppresses) the activity in the VA, particularly, the substantia nigra seems to influence movements of the head, neck, and eye, while the globus pallidus

seems to affect sophisticated gross motor behaviours (Mai and Forutan, 2012).

The sensory thalamus consists of numerous nuclei including the pulvinar and the geniculate body, among many others. The pulvinar is one of the largest regions in the thalamus, particularly in primates, where the human pulvinar makes up to 30% of the thalamus. The pulvinar is primarily involved in the processing and relay of visual information, especially to visuo-spatial areas, visual abstraction and attention, and is to a lesser degree involved in the processing of somatosensory, and multisensory information (Mai and Forutan, 2012). The medial part of the geniculate body is involved in the auditory pathway while the lateral portion relays visual information from the optic tract to the primary visual cortex. The ventroposterior complex projects to the primary sensory area in the cortex. The dorsal medial nucleus integrates somatic and visceral ascending projections and projects to the associative areas in the prefrontal cortex. The dorsal tier of lateral nuclei is involved in the analysis and integration of sensory information. It has reciprocal projections with the associative areas in the cortex. The intralaminar nuclei have nociceptive afferents and are involved in autonomic responses to visceral pain (Mai and Forutan, 2012).

The epithalamus lies above the thalamus and includes the habenular nuclei, posterior commissure and pineal gland. It receives projections from the hypothalamus, the olfactory system, the limbic systems via the stria terminalis, and the hippocampus via the fornix. The habenular nuclei projects to the midbrain creating a neural circuit by which the olfactory and affective system affect visceral responses such as salivation, gastrointestinal motility and secretion. Lesions in the habenular nucleus of the epithalamus affect metabolism, endocrine and thermal regulation. The pineal gland produces melatonin (derivative of serotonin). It is involved in the regulation of seasonal patterns such as mating and circadian rhythm (wake/sleep patterns) along with the hypothalamus (Wilkinson, 1992a).

2.3. The Limbic System

Although the term *limbic system* is broadly known and used, it is ill-defined. The existing definitions involve grouping a heterogeneous and varying number of brain structures that are hard to describe under a single criterion (Kötter and Meyer, 1992). The ever increasing quality of research tools have increased our understanding of the brainstem but it has not helped to improve the definition or to support the concept of the limbic system. Thus, many authors (Blessing, 1997; Blessing and Benarroch, 2012; Kötter and Meyer, 1992, among others) have suggested to rather focus on the individual role of different regions in the brainstem. Nevertheless, from a conceptual point of view, the concept of the limbic system is still very useful because of its simplicity, broad recognition and explanatory power

of "poorly understood brain functions" (Kötter and Meyer, 1992, p. 105). Or as indicated by Kötter and Meyer (1992, p. 124) "The term, however, is simple and enjoys universal recognition: everyone thinks he knows what is meant when he hears it".

The limbic system is generally described as a complex group of subcortical structures and their interconnections which includes the hypothalamus, the thalamus, the amygdala, the hippocampus and other neighbouring regions (Blessing, 1997). Figure 2.5 depicts a possible organization of the limbic system. It is an old system shared by all mammals. It is believed to be responsible for the modulation of non-motor aspects of the bodily homeostasis, i.e. autonomic regulation, modulation of mood, motivational and affective states and memory (Blessing, 1997; Mega et al., 1997). From the different roles attributed to the limbic system maybe the most commonly highlighted is its role in emotional behaviour and higher cognitive functions, in which the amygdala is crucial (Purves et al., 2012, p. 652).

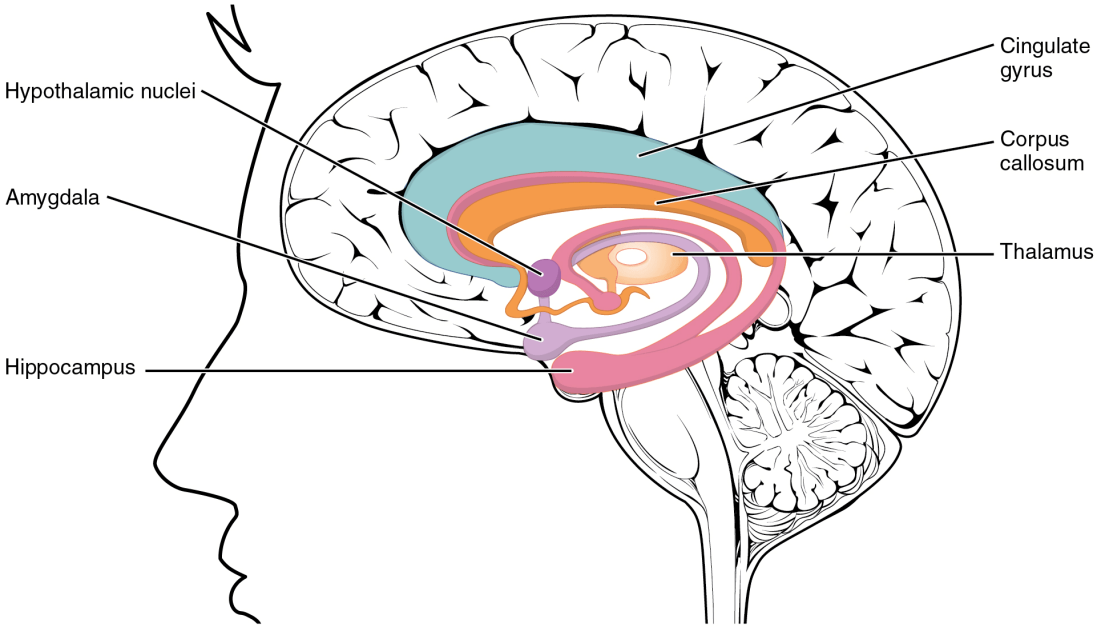


Figure 2.5 – Possible organization of the limbic system including the hypothalamus, the thalamus, the amygdala, the hippocampus, cingulate gyrus and corpus callosum. By OpenStax College (2013, p. 642)[9].

[9]Under Creative Commons Attribution License (CC-BY 3.0).

2.4. The Amygdala

The amygdala, also known as amygdalar complex, is an old brain structure located in the anterior and medial part of the limbic system, see Figure 2.6. It is present in both hemispheres and although response differences have been reported, they are often modelled as a single structure (Swanson and Petrovich, 1998). As is also true for many other brain structures, better instruments and methodologies have helped to reformulate the traditional definitions of the amygdala based primarily on shape, density, chemical signatures and other criteria (LeDoux, 2007; Pessoa, 2010; Swanson and Petrovich, 1998). This has led to an ongoing controversy about the neurophysiology of the amygdala, i.e. number, type of subdivisions and their relation to other brain structures.

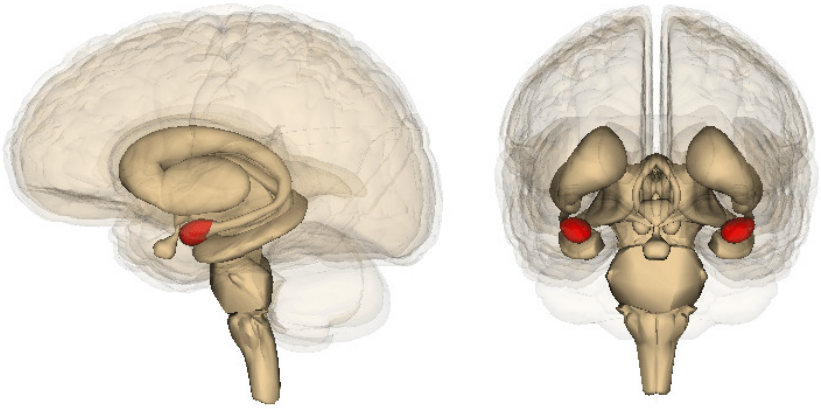


Figure 2.6 – The amygdala in the brain. By Life Science Databases (LSDB)[10] and Mitsuhashi et al. (2009)[11].

The amygdala has been persistently implicated in the processing of stimuli of biological relevance, affective modulation of behaviour and cognitive processes, e.g. homeostatic regulation, motion control, memory formation, behaviour on many reward-based decision-making tasks, among many other cognitive processes (Bechara et al., 1999; Fellous et al., 2002; Pape and Pare, 2010). The numerous and diverse processes affected by the amygdala can be explained by the rich connections existent from and to many neocortical and subcortical regions, including the thalamus, hypothalamus, hippocampus and cortex (Pape and Pare, 2010), see Figure 2.8 and Figure 2.9.

A substantial amount of evidence supports the idea that the amygdala acts as an early detection system for biologically significant stimuli (Belova et al., 2008; Germana, 1969; LeDoux, 2000; Whalen, 1998). The amygdala is particularly

[10] `http://lifesciencedb.jp/`

[11] Under Creative Commons Attribution-Share Alike 2.1 Japan License (CC-BY-SA-2.1).

sensitive to ambiguous stimuli, showing even higher activation than to known relevant stimuli (Pessoa, 2010; Whalen, 1998). Once a significant stimulus is detected, the amygdala is involved in the organization of bodily resources to gather additional information about the particular event in a process known as *affective attention* (Pessoa, 2010), but rather than trying to exactly identify the stimulus, the amygdala helps to prepare the body to quickly react to it, e.g. by increasing heart rate, hormonal levels, among other responses (Germana, 1969). During affective attention, sensory processing and memory consolidation are enhanced facilitating learning and memory retrieval of the significant cues (Anderson and Phelps, 2001; Pessoa, 2010).

Besides its well-known role in fear conditioning and in general threat evaluation, the amygdala also contributes to appetitive evaluation of stimuli such as the regulation of eating behaviours based on energetic and nutritional demands (LeDoux, 2012). Yet appetitive and aversive stimuli activate different circuits within the basolateral complex of the amygdala (Belova et al., 2008; Morrison and Salzman, 2010).

2.4.1. Anatomical Organization and Connectivity

Although the amygdala consists of numerous nuclei and cortical-like structures, it is simpler to think of it as two complexes namely the basolateral complex (BLA) and the central nucleus (CeA), as shown in Figure 2.7.

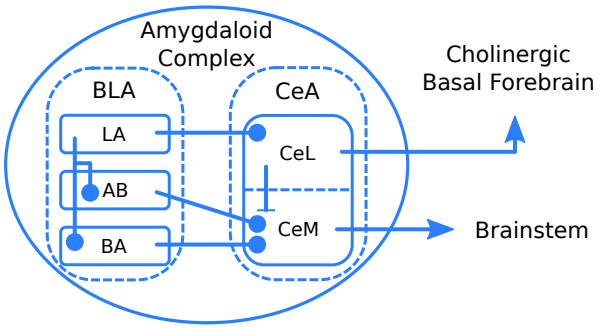


Figure 2.7 – Schematic overview of the principal connections within substructures of the amygdala. BLA, basolateral complex; AB, accessory basal nucleus; BA, basal nucleus; LA, lateral nucleus; CeA, central nucleus; CeL and CeM, lateral and medial part of the CeA, respectively. Based on Pape and Pare (2010) and Pape (2010).

The main nuclei in the BLA are the lateral (LA), basal (BA) and accessory basal (AB) nuclei. Even though the neurons in the BLA can be roughly grouped into two categories, i.e. glutamatergic and GABAergic, they are very diverse morphologically, electrophysiologically and neurochemically speaking, much like in the cortex. However, contrary to those in the cortex, the neurons in the BLA

are randomly oriented[12]. Neurons in the BLA are predominantly glutamatergic constituting approximately 80% of them. This group innervates not only the amygdala region but also other brain structures such as the striatum and some cortical areas. The remaining neurons are GABAergic, have short axons and thus only form local neuronal circuits (Pape and Pare, 2010).

The CeA consists mainly of GABAergic neurons, but in contrast to those in the BLA these neurons do not tend to form local circuits and rather project to different regions in the brain such as the cholinergic basal forebrain, the bed nucleus of the stria terminalis (BNST) and other structures in the brainstem (Pape, 2010; Pape and Pare, 2010).

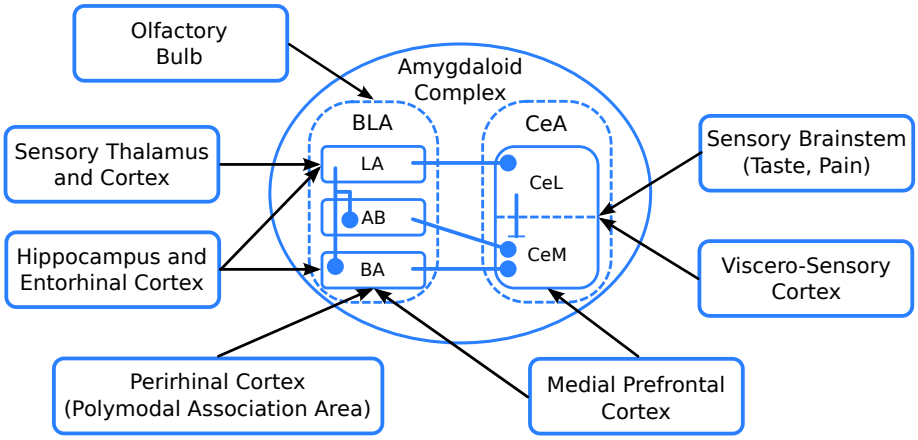


Figure 2.8 – Principal afferent projections to the amygdala. The amygdaloid complex is enclosed by an ellipse. BLA, basolateral complex; AB, accessory basal nucleus; BA, basal nucleus; LA, lateral nucleus; CeA, central nucleus; CeL and CeM, lateral and medial part of the CeA, respectively. Adapted from LeDoux (2007) and Pape and Pare (2010).

At a nuclei level, projections within the amygdala are chiefly unidirectional going from the BLA to the CeA. Even within the BLA, a feedforward-like organization can be distinguished; e.g. dense glutamatergic projections go from the LA to the BA, both LA and BA project to the CeA (Pape, 2010; Pape and Pare, 2010). Due to this organization, the BLA receives most of the sensory information reaching the amygdala, and consequently, the CeA is considered the main output and responsible for the modulation of autonomic and behavioural responses (Pape, 2010).

The amygdala receives multisensory information from both subcortical and cortical structures, but interestingly not directly from primary sensory areas, but rather from the associative cortex and multi-modal areas of the thalamus (Pape and Pare, 2010). This suggests and supports the idea that the amygdala is not responsible for the detection of neutral stimuli or the association of them with biologically significant stimuli (Pessoa and Adolphs, 2010; Weinberger, 2011), but

[12]Axons and dendrites are not oriented towards any specific direction.

rather to assign an affective value to them and modulate appropriate behavioural responses (Blessing, 1997; LeDoux, 2012; Murray, 2007).

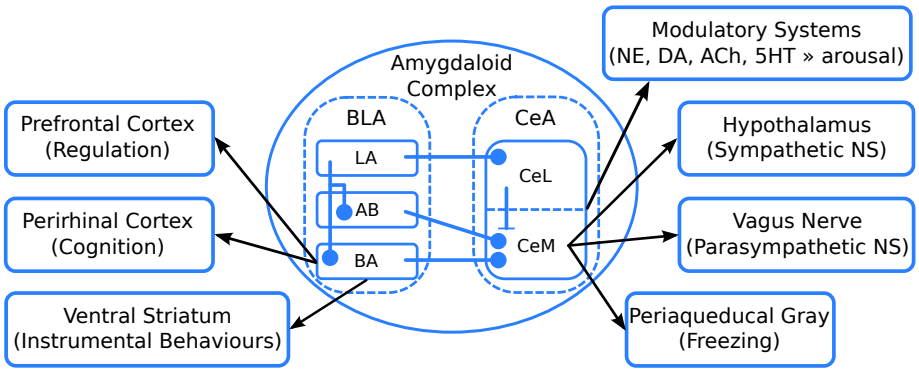


Figure 2.9 – Principal efferent projections from the amygdala. The amygdaloid complex is enclosed by an ellipse. BLA, basolateral complex; AB, accessory basal nucleus; BA, basal nucleus; LA, lateral nucleus; CeA, central nucleus; CeL and CeM, lateral and medial part of the CeA, respectively; NE, norepinephrine; DA, dopamine; ACh, acetylcholine; 5HT, serotonin; NS, nervous system. Adapted from LeDoux (2007) and Pape and Pare (2010).

2.4.2. Functions of the Amygdala

One of the most studied functions of the amygdala is its role in the evaluation and attribution of biological significance to neutral stimuli. This is achieved primarily via conditioning and allows animals to organize innate and learned behaviours in response to environmental changes. Neutral cues perceived along with biologically significant stimuli acquire a value equivalent to the innate significant stimuli, a process also known as *value encoding*. The selection of neutral stimuli that would acquire a biological significance strongly depends on its ambiguity and unpredictability, therefore it is tightly connected to the concept of *prediction error* (Pessoa, 2010).

While most of the knowledge about the amygdala has been obtained via fear conditioning studies, the aymgdala does not only encode information regarding aversive stimuli. On a more general scale, the amygdala, particularly the basolateral complex, seems to encode value intensity rather than valence (appetitive vs. aversive) of biologically relevant stimuli (Murray, 2007). During value encoding, the amygdala boosts memory formation and biases sensory processing contributing to long-lasting associations between neutral and significant stimuli (Pape, 2010). A closer look at the basolateral complex reveals its impact on memory formation in the caudate nucleus and the hippocampus by influencing the release of cortisol and adrenaline (Mcintyre et al., 2003; Paré, 2003).

Evidence from lesion studies indicates that value representations, i.e. learned associations between neutral and biologically significant stimuli, are not stored in the amygdala (Anderson and Phelps, 2001; Pessoa, 2010), because expression of fear responses is still possible when the amygdala is damaged. Nonetheless, lesions in the amygdala have a detrimental impact on both the acquisition of new fear memories (BLA lesions) and expression of fear (CeA lesions) (Pape and Pare, 2010).

Value encoding and the organization of innate and learned behaviours is a complicated process and requires the involvement of other systems such as the mesolimbic dopamine system, and the prefrontal cortex, among others (Pessoa, 2010). The intricate interrelation of these systems has far reaching consequences extending from adaptive survival circuits to attention, decision making and affective processing (LeDoux, 2000; Pessoa, 2010). The role of the amygdala in all these cognitive processes depends on the combination of three pieces of information primarily originating in subcortical regions (Morén, 2002, p. 27):

- information regarding the current homeostatic state of the organism, e.g. current energetic and nutritional demands.
- external information regarding stimuli of innate biological significance such as the distinctive odour of a predator.
- information regarding salient sensory cues that may help to anticipate the occurrence of stimuli with innate biological significance.

The role of the amygdala in adaptive survival circuits can be substantiated through descending projections of the central nucleus to the hypothalamus, the periaqueductal gray, the reticular formation and the pituitary gland, among others (Pape, 2010; Pessoa, 2010). Projections upstream are indirectly involved in certain aspects of vigilance, alertness and affective attention, more specifically in the aspects dealing with bottom-up competition under ambiguous or unexpected outcomes (Pessoa, 2010; Whalen, 1998). The role of the amygdala in these processes could be considered indirect, because it relies on its intricate connectivity with both subcortical and cortical structures. For example, the central nucleus via its dense projections to basal forebrain is involved in arousal, enhancement of sensory processing and active behavioural responses (Pape, 2010; Pessoa, 2010). Furthermore, the basolateral complex has been implicated in the modulation of awareness and perception (Pessoa, 2010).

From a behavioural point of view, activity in the central nucleus determines the magnitude and type of conditioned responses (Pape, 2010). For instance, active behavioural responses such as exploration and cortical arousal are linked to projections from the central nucleus to the cholinergic basal forebrain, whereas projections from the central nucleus to the brainstem are linked to passive behavioural

reactions such as freezing. It is also worth noting that they are expressed in a mutually exclusive fashion (Pape, 2010).

Cortical projections from the amygdala are also closely linked to psychological phenomena such as *discounting*. Discounting is the mechanism by which the perceived value of a reward decreases with time. The basolateral complex reduces discounting and thus helps to establish longer temporal relationship between decision and reward (Pessoa, 2010). Consequently, damage in the basolateral complex is associated with impulsiveness, risk-aversion and laziness (Pessoa, 2010).

The important role of the amygdala in fear conditioning dynamics has contributed to the misconception that the amygdala is involved or responsible for emotion processing (LeDoux, 2007). While the amygdala is crucial for affective information processing by associating a biological significance to neutral stimuli, it does not necessarily influence mood (Whalen, 1998). However, the process of value encoding is considered an essential building block in the development of cognition, emotions and the development of intelligence in general (Pessoa, 2010).

2.5. Conditioning

Animals need to be able to identify biologically significant stimuli in order to properly interact with their environment, e.g. home, food, partners, and predators. Certainly a large number of these evaluations correspond to or are directed by instinctive and subconscious processes. The amygdala is one of the main structures responsible for constantly monitoring internal and external states and promptly eliciting proper behaviours, see Section 2.4. Even though many of those responses are innate there are clearly not enough for self-preservation in complex living forms, consequently, most animals must be able to judge and adapt properly to new or unexpected situations during their lifetime. One of the main mechanisms of adaptation is conditioning. Conditioning is a type of associative learning based on rewards and punishments and constitutes the basis for reinforcement learning paradigms (Morén, 2002, pp. 44, 62).

During conditioning, learning meaningless stimuli, called *conditioned stimuli* (CS), are associated with biologically significant stimuli, called *unconditioned stimuli* (US). After the association, the CS will be able to produce the same responses that can the US alone. The amygdala is well known for its role in conditioning, especially in fear conditioning dynamics (LeDoux, 2007; Pessoa, 2010). As introduced in Section 2.4, when a biologically significant stimulus is detected, an enhanced sensory processing takes place. Here, the cues with more predictive power, i.e. ambiguous or unexpected, are selected for association with the US. Conditioning dynamics can be classified at least by the temporal relationship between CS and US and the

kind of behavioural responses elicited.

In terms of temporal separation of stimuli, conditioning paradigms are classified into delay and trace conditioning. In *delay conditioning*, the CS and the US may differ in their onset and offset, but they are present concurrently for at least a small fraction of time. In contrast, *trace conditioning* is characterized by the CS and the US being separated by a time gap, which makes it more difficult to learn (Morén, 2002, p. 72).

Depending on the type of responses elicited by the US, conditioning can be classified into Pavlovian and instrumental conditioning. *Pavlovian or classical conditioning* involves innate responses such as increase of the heart rate, salivation, etc. These responses prepare the agent's body to interact with the unconditioned stimulus. *Instrumental or operant conditioning* elicits behavioural responses to avoid or approach the unconditioned stimulus. With those behavioural responses the agent can, to some extent, influence the presentation of, or not, the US (Morén, 2002, p. 74). Instrumental conditioning can be seen as a two-step process. Firstly, classical conditioning associates a conditioned stimulus with the unconditioned stimulus. Secondly, the appropriate behavioural response is learned via instrumental conditioning (LeDoux, 2007; Morén, 2002, p. 77).

Furthermore, the discovery of the appropriate behavioural responses by instrumental conditioning is based on the perceived feedback or reinforcement. Primary reinforcers are the stimuli that trigger innate responses or satisfy survival needs such as food, water, sex and pain. Secondary reinforcers are stimuli that have acquired a biological significance during the animal's lifetime (LeDoux, 2012). The quality and speed of learning is affected by the frequency with which the reinforcer is presented. As a result, *reinforcement policies* can be classified into three core types which can be varied and combined to create further reinforcement schedules. If the reinforcer is consistently delivered every time the appropriate response is performed, then the policy is known as *continuous reinforcement*. Alternatively, if the reinforcer is presented only after a given time, then the policy is called *interval scheduling*. Finally, if the reinforcer is delivered only after the appropriate responses are performed a number of times, then it is named *ratio scheduling* (Morén, 2002, p. 76). As discussed before, ambiguous or unexpected events boost learning, leading to stronger and longer lasting associations (Pessoa, 2010).

Chapter 3
Biologically-Inspired Self-Preservative Mechanisms for Robots

Self-preservative mechanisms belong to the most important and essential capabilities of any organism. Mechanisms such as those involved in homeostatic, defensive, and sexual behaviours are fundamental to ensure individual survival and the perpetuation of the species (Canteras, 2002; LeDoux, 2012; Mirolli et al., 2010; Sternson, 2013). But living systems are fundamentally different from artificial systems, both internal and external differences in their embodiment lead to significant differences in their mechanics and computations capabilities (Arbib and Fellous, 2004), which raises the question of whether artificial systems require similar mechanisms.

Survival is dependent on efficient and timely satisfaction of a multitude of physiological deficits, thus through evolution, animals have developed and perfected a number of innate responses to comprehensively satisfy these needs (Mirolli et al., 2010; Prescott et al., 1999). However, a simple mapping between stimuli and reactions is not enough to guarantee self-preservation in an ever-changing environment. Consequently, adaptability becomes an essential component for survival. Vertebrates have developed common sub-cortical mechanisms, both neural and non-neural, that adjust and coordinate innate behaviours to cope with a dynamic environment (Krichmar, 2008; Ziemke and Lowe, 2009). These systems are hierarchically organized in increasingly sophisticated and nested control loops that can be independently activated by specific internal or external events (Blessing and Benarroch, 2012; Canteras, 2002). They modify pain thresholds, shift attentional effort, drive motivation and elicit goal-directed behaviours, thereby establishing the basis for cognitive processes such as planning, decision making, mood and emotions (Arbib and Fellous, 2004; Krichmar, 2008; Ziemke and Lowe, 2009). Moreover, a growing corpus of evidence shows the striking and intertwined relationship between effort for self-preservation such as homeostatic regulatory processes and evaluation of biologically significant stimuli, and development of intelligence (Arbib and Fellous, 2004; Parisi, 2004; Ziemke and Lowe, 2009).

However, as Ziemke and Lowe (2009) summarize, it is not that the brain and

cognition can simply be mounted on a body, nor that cognition is built on top of autonomic and homeostatic regulatory systems, but rather that cognition is deeply interwoven and inseparable from the concepts of brain, body and autonomic regulation. Furthermore, the current approach of studying neural phenomena in isolation from other neural circuits and neuromodulatory systems needs to shift towards a more holistic view of organisms. This shift will enable a new crucial step towards a better understanding of autonomous and cognitive systems (Ziemke and Lowe, 2009). Cognitive neuroscientific approaches are increasingly viewing cognition as fundamentally dependent upon embodiment and distributed neural circuitry (Adolphs, 2010).

Now, under the assumption that autonomy, cognition, and consciousness are indeed products of this deeply interwoven and inseparable interaction of self-sustained processes, if general purpose artificial systems are to become truly autonomous systems and act appropriately in highly dynamic environments, and co-exist with other autonomous systems and humans in a natural way, they will require a set of homologous mechanisms, i.e. systems that help them to be more efficient and allow them to adapt to changes in their surroundings and elicit suitable or successful behaviours. Therefore, the question changes to whether artificial systems need biologically inspired mechanisms for facilitating the development of self-preservative and artificial autonomous systems.

Undeniably, there have been and there still are many great advances to come in the field of artificial autonomous systems, both in terms of hardware and software. However, current commercial artificial autonomous systems are tailored solutions to specific tasks and they have little potential to become the general purpose robots envisioned by many. The reasons for this are diverse, but from a technical point of view, they can be summarized into two factors. Firstly, it is becoming increasingly difficult to foresee all possible variations of a scenario in which a particular robot may operate, in other words, to identify the frame or required world model, see the frame problem (McCarthy and Hayes, 1969). Secondly, the cost of designing, implementing, maintaining, upgrading and integrating the appropriate behaviour (control system) for the anticipated use cases with the rest of the system is prohibitive. On the other hand, the robustness, adaptability and effectiveness of biological agents has long been acknowledged, as well as, the weaknesses of current artificial autonomous systems. This has aided the development of new research directions motivated by the admirable evaluative and learning mechanisms of animals such as embodiment theory (Pfeifer and Bongard, 2006) and developmental robotics (Asada et al., 2001).

These new paradigms, as compared to loosely bio-inspired approaches common to the initial attempts outside the domain of the *Good Old Fashioned Artificial Intelligence* (GOFAI) (Haugeland, 1989, p. 112), take into consideration and, moreover, exploit the attributes and limitations of the system without being deceived

by its appearance. For instance, if an artificial system mimics characteristics like morphology, some learning rules and maybe other aspects of a biological agent, it does not imply that it will eventually develop any arbitrary behaviour of the emulated biological agent. The reason for this is that the underlying principles ruling those behaviours have not yet been discovered nor understood nor can they can be artificially emulated, at least with today's technology (Bovet, 2007, p. 5). Unfortunately, even under these more integrative research paradigms the focus is still biased towards the interactions of the system with the external environment. Ziemke and Lowe (2009, p. 105) describe them as "sensorimotor embodiment and the grounding of cognition in perception and action", and emphasize that these approaches neglect what happens in the 'internal' environment of the system by not having an organismic view of the system (Parisi, 2004).

One approach towards the development of truly artificial autonomous systems, i.e. those able to adapt in a timely manner to changing internal and external circumstances in an unsupervised fashion, would be to mimic not only morphological characteristics of biological agents, but also the building blocks that led to such sophisticated biological systems. For example, something along the lines of defining *needs* and the related innate *appetitive* and *aversive* stimuli, providing a minimal repertoire of built-in responses, the sensory capabilities to monitor the state of those needs and the mechanisms to organize, mix and develop new behaviour to fulfil them. These components may resemble the ones used in current approaches; however, the difference in the implementation is as subtle as it is fundamental. For instance, consider the case of a depleted battery; in a traditional approach the solution would be to explicitly design a recharging procedure, whereas the suggested approach is to attempt to ground the sensation of *being hungry* and the means to satisfy it. Ultimately, this approach would unavoidably force designers to give up their control over certain aspects of the system to eventually achieve truly autonomous artificial systems (Morén, 2002, p. 11).

As part of the European project *Integrating Cognition, Emotion and Autonomy* (ICEA)[1], Ziemke and Lowe (2009) developed a cognitive architecture for artificial autonomous systems profoundly based on principles like those introduced above, see Figure 3.1. Particularly, they based their architecture on the "architecture and physiology of the mammalian brain" (Ziemke and Lowe, 2009, p. 111) and tried to integrate autonomic/homeostatic, cognitive and emotional mechanisms, i.e. developed a hierarchically organized architecture of nested systems that control the basic aspects of the agent – such as autonomic/homeostatic processes – in which complex cognitive processes, like emotions, are naturally embedded and are an integral part of the system, and not simply built as independent modules. With this architecture Ziemke and Lowe (2009) highlight the self-organisational and interwoven nature existing between neural and non-neural autonomic/homeostatic

[1] `http://www.iceaproject.eu/`

regulatory mechanisms and sensorimotor activity that give rise to living autonomous systems, where no causal relationship between homeostatic regulation and cognition can be drawn.

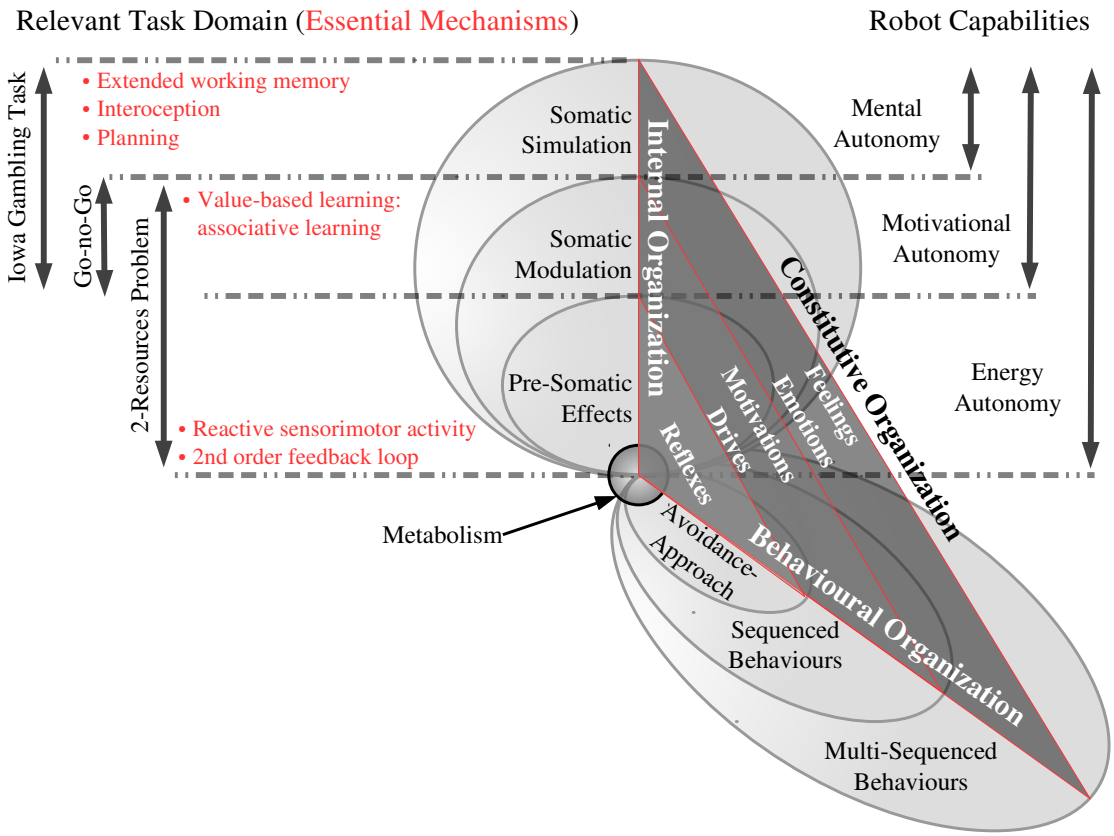


Figure 3.1 – Schematic representation of a cognitive architecture for the organization of autonomous behaviour. On the right-hand side are the potentially achievable robot capabilities based on the three core types of autonomy defined by McFarland (2009, p. 15) which are dependent upon the degree of constitutive organization. On the left-hand side are some essential mechanisms and potentially achievable behavioural tasks. Adapted from Ziemke and Lowe (2009).

This is still a descriptive model and as such it is not directly testable, but it elegantly summarizes and delineates the critical aspects that *cognitive architectures* for artificial systems may require to achieve autonomy. Certainly, the challenges to develop such architecture are numerous. However, survival needs and their regulatory processes – although specific to current pressing needs – have system wide implications, and they impose ground rules for more complex and motivated behaviours (Sternson, 2013). Therefore, they seem to be the right place to start from. Nevertheless, as Ziemke and Lowe (2009) indicate a few questions remain. For instance, what survival mechanisms present in biological systems are worthwhile transferring to artificial systems? Is there a minimal set of mechanisms that enable

autonomy? How much biological detail is needed in the models? Should these models be based on a particular living system?

From an engineering point of view, there is a list of highly desirable and convenient mechanisms that could be modelled. Arguably, energetic autonomy, damage prevention, and self-reparation are of special interest to robot applications. Advances on these areas would undoubtedly represent a step towards truly artificial self-preservative and autonomous systems, i.e. artificial systems that behave in accordance with relevant internal and external circumstances. Consequently, a main focus of research in this regard should aim at a better understanding of homeostatic processes and neural circuits involved in those mechanisms defined as *of particular interest* or *highly desirable*. Regarding brain areas, of particular interest are those areas with a high degree of connectivity, which either distribute (hubs) or aggregate (authorities) information, e.g. the thalamus, the basal ganglia and the prefrontal cortex (Modha and Singh, 2010). The thalamus, the basal ganglia and the amygdala may constitute the primary brain regions to model, because they are a homogeneous group of brain regions across mammals, disregarding the prefrontal cortex, which is both disproportionately large in primates, particularly humans, and the latest level of complexity in the organization of behaviour. With regard to autonomic, homeostatic and behavioural aspects, those involved in energetic autonomy and damage prevention may arguably be considered as of higher relevance.

Motivated by the above-mentioned concepts, the following sections of this chapter attempt to give an overview of current research on the understanding and development of artificial self-preservative and autonomous agents, primarily humanoid robots. The presentation of these studies will be roughly grouped and mapped to the different stages shown in Figure 3.1.

3.1. Energetic Autonomy

Although definitions of autonomy are very flexible, in the field of robotics and artificial intelligence this term is mostly used to describe agents that can perform a task without outside control (McFarland, 2009). Therefore, many levels or types of autonomy are possible. For instance, McFarland (2009, p. 15) defined three core types of autonomy: energy, motivation and mental autonomy.

Energetic self-sufficiency is an important and current problem for both research on autonomous artificial systems and engineered applications. At first glance, this concept may seem trivial. However, even for technically driven solutions it is still a challenge. From a technical point of view, mobile robots can perform a finite amount of work in a single charge cycle which is determined by the capacity of their batteries. An increasing number of tasks that robots are expected to perform,

and the steadily growing number of 'energy hungry' sensors and actuators that are required to perform these tasks, exacerbate this problem. Therefore, robots should plan their actions taking into consideration the available energy and schedule recharging to maximize operational efficiency. Furthermore, from the point of view of autonomy, robots should be able to obtain their required energy directly from their environment without the need of human intervention. Endowing robots with these capabilities will contribute to the development of truly autonomous robots and permit the use of robots in inhospitable or remote environments.

3.1.1. Homeostatic and Metabolic Energy Management

Animals forage, eat and digest organic matter to obtain the energy and nutrients they need to subsist. In the same sense do robots require a source of energy to function. The skills of animals of extracting energy directly from their surroundings have motivated the development of a new class of robots known as Symbots or symbiotic robots (Melhuish et al., 2006). Symbots are robotic systems that use microbial fuel cells (MFCs) as an energy source[2], thus capturing both metabolic and behavioural aspects of energy management of living systems (Ieropoulos et al., 2010, 2003, 2005). The robot is forced to maintain a metabolic equilibrium in order to obtain the energy required to function, specifically, the robot needs to keep the internal temperature, pH and water level within an ideal range to keep the microorganisms in the fuel cells alive and healthy. Current developments use open loop systems much like animal digestive systems, i.e. they need to forage fresh water and organic matter, and eliminate the by-products or indigestible matter. The used microorganisms may withstand harsh environmental conditions which allows for great flexibility in the refuelling capacities of the robot as it can process raw substrate of the type indigestible to biological agents and thus may feed in a highly opportunistic manner.

The series of EcoBots (Ieropoulos et al., 2010, 2003, 2005; Melhuish et al., 2006) are one of the earliest examples of Symbots and are still under active development. Limitations in current microbial fuel cells technology impose some additional constraints. For instance, they are not capable of producing enough energy to sustain a continuous operation of standard motors and sensors, thus EcoBots show a 'pulsed behaviour', i.e. robots operate intermittently with short 'awake-like' stages and longer recovery periods. Other constraints originate from the kind of microorganisms in the system that may produce some kind of food preference or selectivity. Finally, the 'food' needs to be pre-processed before it is fed to the fuel cells.

[2]The use of fuel cells is an environmentally friendly source of energy for robots that can operate with organic waste and could even be used to process polluted water.

Experiments with EcoBots provide strong evidence that simple threshold indicators of energy levels are not enough to achieve energetic autonomy (Ieropoulos et al., 2003), because symbots require a wide range of complementary systems to work. Many of these systems are required for food intake and waste disposal, others are needed to monitor and control temperature, pH and water levels. Naturally, most of these complementary systems require energy and have to be effectively and efficiently organized alongside foraging behaviours and potentially other 'higher' cognitive behaviours. It can be speculated that a rather more suitable energy management system would learn to adapt to the dynamics of the bio-electrochemical processes that produce the energy and processes that use the energy, firstly, because natural fluctuations in energy production are inherent to the nutrient concentration and waste accumulation in the fuel cells, and secondly, because energy consumption of sensors and actuators changes under different work regimens. The heterogeneity of the time constants of these processes may become an interesting challenge to the development of such artificial energy management systems. The hypothalamus is crucial for energy management in mammals and effectively integrates fast changing neural information with slow changing hormonal information via the endocrine system, see Section 2.2.2.

Another interesting future development originating from MFC powered robots would be the incorporation of 'artificial immune systems' (Neal et al., 2006) that could look after the well-being of the microorganisms in the fuel cells. An effective artificial immune system would be able to detect and, whenever possible, tackle issues locally without the need of high-level software and control systems (Neal et al., 2006). For instance, issues such as temperature, pH or water level regulation could be under the control of an artificial immune system. In natural systems, the sustained activation of the innate immune system elicits a 'stress response' that has system-wide implications affecting physiological, behavioural, and psychological states (Cummins, 2012; Neal et al., 2006). Similarly, a stress response could be used to draw attention to exceptional events that a low-level artificial immune system can not resolve.

But as Ziemke and Lowe (2009) indicate, it still remains to be seen how much biological detail would be necessary or useful for the development of artificial autonomous systems. Furthermore, regarding energetic autonomy different questions arise, e.g. whether fuel cells will become the most important source of energy for robots or whether more diverse energy sources, such as fuel cells, solar cells, and batteries, with a flexible recharging strategy are required or some hybrid thereof.

3.1.2. Recharging and Goal-Driven Behaviours

Regardless of the source of energy, robots will likely still require recharging behaviours, either by foraging organic matter or seeking solar radiation or a dedicated recharging station or connecting directly to a standard power outlet. For domestic service robots, energetic autonomy is paramount, because the user should not face the burden of replacing the robot batteries or plugging it to a power outlet. The autonomous recharging problem of robots is a difficult and unsolved challenge that involves the effective coordination between task execution, recharging, idle time, localization and navigation. The required hardware for recharging has to be specifically designed to match task, scenario and robot embodiment, which adds a greater level of complexity. These two aspects, hardware and behavioural, are ideally tackled together within one solution. Unfortunately, the development of complete solutions is in most cases infeasible, and the development of complementary hardware for existing robotic platforms is greatly hindered for lack of information and license issues, thus solutions tackling the behavioural aspects of energy autonomy predominate the robotic literature.

One attractive technical approach consists of an especially designed battery bay, which may permit autonomous battery swapping (Suzuki et al., 2012; Zhang et al., 2013). Besides being a quick way to 'refuel' a robot, it solves the problem of battery end of life that other approaches such as wireless recharging do not solve. Unfortunately, battery swapping requires tailored hardware; moreover it is not clear how the robot could be kept functional while the swapping is carried out. For instance, should the robot use two batteries or have an empty space to accommodate simultaneously both the depleted and the charged battery. Besides, complementary systems may be required such as solar cells or even the capability to recognize and plug itself to a standard power outlet. Nevertheless, although autonomous battery swapping does not provide a complete solution, it represents an important step in the right direction with respect to the hardware aspect of energetic autonomy.

Ideally, the behavioural aspect of autonomous recharging requires the robot's awareness of consumption, correlating task execution and energy usage to decide when, where and for how long to recharge to provide a reasonable service time. This requires efficient coordination of different tasks, e.g. task execution, recharging or battery swapping, idle time, etc. To navigate the robot to the charging station requires estimation of its position. Occlusion of recharging location, slipping and skidding effects, positioning error, manoeuvring speed and jerk make the precise docking to a charging station a difficult problem. Precise navigation is of general interest. It can reduce jerk and thus make robots safer for navigation reducing the risk of balancing and falling over, and by this potentially hurting the user, or damaging the robot or the surrounding environment. Precise navigation or docking

is also relevant for navigation in narrow spaces common in home environments (Ren et al., 2012).

Although, many dedicated pre-programmed autonomous recharging solutions exist, they are not ideal, because they cannot adapt or easily accommodate other behaviours or integrate with other systems, in other words, they have to be learned and not hard-coded. Many design considerations remain to be addressed such as how to design robots that can operate for long periods of time without human intervention? How would they know when it is the best time to 'refuel'? Also, the developmental aspects of these problems still need to be studied: how can the feeling of being hungry be grounded rather than being a simple collection of thresholds? Learned and developmental solutions are difficult to develop. Our current strategies create an interesting 'chicken-egg' dilemma, i.e. the agent needs to learn to survive, but most learning approaches rely on experiences to learn, therefore, may place the agent in harmful situations unnecessarily and even cause irreparable damage. Clearly, learning is needed but a number of other mechanisms also need to exist to keep the agent *alive* while it learns to survive more efficiently. A set of hard-coded mechanisms is needed; we could look at innate responses and find an equivalent set of mechanisms for robots.

3.2. Damage Prevention

Current robot platforms have a vast number of sensors and actuators meant to fulfil the growing number of tasks expected from general purpose robots but, surprisingly, most of them do not incorporate adaptive mechanisms of self-protection, if any at all, either for human safety or robot self-protection. So far, the most common approach has been to physically separate the robot's workspace in time and space from the human's workspace to ensure safe operation in industrial applications. However, the inexorable need for robots to coexist with humans in order to tackle a broad and steadily growing number of tasks, such as helping the elderly (KSERA), inevitably requires a different paradigm. The initial attempts focus mainly on limiting robot size, weight and power[3], and increasing robot compliance[4]. Although all these measures are important for safety, they are engineered for pre-defined situations or based on reflexes with little adaptation or learning. Unfortunately, these solutions are not enough to ensure human, infrastructure, and robot safety both inside and outside the laboratory.

In recent years, there has been a growing interest in the development and design of safety mechanisms and standards for human-robot interaction, mainly focusing

[3]Nao, Paro, Roomba
[4]Baxter, Mekabot, REEM-C

on human safety (Harper and Virk, 2010; Murphy and Woods, 2009). However, it is important to highlight that self-protective mechanisms could not only help to protect robots, and thus contribute to the development of more autonomous artificial systems, but also and more importantly, they could constitute an additional safety layer for service robot applications. For instance, sharp edges or deformations on the 'robot's skin' resulting from accidental collisions or natural material gear could easily lead to accidents posing imminent danger to the users. These are the type of events where biologically inspired mechanisms such as pain, autonomic reflexes and fear could be used to enhance robot autonomy, reducing maintenance costs and preventing accidents in the long run.

3.2.1. Pain Modelling

When referring to self-protective mechanisms, pain is one of the first concepts that comes to mind, rightfully so, because pain perception, a complex psychological and neurophysiological mechanism, developed to protect the body from injury. The main characteristics of pain are that it elicits immediate autonomic responses, attracts attention and can have long-lasting behavioural repercussions such as chronic pain. However, in biology, pain is always considered as a subjective experience that goes far beyond the sole activation of pain receptors and thus it can be argued that robots are incapable of perceiving pain. Nevertheless, the mechanisms involved in pain management, e.g. physiological and autonomic responses, attention and learning, may be modelled and taken as guidance to build safer robots, as it will be exemplified in this section.

There are many stimuli and events that pose a threat to the well-being of living creatures; homologically, there are numerous conditions that pose a threat to the functioning of robots and artificial systems. The identification of homeostatic-like conditions for robots could be accomplished by looking at the operational limits of the robot's components, such as temperature, pressure, acceleration, humidity, etc. and developing around these variables/components biologically inspired regulatory mechanisms to ensure the robot's proper functioning and the user's safety.

In order to design self-protective mechanisms based on pain-like signals, it is necessary to establish a model of pain perception. Towards this goal, Kawaji and colleagues (Akayama et al., 2006; Matsunaga et al., 2008, 2012) have developed detailed models of pain perception produced by mechanical stimuli. Kawaji et al. focused on touch, pressure and brief impacts on skin of human upper limbs, i.e. fingertips and arms. They considered how physiological aspects of the human skin, such as elasticity, sensitivity and nociceptor distribution, influence pain perception in terms of intensity and duration. Pain perception triggered by mechanical stimuli can produce fast and slow pain. Fast pain is primarily attributed to the activation of

superficial mechanical nociceptors ($A\delta$) that produce an intense and localized pain. Slow pain is attributed to the activation of subcutaneous polymodal nociceptors (C) that produce a diffuse and longer lasting pain. High-frequency mechanical stimuli are attenuated by the most superficial regions of the skin and activate nociceptors of the $A\delta$-type, whereas low-frequency stimuli can easily activate nociceptors of the C-type. The amplitude of the stimuli also influences pain perception. For instance, high-frequency mechanical stimuli that also have high amplitude can elicit both fast and slow pain. On the other hand, high amplitude low frequency stimuli typically produce only slow pain (Matsunaga et al., 2008, 2012).

To capture all these features, Kawaji and colleagues (Akayama et al., 2006; Matsunaga et al., 2008, 2012) have proposed a 2-DoF mass-spring-damper system model of mechanical pain that can be tuned to emulate fast and slow pain responses as described above, see also Figure 3.2. The pain level is expressed by the position of the outermost mass (m_2). The pain model based on a 2-DoF mass-spring-damper system model can generate pain intensity responses proportional to the force of impact, which also influences the duration of slow pain. The model was applied to a simple nociceptive withdrawal reflex (Matsunaga et al., 2005). However, such a detailed model could also be applied as a feedback signal for learning and it can even be speculated that such models may contribute to the development of some sort of empathic robots, i.e. robots that could judge subjective pain or tissue damage of others based on the observed speed, shape of the objects involved, among other variables.

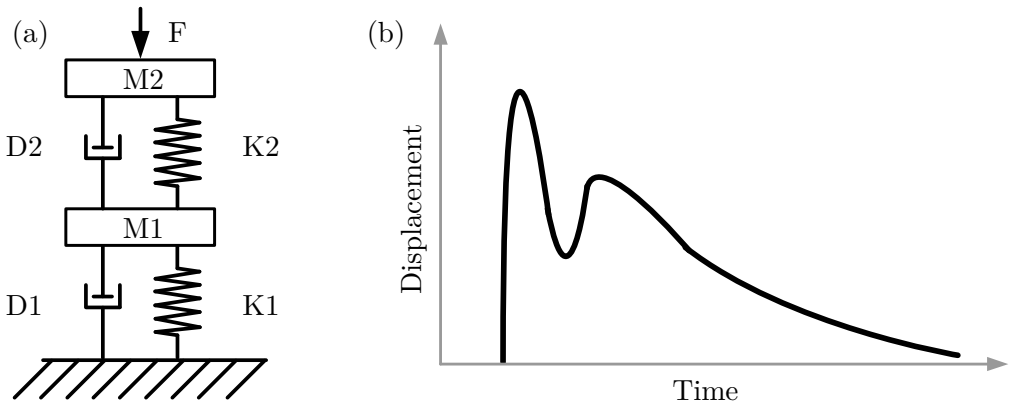


Figure 3.2 – Detailed pain model of mechanical stimuli on human skin. (a) Model based on a 2-DoF mass-spring-damper system. (b) Qualitative dynamical response of the position of the outermost mass of the system. The first and second peak from left to right represent fast and slow pain, respectively. Based on Matsunaga et al. (2005).

3.2.2. Autonomic Reflexes

As discussed previously in this chapter, robots are becoming increasingly complex and this is making it more difficult to anticipate all possible scenarios in which they will operate. However, just as living organisms, robots can directly benefit from hard-wired responses to harmful stimuli or potentially dangerous events. Much like in living organisms, those hard-wired responses could eventually become the basis for more sophisticated and adaptive behaviours through learning. This section will present an overview of biologically inspired self-protective reactions and learning mechanisms associated with painful stimuli or negative outcomes.

Detection and Management of Collisions and Falls

Simple and useful self-protective mechanisms are withdrawal reflexes. In general, reflexes are hard-wired responses elicited by painful stimuli designed to prevent or reduce damage. These painful stimuli could be originated by self-collision and collisions with external objects. Either way, robots need to detect and appropriately respond to these exceptions before they damage themselves, damage surrounding infrastructure or hurt a human user. More complex mechanisms are fall reactions. Fall reactions are meant to prevent a fall or to protect vital organs in an imminent fall. Due to the inherent instability of humanoid robots it seems reasonable to endorse them with fast and adaptable collision and fall management systems instead of just reduce operational speed or pose restriction on the robot's workspace.

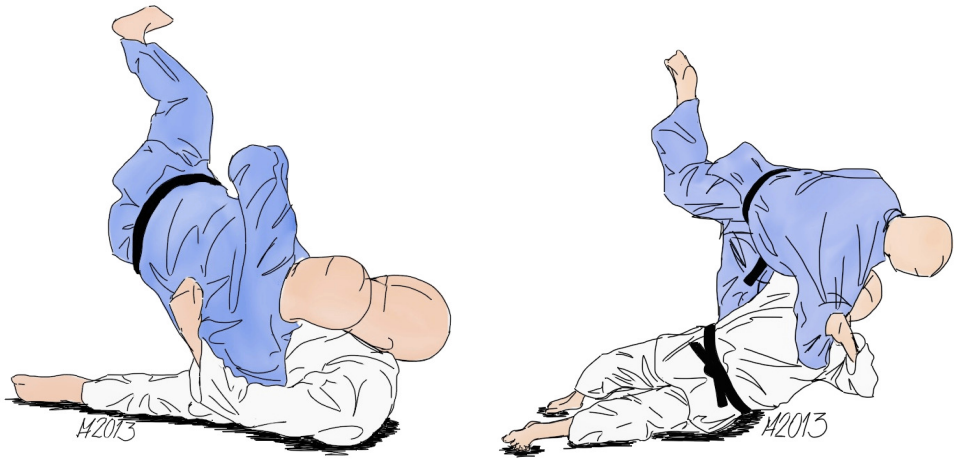
The ideal collision and fall management system for humanoid robots should at least consider a number of heterogeneous aspects (Ruiz-del-Solar et al., 2010):

- The first aspect is related to the mechanics of the robot. In animals, the musculoskeletal systems play a crucial role in dissipating the energy of impacts, thus reducing damage from collisions and falls. Similarly, robots should be designed with soft bodies or equivalent mechanisms that allow a quick and efficient dissipation of kinetic energy such as the one originated in collisions, falls or even locomotion.
- The second crucial aspect involves accurate proprioception and fast instability detection. Both types of information are crucial to choose the optimal collision avoidance or falling sequence.
- The third aspect is concerned with whole body kinematics and the repertoire of falling sequences. It is important to endow robots with a number of hard-coded reflexes from which new learned sequences can derive. These reactive responses should be designed to prevent collisions and falls whenever possible or to protect the most vulnerable components of the robot and reduce overall damage in unavoidable collisions or falls.

- A final aspect involves surrounding awareness. This is crucial to ensure user safety. Collision avoidance and falling sequences should be controlled to prevent hitting and hurting users or, if conditions allow, to facilitate recovery.

Many of the behavioural aspects of controlled falls have been perfected in many martial arts. Full body control and accurate proprioception allow controlled falling sequences which efficiently dissipate kinematic energy and permit fluid recovery. These techniques typically do not intend to disrupt the flow of the motion but rather modify the body posture to roll or maximize the contact area to minimize damage. Similarly these techniques can efficiently deal with obstacles and could be used to improve robots' fall management systems, see Figure 3.3.

Shimizu et al. (2011) designed a reflex management architecture for the humanoid robot iCub (Shimizu et al., 2011, 2012). The architecture efficiently orchestrates pre-programmed responses against collisions and falls. The system was optimized for falls that start from a still and upright pose. Robot responses are divided into global and local reactions. Those reactions are meant to protect the robot's head and torso while reducing the overall damage. Global reactions were designed to provide a whole body reaction when falling and local reactions were designed to respond to particular conditions while performing a global reaction sequence.



(a) (b)

Figure 3.3 – Schematic representation of a fall management procedure based on martial arts falling techniques. (a) Uncontrolled falling on the user may lead to user injury and robot damage. (b) Controlled falling over and away from the user to prevent/reduce user injury and reduce robot damage. By Michael Hultström[5].

Global reaction consists of three sub-reactions executed simultaneously: firstly, extend the arms in the direction of the fall; secondly, lower the legs to reduce

[5]Retrieved from `https://www.flickr.com/people/hultstrom/`
under Creative Commons Attribution-ShareAlike 2.0 Generic (CC BY-SA 2.0)

potential energy; finally, incline the head in the opposite direction of the fall. If the robot fails to detect the direction of fall or an uncategorised exception is detected, the robot performs a general guard pose. The guard pose was intended to protect the robot's head and torso from impacts regardless of direction. To create the guard pose, the head is inclined to the front, while the arms cover the head and legs are folded. Additionally, due to the possible presence of obstacles in the direction of the fall, the robot's limbs could eventually be blocked or the direction of the fall could change increasing damage, thus a number of local reflexes were also pre-programmed to cope with these situations.

Local reflexes, although pre-programmed, can be automatically adapted and work similarly to spinal reflexes (Shimizu et al., 2011). Shimizu et al. developed three types of local reflexes, i.e. a tonic reflex and a pair or myotatic reflexes. The tonic reflex acts as an emergency stop for an individual limb when a collision is detected. The myotatic reflexes are divided in two independent reflexes, i.e. a forward reflex and an inverse reflex. Both myotatic reflexes act synergetically to control the stiffness of individual joints. This control allows joint positioning, gravity compensation and joint compliance using an optimal amount of torque (Shimizu et al., 2011, 2012).

Another sophisticated example of a fall management system for robots was developed by Ruiz-del-Solar and colleagues (Ruiz-del-Solar et al., 2010, 2009). Similarly to the system developed by Shimizu et al. (2012), this system focuses on the falls produced by external events and not inherent in the robot's locomotion. It employs whole body reactions and the system can naturally cope with a fall from any robot position. But contrary to Shimizu et al.'s system, here, the instabilities are evaluated in real time and thus can be used regardless of the robot's gait or pose. The main drawback of the system is that it was optimized to robot soccer applications and thus suitable only on even surfaces.

3.3. Amygdala and Conditioning

In the brain, the amygdala plays a key role in self-protective and affective systems. It sets an affective valence of situations, a 'state-value' necessary for coordinating physiological, behavioural and cognitive responses. Furthermore, recent evidence suggests that the human amygdala, in addition to its important role in cue-dependent conditioning, contributes to many cognitive processes such as reward-based decision-making tasks (Gupta et al., 2011). Understanding these mechanisms based on computational modelling is of scientific interest not only in order to obtain a better interpretation of neuropsychological findings, but also to design autonomous and safer robot assistants.

Computational embodied models dealing with self-protective and fear circuits in the brain are still rare. Currently, most amygdala models are described as abstract models of cued fear conditioning. Many of these models (e.g. Armony et al., 1995, 1997; Balkenius and Morén, 2001) are based on the dual-route hypothesis proposed by LeDoux (1992). The dual-route hypothesis explains parallel processing of biologically relevant stimuli via a subcortical and a cortical pathway. The sub-cortical pathway is characterised by fast and coarse evaluation of stimuli and the elicitation of related autonomic and behavioural responses, whereas the cortical route may be comparatively slow but more accurate. Besides, the cortical route may also have a modulatory effect on the responses elicited by the subcortical route, i.e. enhancing or inhibiting them. And it would most likely elicit additional or corrective responses to those produced via the subcortical route (Resnik et al., 2011).

Another recurring feature of most models of the amygdala and fear conditioning dynamics is the use of oversimplified input and output signals, typically binary or abstract numerical inputs and outputs (Balkenius and Morén, 2001; Krasne et al., 2011; Lowe et al., 2009; Mannella et al., 2008; Vlachos et al., 2011), thereby neglecting the nature of the information, e.g. interdependence, temporal scales, and quality of pre-processing. In other words, these are, to a great extent, disembodied computational models. However, the virtue of these early models lies in the identification of essential components that enable conditioning and affective processing.

An early anatomically constrained model for conditioned fear was suggested by Armony and colleagues (Armony et al., 1995, 1997; Armony, Servan-Schreiber, Romanski, et al., 1997) to investigate information processing in two afferent pathways to the amygdala, one originating in the auditory thalamus (subcortical pathway), the other in the auditory cortex (cortical pathway). This model is interesting due to its modularity and, to a certain extent, biological plausibility. Armony et al. (1995) showed that the subcortical route is enough for conditioning dynamics and speculated that the cortex may provide finer discrimination capabilities when handling complex sensory stimuli such as temporal patterns, although this was not explored.

Recently, this model was replicated and further analysed by Lowe et al. (2009). Lowe et al. (2009) found that the combined use of population coding, modularity and redundant connections (dual-route) improved robustness, stimuli discrimination and speed of acquisition. Lowe et al. (2009) also found that the model of the ventral division of the medial geniculate body (MGv) of the thalamus fed 'noisy' signals to the cortex module and thus impacted the model negatively in terms of conditioning acquisition. Additionally, the MGv, via the thalamo-cortico-amygdala pathway, seems to have an inhibitory effect on the conditioned response. The inhibitory effect is not necessarily detrimental, but it requires a more sophisticated model

of cortical areas in order to be potentially exploited. Other important criticism made by Lowe et al. was regarding the complexity of the model. Although deemed sufficient is perhaps not necessary for reproducing behavioural data qualitatively, thus impacting negatively of the insightfulness of the model results. Finally, Lowe et al. (2009) criticized the lack of recurrent connections within the amygdala and reciprocal connections from the amygdala to both the thalamus and the cortex, not to mention the disembodied aspects of the models, the synergetic interplay between the full connected network and the input signals with respect to generalisation.

A more comprehensive computational model for conditioning, including acquisition and habituation, was presented by Balkenius and Morén (2001). The authors focused on the interaction between the amygdala and the orbitofrontal cortex, where the former serves as the locus for acquisition and the latter for inhibition of fear responses. Balkenius and Morén (2001) also reported that the model was capable of learning conditioned acquisition with no more than the subcortical route. Moreover, acquired conditioned responses were not forgotten by the model, but only inhibited by the prefrontal cortex. Additionally, the model was able to partially 'block' the learning of a new stimulus (CS2) when presented simultaneously with the primary conditioning stimulus (CS1). This partial blocking[6] did not prevent the activation of the amygdala, but its response decreased quickly to a very low activation. Balkenius and Morén (2001) speculated that the simplicity of the thalamus and sensory cortex model prevented full 'blocking'. This model has been recently extended with a hippocampus model and then used to implement a dynamic associative memory (Kuremoto et al., 2009). Although this new extended model does not focus on conditioning acquisition or expression, it does use the *affective* responses of the amygdala model to improve memory formation in the hippocampus model.

Lowe et al. (2011) proposed an interesting model that sheds light to the temporal aspects of conditioning acquisition and inhibition. The model is based on the dopaminergetic dynamics between the amygdala, ventral tegmental area (VTA) and prefrontal cortex (PFC). Similar to the model of Balkenius and Morén (2001), conditioning acquisition is obtained via a subcortical route and inhibition of conditioned responses originates in the prefrontal cortex model. However, the main difference lies in the inclusion of a biologically plausible way of handling the association of temporally distant stimuli, and an inhibition of conditioned responses via a reservoir computing model of the PFC.

Less research has been done on computational models of context conditioning. Context conditioning is important for triggering defensive behaviour only in appropriate circumstances, allowing the organisms to switch between defensive and

[6] *Blocking* refers to the phenomenon observed when a successfully associated conditioned stimulus (CS) with a particular unconditioned stimulus (US) prevents the acquisition of additional conditioned stimuli with the same US (Rescorla and Wagner, 1972).

normal behaviour efficiently. Among the sparse literature, the work by Krasne et al. (2011) and Vlachos et al. (2011) can be highlighted. Both developed biologically plausible models of the amygdala for cued and contextual fear conditioning. Vlachos et al. (2011) presented a spiking neural model of the amygdala's basal nucleus for fear memory encoding. Despite the detailed model of the basal nucleus, this model neglected the sensory input pathways for cue and context information as well as the interaction with downstream structures. Krasne et al. (2011) described a rate-coded model of three amygdala nuclei, the lateral nucleus, the basal nucleus and the medial central nucleus of the amygdala. In contrast to Vlachos' work, they addressed fear conditioning in a more integrative manner, modelling not only one amygdala's input nucleus but also nuclei involved in the expression of fear responses. Despite the broad dynamics captured, such as fear acquisition, consolidation, and extinction, they used an abstraction of sensory and contextual information that prevents the embodied robotic examination of the model.

In general, there is a lack of embodied neural computational research on amygdala modelling with realistic sensory input taken from a physical environment. In one rare example, Mannella et al. (2008) performed a robot simulation where a mechanical rat is used in experiments of first and second order conditioning. Mannella et al. (2008) simulated the two-step mechanism of conditioning acquisition, i.e. a stimulus-stimulus associations (CS-US-UR) stored in the basolateral complex of the amygdala (BLA) and a direct stimulus-response association (CS-UR) stored in the pathway formed by the lateral (LA) and central nucleus (CeA) of the amygdala (Cardinal et al., 2002). This mechanism explains how first order conditioning (CS1-UR) allows the acquisition of a second order association (CS2-UR)[7]. Similar to the disembodied models presented previously, this model also uses binary sensory input and hard-coded conditioned responses. Additionally, the model cannot reproduce conditioned inhibition.

Other embodied models exist (Alexander and Sporns, 2002; Zhou and Coggins, 2002), even though, they primarily focus on behavioural aspects that may be linked to conditioning dynamics, but do not attempt to be neurocomputationally realistic. Moreover, most of these models (Alexander and Sporns, 2002; Zhou and Coggins, 2002) consist of feed-forward networks and do not capture as rich a variety of dynamics as other research (Krasne et al., 2011; Mannella et al., 2008; Vlachos et al., 2011). However, the embodied approach makes them attractive and encourages the development of more sophisticated and biologically plausible embodied models.

[7]The first conditioned stimulus (CS1) anticipates the occurrence of the unconditioned stimulus (US), whereas the secondary conditioned stimulus (CS2) predicts CS1

3.4. Synopsis

The presented computational models account for many aspects that are critical to ensure individual survival. Although the available knowledge of many of these mechanisms is insufficient to produce accurate models capable of reproducing rigorously the same observed behaviour or generating comprehensive quantitative predictions, it is still possible to emulate some neural circuitry and functional aspects of the mechanisms observed in biological systems resulting in models capable of producing qualitative results reliably. Thus computational models can contribute to the interpretation of unexplained phenomena, or become powerful frameworks that can be used to test hypotheses impossible or difficult to examine otherwise. For instance, they can be used to integrate multiple aspects of a single phenomenon or interdependent systems and examine the plausibility of a synergistic relationship or to shed light on minimal conditions that would allow them to work undisturbed within the same system. From this point of view it is clear that biologically inspired computational models not only can contribute greatly to the development of the future robot generations as stressed in this chapter, but also have the potential for a better understanding of human cognition which is far beyond what could be obtained via informal reasoning, because of the number of interdependent systems involved.

The objectives of this research were to develop functional computational models complementary or alternative to many of those presented in this chapter, with the aim of exploring the benefit of autonomic and homeostatic mechanisms in the context of humanoid service robots.

Chapter 4
Methodologies: An Introduction to the Main Techniques Used

This chapter discusses the main machine learning techniques used in the course of this research towards the development of adaptive self-preservative robots. Overviews of the unmodified versions of the algorithms are introduced to facilitate the explanation of the suggested modifications, if any, in the following chapters.

4.1. The Perceptron and Artificial Neural Networks

The perceptron is a mathematical model of a single biological neuron and it is the elemental computing unit of current artificial neural networks (ANNs). The perceptron consists of a vector of synaptic weights $\mathbf{w}$ that multiply the vector of input signals ($\mathbf{u}$) and a bias unit b typically of value -1. The resulting values are then combined into a single value x using a function h (usually sum). Finally, this value is fed into an *activation function* f, also called squashing or transfer function, that produces an output signal y. Perceptron units are usually grouped into layers l. In the special case, when these layers are sequentially arranged, and no connections within layers and cyclic connections between layers exist, they are called multilayer perceptron (MLP) neural networks or feed-forward neural networks, see Figure 4.2.

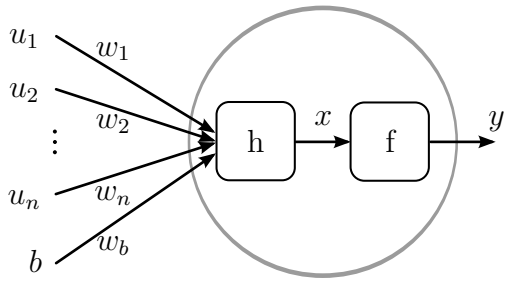


Figure 4.1 – Perceptron neural model, based on López-González (2008, p. 25).

Unfortunately, in the literature, there is an inconsistent notation for the naming and numbering of the network layers. The input units are frequently considered as a perceptron layer, although no computation is carried out there. In this thesis, the input units will be considered as network layers to simplify the mathematical and programming notation. Specifically, the input layer will be given the number 0. The output layer is the layer where the result of the network is read out. All other layers of perceptron units between the input units and the output units are called hidden layers. The number of units per layer as well as the number of layers is heavily dependent on the problem and no rule to compute or even to estimate their size is yet known (LeCun et al., 1998, 2012; Marsland, 2009, p. 64). In spite of this, these networks have been and are being successfully applied to a variety of problems of arbitrary complexity such as universal function approximation, data classification, time-series prediction and data compression.

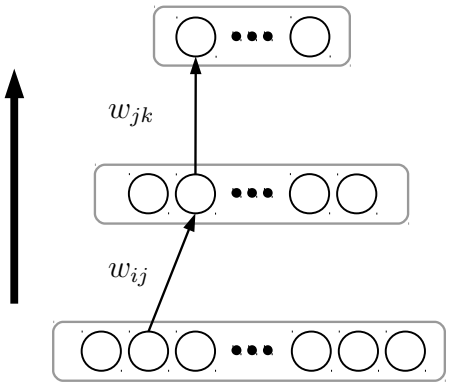


Figure 4.2 – Example of the general structure of a feed-forward neural network.

ANNs can be considered as a neat first step towards embodied systems, because connection weights do not store single data points of a particular data set, but rather their combined values shape the way input patterns are processed to produce a certain output. In other words, ANNs require a certain level of interaction with the real world, on the one hand, to obtain their inputs and on the other hand, to be able to interpret the output. Nevertheless, this was referred to as *a first step* mainly because according to embodiment theory (Pfeifer and Bongard, 2006), the interaction with the world is essential but not sufficient. Based on embodiment theory the design of artificial embodied systems has to mindfully consider the physical and autonomic attributes and limitations of the body. The main aspect that limits the level of embodiment of ANNs is the fact that they highly depend on the designer not only because of the design and configuration needed but more importantly for data preparation and interpretation.

MLPs and ANNs in general, despite being abstract models of neurons, process information similarly as in brain circuits replicating many desirable properties such

as being massively parallel, capable of adaptation over time, robust against noise, fault tolerant and capable of generalization. These properties originate mainly from the topology of the network and also from its capacity to learn. Learning occurs by adapting the parameters of the network, i.e. number of layers, number of units per layer, activation function, connection weights, etc. in order to obtain a determined result.

There are many learning algorithms which can be classified into two main categories, i.e. *supervised* and *unsupervised* methods (Rojas, 1996, p. 78). Supervised or corrective learning is a fully guided method where the designer explicitly indicates the desired output or target vector for every input pattern. In supervised learning the network's parameters are adjusted based on the error between the current network's output and a desired target output. Here, the designer knows in advance the exact output for every input pattern. Maybe the best example of this kind of learning method is the back-propagation algorithm (Rumelhart et al., 1986). On the contrary, unsupervised learning tries to find the underlying probabilistic model of a data set without external guidance. One of the most known unsupervised learning algorithms is Hebbian learning introduced below in Section 4.2. Another popular example is Kohonen's self-organising map, a.k.a. SOM (Kohonen, 1990, 2001). Another well-known group are reinforcement learning (RL) methods, see Section 4.4 below. Reinforcement learning stands between supervised and unsupervised learning methods and it is characterized by the use of sparse and sporadic feedback signals, which may be rewarding or punishing. These feedback signals are used to steer the learning process without explicitly or strictly forming input-output pairs (Sutton and Barto, 1998, p. 4).

4.2. Hebbian Learning

In 1949 the psychologist Donald Hebb postulated a simple, but powerful, reinforcement mechanism between neurons to encode stimulus information leading to memory formation and associative learning (Hebb, 1949). Hebb's postulate accounted for the aspects of synchronous firing of pre- and postsynaptic neurons in learning. This formulation was later completed and refined by Gunter Stent in 1973. Stent hypothesized and showed that the same principle also applies to the efficacy of inhibitory synapses. This concept, sometimes called the Hebb-Stent rule, has been widely applied in computational models of neural networks with great success.

Unfortunately, the basic Hebb-Stent rule is unbounded and it may lead to saturation of the connection weights, but its attractive unsupervised and competitive learning nature driven by neural correlation has motivated the development of several modifications that avoid this problem (Miller and MacKay, 1994).

The modification used in this thesis was taken from the implementation of Armony et al. (1995). Each weight is updated proportionally to the pre- and postsynaptic activation as follows:

$$w_{rs}^{'} = \begin{cases} w_{rs} + \epsilon \cdot a_s \cdot a_r & \text{, if } a_s > \overline{a} \\ w_{rs} & \text{, otherwise,} \end{cases} \tag{4.1}$$

where w_{rs} represents the connection weight between the presynaptic unit s and postsynaptic unit r, ϵ is the learning rate, a_* is the activation of unit $*$, and $\overline{a}$ is the mean activation of the pre-synaptic layer.

To deal with the instability of the basic Hebb-Stent rule, all the incoming connections to a unit are normalized (von der Malsburg, 1973) as follows:

$$w_{rs} = \frac{w_{rs}^{'}}{\sum_{j \in S} w_{rs}^{'}}. \tag{4.2}$$

4.3. Back-Propagation

Back-propagation (BP) is a supervised training algorithm for multilayer perceptron networks suggested by Rumelhart et al. in 1986. Back-propagation is probably the best known training algorithm for artificial neural networks. Its success can be attributed to two factors, the simplicity of the concept and the efficiency with which it can be implemented. However, these two characteristics have also lead to careless use often with sub-optimal results or even failed implementations (Marsland, 2009, p. 54). There are still many intricacies that are not well understood and no principled way of making design choices exists. Hence, their design and set-up has been even considered more an art than a science (LeCun et al., 1998, 2012), because many seemingly arbitrary choices have to be made, e.g. number of neurons and hidden layers, learning rates, etc. Some heuristics to help choose some of these parameters are presented in the following sections.

The idea behind back-propagation is to compute the error E between the current network output y and the desired output t, a.k.a. target, and make it as small as possible by adjusting the connection weights of the network. The error at the output layer is simple to compute, yet the problem remains for the hidden layers, because there is no explicit target for those neurons, hence the name hidden layers (Marsland, 2009, p. 48). However, it is still possible to know how a particular weight needs to be modified to reduce the error E. The derivative of E with respect to this connection weight needs to be computed. So the following update rule is obtained:

$$w_t = w_{t-1} - \eta \frac{\partial E}{\partial w} \,, \tag{4.3}$$

where w_t is the updated connection weight, w_{t-1} is the connection weight's previous value, η is the learning rate, and $\frac{\partial E}{\partial w}$ is the rate of change of the error with respect to weight w. The most common error function to determine the difference between y and t is the sum-of-squares error function. The individual errors for every output unit are combined into a single scalar value as follows:

$$E(t, y) = \frac{1}{2} \sum_{k=1}^{n} (t_k - y_k)^2, \tag{4.4}$$

where the factor 1/2 is used to simplify the following calculations and has no major effect on learning, t_k is the target value for the k-th output unit, and y_k is the activation of the k-th output unit. Consider a 3-layered network as the one shown in Figure 4.2. The network activation can be written as:

$$y_k^2 = \mathrm{f}_2 \left(\sum_j w_{jk}\ y_j^1 \right), \tag{4.5}$$

$$y_k^2 = \mathrm{f}_2 \left(\sum_j w_{jk}\ \mathrm{f}_1 \left(\sum_i w_{ij}\ u_i \right) \right), \tag{4.6}$$

where the superscript for unit activation y indicates the layer where the unit is located. Similarly, the subscript for activation functions f indicates the network layer where the function is applied, y_k^2 is the activation of output unit k, f_2 is the activation function for all units of the output layer, w_{jk} is the connection weight between the k-th output unit and the j-th hidden unit, y_j^1 is the j-th output of unit in the hidden layer 1, f_1 is the activation function for units in hidden layer 1, w_{ij} is the connection weight between the j-th hidden unit and i-th input unit, and x_i is the i-th input unit. Because the update rule shown in Eq. (4.3) depends on the derivative of the error, the activation functions f_* need to be differentiable.

$\frac{\partial E}{\partial w_{jk}}$ needs to be computed to update the output weights. This can be achieved by deriving Eq. (4.4) after y_k is substituted with Eq. (4.5). Here, it is necessary to apply the chain rule to obtain:

$$\frac{\partial E}{\partial w_{jk}} = \frac{\partial E}{\partial y_k^2} \frac{\partial y_k^2}{\partial w_{jk}} = \underbrace{(t_k - y_k^2)\ \mathrm{f}_2' \left(\sum_l w_{lk}\ y_l^1 \right)}_{\delta_{ok}}\ y_j^1\ . \tag{4.7}$$

It is useful to define a term δ to simplify the implementation of back-propagation. This will become clear when deriving the error term with respect to the weight w_{ij} between the hidden layer and the input layer. Now, if y_k^2 is substituted by Eq. (4.6) in Eq. (4.4), the gradient for the weights w_{ij} between the hidden layer and the input layer can be computed as:

$$\begin{aligned}\frac{\partial E}{\partial w_{ij}} &= \sum_k \frac{\partial E}{\partial y_k^2}\frac{\partial y_k^2}{\partial w_{ij}} = \sum_k \frac{\partial E}{\partial y_k^2}\frac{\partial y_k^2}{\partial y_j^1}\frac{\partial y_j^1}{\partial w_{ij}} \\ &= \sum_k (t_k - y_k^2)\ \mathrm{f}_2'\left(\sum_l w_{lk}\ y_l^1\right) w_{jk}\ \ \mathrm{f}_1'\left(\sum_m w_{mj}\ u_m\right) u_i \\ &= u_i\ \underbrace{\mathrm{f}_1'\left(\sum_m w_{mj}\ u_m\right)\sum_k w_{jk}\ \delta_{ok}}_{\delta_{hj}}\ . \end{aligned} \tag{4.8}$$

Back-propagation could theoretically be applied to networks with any number of hidden layers. However, two hidden layers are enough to approximate any function with arbitrary accuracy (Marsland, 2009, p. 64). Furthermore, back-propagation, as any gradient-based training algorithm, becomes rapidly ineffective to update weights in lower layers (close to the input) of a network. This is due to the fact that the second derivatives of the cost function in lower layers are often smaller than in higher layers. In other words, the contribution to the error of each connection weight in lower layers is smaller and thus the changes in weight value are marginal (LeCun et al., 1998, 2012).

4.3.1. The Hessian Matrix in Multilayer Networks

The Hessian matrix H is a measure of the curvature of the error. The eigenvectors are a 'coordinate system' of the dimensions of the error E and the eigenvalues are an estimation of the slope of the error function and the corresponding eigendirection. The eigenvalues are also a measure of the inputs' covariance along the corresponding eigendirection (LeCun et al., 1998, 2012).

When applying back-propagation, typically, a single learning rate η is used to adjust the connection weights of the network. Unfortunately, this poses a problem when working with multidimensional problems, because, depending on the curvature of the error, some weights may require a smaller or larger learning rate to converge or even prevent divergence. Nevertheless, if the Hessian is diagonal, or diagonalized, it can be shown that divergence can be avoided when the single learning rate η is strictly smaller than $2/\lambda_{max}$ and optimal (λ^*) when equal to $1/\lambda_{max}$, where λ_{max} is the largest eigenvalue of H (LeCun et al., 1998, 2012). However, if H is diagonal, a better solution is to assign a different learning rate to every eigendirection where the optimal learning rate is derived from the corresponding eigenvalue (LeCun et al., 1998, 2012).

The ratio between the largest (λ_{max}) and the smallest (λ_{min}) eigenvalue are an indication of the shape of the local minima and thus the speed of convergence.

Large ratios imply very slow convergence along the direction of the λ_{min}. Several techniques have been developed to estimate the Hessian or particular derived information without explicitly computing it. Those techniques have enabled the development of alternative learning algorithms such as conjugate gradient, Gauss-Newton, Levenberg Marquardt and the Quasi-Newton (BFGS) method (LeCun et al., 1998, 2012).

4.3.2. Design Consideration and a Few Practical Tricks

MLPs and back-propagation may be the most popular neural networks and training algorithm. Unfortunately, most of the time, they are applied carelessly leading to suboptimal results (Marsland, 2009, p. 54). This section briefly introduces a few techniques that may help to take advantage of MLP and back-propagation.

4.3.2.1. Stochastic Versus Batch Learning

The gradient $\partial E/\partial w$ of Eq. (4.3) can be computed after either the whole data set has been presented to obtain an average gradient (*batch*) or after the presentation of single examples (*stochastic or online*).

Batch learning enjoys higher popularity, mainly because the convergence properties are better understood which has helped the development of many acceleration techniques and alternative algorithms (e.g. Riedmiller and Braun, 1993). However, the impact on redundancies in the training data, typical of large data sets, may outweigh the speed gained with these acceleration techniques and make them less effective. Batch learning also tends to 'get stuck' at local minima, because of the use of the average gradient which prevents the weights to jump to other regions of the solution space (LeCun et al., 1998, 2012).

On the other hand, online learning is faster than batch learning particularly in large redundant data sets, because learning occurs after every sample is presented rather than after the entire data set. Additionally, due to the constant update, the weights tend to move over the solution space often leading to better solutions. Online learning can also track small changes in data distribution over time. Unfortunately, the same properties that make online learning so advantageous prevent it from full convergence. The weights approach the minimum until an oscillatory state around the minimum is reached, the size of the oscillations is proportional to the learning rate. A counter-measure is to decrease the learning rate as learning progresses. Fortunately, finding the optimal learning rate at every time step is not critical, because over-training may occur before the oscillatory regimen becomes a problem (LeCun et al., 1998, 2012).

4.3.2.2. Choosing the Activation Function

MLPs benefit from certain properties of the input variables, such as mean values around zero and similar covariance for every input variable, see Section 4.3.2.7. Activation functions that are symmetrical with respect to the origin help to keep the mean of the output values close to zero, thus should be preferred (LeCun et al., 1998, 2012).

Sigmoid functions are among the most common activation functions. Sigmoid functions are a family of s-shaped functions characterized by being bounded and differentiable functions. Typical examples are the logistic and the hyperbolic tangent (tanh) functions. A recommended sigmoid is proposed by LeCun et al. (1998, 2012):

$$\mathrm{f}(x) = 1.7159 \ \tanh(2/3 \ x). \tag{4.9}$$

The particularity of the sigmoid presented in Eq. (4.9) over the standard logistic function, or other functions non-symmetrical with respect to the origin or x-axis, lies in the fact that the output's mean will be close to zero and the variance will be close to one, see Section 4.3.2.7. These particular constants maximize the second derivative at $x = 1$ which gives 'dynamism' to the weight updates, see Eq. (4.3). Besides, *saturation* of the output is prevented when the activation function operates outside its asymptotic range. This has two direct benefits: firstly, it prevents the weights from becoming unnecessarily and dangerously large; and secondly, it helps to increase the output difference between patterns which facilitates classification or decision. The remaining areas of small derivatives (flat areas) close to the origin can be addressed by horizontally shifting the sigmoid in a term C, e.g. $\mathrm{f}(x) = A \ \tanh(Bx) + Cx$.

4.3.2.3. Initializing the Weights

The starting values of the connection weights also play an important role in learning. The activation function and distribution function of the training set are important factors for choosing good starting values. The effect of these factors can be inferred from the shape of the activation function. For instance, let us consider that the activation function is a sigmoid function, when the weight values are very large the neurons' output will likely be within the *asymptotical* range from the sigmoid. Operation in this region of the function leads to very small gradients and thus negligible weight updates. Similarly, if the connection weights are too small, the neuron's activation will be close to zero which also leads to very small gradients and thus learning will be slow.

Assuming that the activation function recommended in Section 4.3.2.2 is used and the input vectors are normalized as explained below in Section 4.3.2.7, i.e. have

mean close to zero, covariance ~ 1 and are possible decorrelated, then the initial weight values should be randomly drawn from a distribution with zero mean and a standard deviation given by $\sigma_w = m^{-1/2}$, where m is the number of inputs to the corresponding unit (LeCun et al., 1998, 2012).

4.3.2.4. Momentum

Momentum can best be explained with the analogy of a ball rolling down a hill typically used when introducing back-propagation. When rolling down a hill a ball gains momentum and hence is able to overcome small irregularities on the terrain. Similarly, a momentum term added to the weight update equation, Eq. (4.3), can help overcome shallow local minima, while also acting as a primitive mechanism to adapt the learning rate.

The momentum term consists of the previous value used to update the weights multiplied by a small factor μ. The value of μ typically ranges between $[0, 1[$.

$$\Delta \mathbf{w}_t = \eta \frac{\partial E_t}{\partial w} + \mu \Delta \mathbf{w}_{t-1} \ . \tag{4.10}$$

4.3.2.5. Choosing Learning Rates

As indicated in Section 4.3.1, the learning rate plays a key role in convergence and learning speed. Typically, a single value learning rate is used, but a dedicated learning rate per connection weight can significantly speed up learning, even more if the learning rate adapts to the curvature of the error.

In case a single learning rate is used and if the Hessian has been diagonalized, then the optimal learning rate is given by $1/\lambda_{max}$ (LeCun et al., 1998, 2012). When using dedicated learning rates per weight but with fixed values, they should be proportional to the square root of the number of inputs to the corresponding neuron ($\propto \sqrt{m}$) and decrease in size towards the output (higher) layer (LeCun et al., 1998, 2012), because the derivative of the error is more pronounced closer to the output as mentioned at the end of Section 4.3.

Many methods to automatically adjust the learning rates have been proposed (Darken and Moody, 1991; Murata et al., 1996; Sutton, 1992). Out of these three, the simple and automatic method suggested by Murata et al. (1996) can be highlighted. Murata et al. (1996) suggest to automatically adjust the learning rate without explicitly computing the Hessian. Murata et al.'s method works under the assumption that the smallest eigenvalue of the Hessian is much smaller than the second smallest eigenvalue and thus after a large number of iterations, the weights will converge from the direction of the minimum eigenvector of the Hessian. The

algorithm uses the size of the error to either increase or decrease the learning rate as follows:

$$w_t = w_{t-1} - \eta_{t-1}\frac{\partial E}{\partial w}, \tag{4.11}$$

$$\mathbf{r}_t = (1-\delta)\mathbf{r}_{t-1} + \delta\frac{\partial E}{\partial w}, \quad (0 < \delta < 1), \tag{4.12}$$

$$\eta_t = \eta_{t-1} + \alpha\,\eta_{t-1}\,\left(\beta\,\|\mathbf{r}_t\| - \eta_{t-1}\right), \tag{4.13}$$

where Eq. (4.11) is the equation used to update the weight values, $^{\partial E}/_{\partial w}$ is the average of the gradient of the training episode, $\mathbf{r}$ is an auxiliary variable used to keep track of the size of the weight update, δ controls the influence of the current error in the new learning rate values, α, β are empirically adjusted constants. As a point of reference for these constant values, the size of these constants used by Murata et al. (1996) in a sound separation problem are: $\delta = 0.01$, $\alpha = 0.002$ and $\beta = 20/\hat{\mathbf{r}}$, where $\hat{\mathbf{r}}$ represents the maximal observed value of $\mathbf{r}$.

Besides a proper design there are other factors that can improve convergence and learning speed of the back-propagation algorithm. As most machine learning algorithms, back-propagation benefits from pre-processing the data set (Marsland, 2009, p. 63).

4.3.2.6. Shuffling the Examples

Shuffling the data set can boost learning speed and generalization of online learning. The idea is to produce an *information rich* training data set by taking training examples from different regions of the working space.

A complementary but risky strategy to improve learning speed of online learning is to dynamically adjust the frequency of appearance of input examples. This strategy, known as *emphasizing scheme*, searches for patterns that produce relatively high errors and presents them more often under the assumption that the network has not yet picked up on some of the information in them. This heuristic is particularly beneficial when certain patterns are not well represented in the training data set. However, the danger of the emphasizing scheme becomes evident when the training set has undetected outliers (LeCun et al., 1998, 2012).

4.3.2.7. Normalizing the Inputs

Although input normalization is not essential it is usually beneficial and thus recommended (LeCun et al., 1998, 2012). Ideally, the input variables should fulfil as many of the following requirements as possible.

Firstly, the mean of the input variables should be close to zero. The reason for this is that the *sign* of the input pattern biases the direction of the weight updates thus slowing learning. This requirement also applies to the outputs, because the output of a layer becomes the input of the next layer. Activation functions that are symmetrical with respect to the origin can be used to preserve mean zero at the output values, see Section 4.3.2.2.

Secondly, when all training patterns are of similar significance then the input normalization should be performed in such a way that all input variables have a similar covariance. This helps to increase the relevance of every input variable to roughly the same level, which in turn produces more homogeneous weight updates across the weights connected to the input units. In other words, certain patterns will be captured more quickly and accurately than others, which could reduce the overall network performance. The covariance values should also be within the output range of the used activation function.

A final beneficial transformation consists on decorrelating the input variables, this may be more difficult to achieve though. For instance, Principal Component Analysis (PCA) (Diamantaras and Kung, 1996) can be used to remove *linear* correlations in inputs.

4.4. Reinforcement Learning

The paradigm of reinforcement learning (Sutton and Barto, 1998, p. 18) is a type of associative learning strongly rooted in conditioning (Morén, 2002, p. 61), see Section 2.5. Reinforcement learning (RL) originates from the need to explain and model aspects of conditioning learning and it is particularly useful for investigating instrumental conditioning. Rather than a single model or algorithm, RL represents a complete framework that permits the study of qualitative and quantitative aspects of different phenomena such as conditioning learning, planning, decision making and world representation (Dayan and Niv, 2008). RL focuses on the learning of arbitrarily complex action sequences in a trial and error fashion with sparse and delayed feedback. This is possible by establishing predictive associations between actions and outcomes, integrating gracefully planning and real-time action selection. All these aspects have made it very popular and successful within the machine learning and robotics community.

From a systemic point of view, the reinforcement learning framework consists of the *environment* and *an agent* with a particular goal or task (Wörgötter and Porr, 2005). A continuous interaction with the environment permits the agent to establish a causal relationship between its actions and particular sensory input; in this manner, the agent can learn to move effectively towards its goal. The only

feedback that the agent receives is the one that the agent can perceive with its sensors. The trial-and-error search of the appropriate action for a given situation, together with the temporal association of feedback with actions are the two main characteristics of RL (Sutton and Barto, 1998, p. 4). Despite the fact that this formulation seems to point at an example of unsupervised learning, in reality, subtleties in the definition of the environment and feedback (reward delivery) place RL in an intermediate ground between supervised and unsupervised learning.

Although attractive, the two main characteristics of RL also pose a number of challenges to the implementation of reinforcement learning methods. Some of these are due to the trial-and-error search strategy, which makes it a *relatively slow learning mechanism*. This also restricts RL to *work only with stationary or quasi stationary problems*. The agent's efficiency depends greatly on the selection of the appropriate action sequence under a given circumstance. Hence, a fine discrimination between different states would be ideal. However, this leads to a combinatorial explosion of the solution space, because the agent needs to try actions at each state to determine its suitability. This phenomenon is called the *Curse of Dimensionality* (Sutton and Barto, 1998, p. 17). Another difficulty emerges from the sparse and delayed feedback. Specifically, it is hard to establish a relationship between distant state-action pairs and the received feedback. The feedback is weakened or diluted over time which may lead to convergence problems. This issue is known as the *(Temporal) Credit Assignment Problem* (Minsky, 1961; Sutton, 1984). Another important issue is known as the *Exploration-Exploitation Dilemma*, and it refers to the trade-off between performing actions with known outcomes and exploring other actions to discover alternative solutions. The proper balance of both aspects is crucial for the survival of autonomous systems (Sutton and Barto, 1998, p. 4).

4.4.1. The Reinforcement Learning Framework

In the simplest terms, the reinforcement learning framework consists of an *agent* that makes decisions and performs actions towards a goal, and an *environment* where the agent operates, but there are a number of elements that influence the effectiveness of the agent. Specifically, these are a *reward function*, a *value function*, a *policy*, and the *world representation* (Sutton and Barto, 1998, p. 7).

4.4.1.1. Reward Function

Within the reinforcement learning framework, a reward or feedback function acts as an incentive and a reinforcer for the agent, i.e. they have an intrinsic or learned value for the agent. Rewards (punishments) are single values associated

with different stages of progress or appropriateness of actions, for instance, in the case of a hungry animal smelling the food, seeing the food, tasting of food. Receiving these values can increase or decrease the likelihood of performing certain actions and thus have an influence in the agent's strategy, a.k.a. policy. Reward values are typically denoted by the letter r.

4.4.1.2. Value Function

The value function depends directly on the problem to be solved. It encodes the agent's expectation of receiving a reward when an action a was performed when being in a state s. As such, the value function helps the agent to plan action sequences looking beyond the immediate reward value. The value function assigns internal 'desirability' values to the states depending on its importance when trying to fulfil the task. Value functions are also referred to as *cached values*, because of the way they encode the agent's experience. More specifically, value functions combine reward information with successful action sequences into a single scalar value (Dayan and Niv, 2008).

The value function is a crucial component of reinforcement learning algorithms, because action selection and the success of the agent depend exclusively on its value. For instance, consider the case of a mouse placed in a labyrinth with glass walls, the mouse is able to see the food (reward), however, to reach it, it may need to first go in the opposite direction. Contrary to reward functions, the value function depends solely on the agent's experience and it is created and constantly reshaped.

The main role of the value function is to keep track of the agent's experience under a given strategy and if used appropriately can help to organize the search for alternative strategies. There are two possible methods of storing the agent's experience, i.e. in a *state-value function* or in an *action-value function*. The *state-value function* is used to encode the desirability of state s when following a policy π and it is denoted as $V^\pi(s)$ or simply V^π. Similarly, the *action-value function* is used to encode the desirability of taking action a when in state s under a policy π and it is denoted as $Q^\pi(s, a)$ or simply Q^π (Sutton and Barto, 1998, p. 68).

4.4.1.3. Policy

The policy describes the agent's strategy to solve a task or to reach a goal. Sutton and Barto (1998, p. 7) describe it as "a set of stimulus-response rules". An agent may develop different policies of varying complexity depending on the size and particularities of the problem. The policy is directly related to the above-mentioned *Exploration-Exploitation Dilemma*, i.e. the trade-off between exploration and exploitation.

RL methods can be categorized into on-policy and off-policy methods (Sutton and Barto, 1998, p. 122). On-policy methods try to improve the value function of the same policy that is used to make decisions, and thus directly face the *Exploration-Exploitation Dilemma*. An advantage of on-policy methods is that they can benefit from bootstrapping techniques, i.e. value estimates can be improved based on previous value estimates, see Section 4.4.1.2. The main disadvantage of on-policy methods is that they can only train one value function at a time. Furthermore, they tend to never fully converge to the policy being followed, also called optimal policy (π^*). An example of an on-policy method is SARSA (Sutton and Barto, 1998, p. 145).

On the contrary, off-policy methods try to learn a policy π, a.k.a. *estimation policy*, using information collected while performing actions derived from a different policy π', a.k.a. *behaviour policy*. Therefore, off-policy methods are not necessarily affected by the *Exploration-Exploitation Dilemma*. This permits to learn more than one policy at a time, but the strict division between estimation and behaviour policy makes the use of bootstrapping techniques difficult, because previous estimates of the value function have no influence on action selection. An example of an off-policy method is Q-learning (Watkins, 1989; Watkins and Dayan, 1992).

4.4.1.4. World Representation

The final component of RL algorithms is the internal world representation or model. This representation helps the agent to make predictions about the outcome under different conditions without explicitly performing the actions. The world model may be used in conjunction with the value function for action planning and decision making. However, it is not necessary for all RL algorithms which gives rise to two big categories, i.e. model-based (a.k.a. indirect) and model-free (a.k.a. direct) algorithms. Both types are important for animals and they are used depending on the problem and circumstances. But it is still not clear how both methods may interact, e.g. communicate, cooperate and compete (Dayan and Niv, 2008).

Model-based RL methods build a detailed internal model of the environment based on the agent's experience. This model is then used to efficiently plan future action sequences. Model-based methods require means to efficiently store information from the environment, so that it is a faithful representation of the environment dynamics, quickly searchable, and permit continuous updates to adapt to environmental changes or to correct wrong beliefs. Model-based RL methods are associated primarily with cortical circuitry for planning and decision making (Daw et al., 2005), but they also share some aspects of decision making in planning of the limbic circuitry (Dayan and Niv, 2008). Examples of model-based RL algorithms are Dyna (Sutton, 1991a,b) and prioritized sweeping (Moore and Atkeson, 1993).

On the contrary, model-free RL methods use the agent's experience to learn and adapt a value function and a policy without explicitly using or creating a world model. In model-free RL, information of the environment is directly combined with previous experiences into a value function, which are abstract internal constructs of the world and successes of the agent, see Section 4.4.1.2. Therefore, it is harder to get rid of wrong beliefs or estimates, or to make fine distinctions between rewarding states and particular transitions that lead to high reward. These characteristics have been associated with dopamine activity and conditioning learning (Daw et al., 2005) and the role of subcortical areas such as the striatum and the amygdala (Balleine, 2005). Model-free methods are simpler in terms of online decision-making, however, they require more trial-and-error experience to update the value function and thus make good predictions of future outcomes. Examples of model-free RL algorithms are *Temporal-Difference* (TD) *Learning* (Sutton, 1988) and *Actor-Critic* (Barto et al., 1983).

4.4.2. Temporal-Difference Learning

The most influential variations of reinforcement learning are those based on the error between the predicted ($V(s)$) and received (r) reward. The predicted value is updated with agent experience and thus the error changes over time leading to the name *temporal-difference* (TD) *learning* (Schultz et al., 1997; Sutton, 1988; Sutton and Barto, 1998, p. 133). TD-learning is closely linked to the release of dopamine (DA) in the brain (Schultz et al., 1997) and thus with movement control, reward prediction and motivation (Doya, 2002). However, TD-learning suffers from various limitations with respect to its explanatory power of conditioning, mainly due to the high simplification of reward- and punishment-driven adaptive decision-making mechanisms that models. One of the main disadvantages is that it makes no distinction between Pavlovian and instrumental conditioning. Moreover, it does not model different stages of conditioning such as acquisition and expression (Mirolli et al., 2010).

TD-learning belongs to the model-free RL methods and combines ideas of dynamic programming (DP) and Monte Carlo methods for solving reinforcement learning problems (Sutton and Barto, 1998, p. 133). The model-free aspects of TD-learning are derived from Monte Carlo methods, whereas value function updates are based on bootstrapping from dynamic programming, i.e. updates are based on previous values/estimations. Another important characteristic of TD methods is that they are fully incremental and thus can be easily implemented for on-line learning. TD methods converge asymptotically to any fixed policy π (Sutton and Barto, 1998, p. 138).

The simplest TD-learning method is known as TD(0) and it is described algo-

rithmically in Algorithm (4.1). In Algorithm (4.1) the following variables are used: V is the state-value function, s is the state before performing action a, s' is the resulting state after performing action a, r is the observed reward, α is the leaning rate, and γ is the discount factor. The target for TD updates is given by $r + \gamma V(s')$ (Sutton and Barto, 1998, p. 134).

Algorithm 4.1 TD(0) for estimating V^π.

```
Initialize the value function V(s)
for episode do
    Initialize the state value s
    repeat
        Select an action a following policy π when in state s
        Perform action a
        Sense resulting state s' and received reward r
        update the value function: V(s) ← V(s) + α (r + γV(s') − V(s))
        prepare for next iteration: s ← s'
    until s is terminal
end for
```

4.4.3. Actor-Critic Reinforcement Learning

Actor-Critic methods are on-policy TD-learning methods that have two memory structures, i.e. a dedicated memory for policies and another for value functions, see Figure 4.3. The Actor represents the policy and this is denoted as $A(s)$. The Critic provides a state-value function $V(s)$. The Critic evaluates the outcome of the selected action in the form of a TD error. The TD error is then used to update both the Actor and the critic. If the error is positive, the selected action should be strengthened, whereas a negative error may suggest the opposite (Sutton and Barto, 1998, p. 152).

4.5. Eligibility Trace

Eligibility traces can be seen as primitive memory mechanisms, they keep track of previous activation values. Memory of previous activations can so be used to influence the amplitude of new activations. Eligibility traces are a sort of moving average with a decay factor. Physiological evidence suggests that processes equivalent to long-lasting eligibility traces regulating dopaminergic circuits exist. Moreover, the synergetic work of these long-lasting eligibility-trace-like processes with low per-trial learning rates are essential for reward learning in the brain,

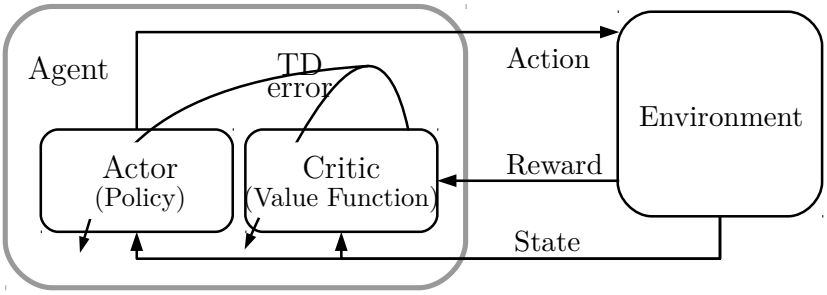


Figure 4.3 – Generic representation of Actor-Critic architectures. Adapted from Sutton and Barto (1998, p. 151).

particularly when the number of examples is limited, i.e. the number of state-reward pairs (Pan et al., 2005).

In the field of reinforcement learning, eligibility traces are described as a basic mechanism for temporal credit assignment and, when applied to TD learning methods, create a continuum between TD and Monte Carlo methods. Eligibility traces are particularly effective when dealing with long-delayed rewards or non-Markov tasks. From a technical point of view, eligibility traces are parameterised by a variable λ which indicates the length of a particular trace in time, and a variable γ which is the *trace decay* and controls the 'vividness' of the memory. Values of λ range from $[0, 1]$, where $\lambda = 0$ means no memory is kept and $\lambda = 1$ depends on the particular time scale of the model. Values of γ also range from $[0, 1]$, where $\gamma = 1$ indicates no decay within the time window defined by λ, and $\gamma = 0$ means that nothing is remembered. Unfortunately, there is no deterministic way of computing the ideal value of neither λ nor γ and thus they need to be empirically determined. Typically, every connection weight or state has an associated eligibility trace. Applications of the concept of eligibility traces to TD-learning methods have shown to speed up learning (Sutton and Barto, 1998, pp. 163, 181).

Eligibility traces can be used, for instance, to remember neural activation or how often a state has been visited. These values can have discrete or continuous values and they are usually bounded. Traces can be implemented in two ways based on how new values are incorporated into the trace value. The first combines the new activation value with the existing trace and is called *accumulating traces*. The second alternative simply replaces the previous trace value with the new activation value and is known as *replacing traces*. In TD-learning, replacing traces have shown considerably better results than accumulating traces (Sutton and Barto, 1998, p. 186).

4.6. Echo State Networks

Recurrent Neural Networks (RNN) differ from MLP in that the perceptron units are not strictly organized in layers or in a feed-forward manner. This means that the activation of perceptron units may depend not only on the current input but also on previous activations. The recurrent connections within a RNN permit the creation of dynamical memories and processing of non-linear time series. Therefore, they may be the most biologically plausible neural network models in terms of their internal recurrent dynamics. Unfortunately, the recurrent connections pose several challenges to training procedures. Specifically, training methods are theoretically poorly understood, are often of high computational complexity, are prone to instability, have complex performance surfaces and training is slow and difficult. Although there are methods that address some of these issues, they are still not suitable for real-world applications, because their set-up requires experienced judgement (Jaeger, 2002; Ozturk et al., 2007). Besides most of these training methods are offline and thus not biologically plausible.

Echo-State Networks (ESN) belong to a novel approach to analysing and training RNN called *reservoir computing* (RC) (Jaeger, 2001, 2002). With the potential of simplifying training of RNN while preserving all dynamical properties of RNN. ESN were conceived for engineering applications and rely heavily on use control and system theory to prove and explain properties and dynamics of the network. A similar approach developed independently and in parallel known as *liquid state machines* (LSM) is better suited for spiking neurons and biologically oriented modelling (W. Maass et al., 2002). Regardless of their origins reservoir computing techniques bare a surprisingly resemblance to the sparse recurrent dynamics of cortical micro-columns (Buonomano and W. Maass, 2009).

The main particularity of ESN is the use of a hidden layer of sparsely and randomly connected neurons, the value of which is fixed upon creation. This hidden layer is called *dynamical reservoir* (DR). The reservoir is connected to a trainable recurrence-free readout network which is used to generate the ESN output. Figure 4.4 presents a conventional ESN architecture. The concept of only training the readout network is based on the observation that connection weights change most at the output layer while deeper layers remain mostly constant, see Section 4.3. This significantly reduces the training complexity from complex gradient descent methods to even simple linear regression, while still preserving the benefits of recurrent networks.

This separation into reservoir and readout is homologous to the process used for kernel methods. The reservoir is used to expand the input history into a set of diverse dynamics, whereas the readout network is trained to produce the desired output using the oscillatory dynamics from the reservoir (Lukoševičius and Jaeger,

2009).

The activation of the reservoir is updated according to

$$\mathbf{x}_t = \mathrm{f}\left(\mathbf{w}^i\mathbf{u}_t + \mathbf{w}\,\mathbf{x}_{t-1} + \mathbf{w}^b\mathbf{y}_{t-1}\right), \tag{4.14}$$

where $\mathbf{u}$ is the input vector, $\mathbf{x}$ is the activation vector of the reservoir, $\mathbf{w}^i$, $\mathbf{w}$ and $\mathbf{w}^b$ are the input, reservoir and output feedback weight matrices with fixed values, respectively, f are the internal units' activation function, and $\mathbf{y}$ is the output vector.

The activation of output vector, for a single layer readout, is computed according to

$$\mathbf{y}_t = \mathrm{f}_o\left(\mathbf{w}^o\left(\mathbf{u}_t + \mathbf{x}_t + \mathbf{y}_{t-1}\right)\right), \tag{4.15}$$

where f_o are the internal units' activation function for the output layer, and $\mathbf{w}^o$ is a trainable weight matrix. This weight matrix is typically adjusted using linear regression techniques.

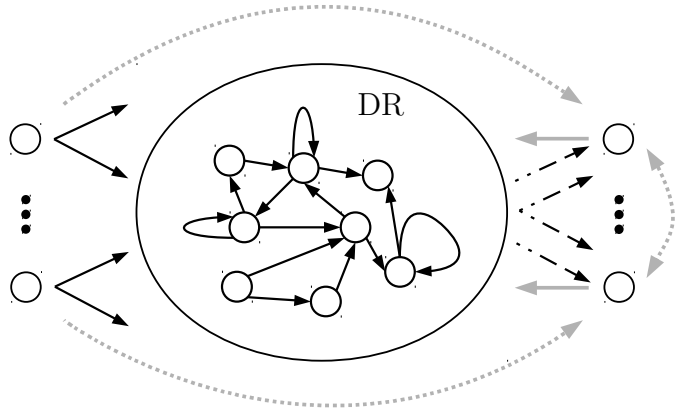


Figure 4.4 – Generic architecture of an Echo State Network (ESN). The basic network architecture is represented by black solid and dashed lines. Solid lines indicate fixed connections, whereas dashed lines represent connections subject to training. Grey solid or dashed lines represent alternative configurations such as direct input-output connections or output feedback connections. Adapted from Jaeger (2001).

4.6.1. Design Consideration of an Echo State Network

The main goal when designing a reservoir is to obtain a rich set of dynamics from a sparsely and randomly connected layer. Typically, the connection ratio within the reservoir is around 20%. Those connections may randomly be drawn from a uniform distribution with zero mean. The size of the reservoir strictly depends on the size of problem and the dynamics of the reservoir are directly driven by the input. An important characteristic for the reservoir is the so-called *echo state*

property (ESP), see Figure 4.5. The ESP, in terms of control theory, refers to the stability of the reservoir, i.e. although a rich set of oscillatory dynamics is wanted it has to remain stable[1]. In other words, if the reservoir is not stable, the responses within the reservoir may grow uncontrollably, and eventually saturate the reservoir units. Consequently, the readout layer may only read constant values (Lukoševičius and Jaeger, 2009).

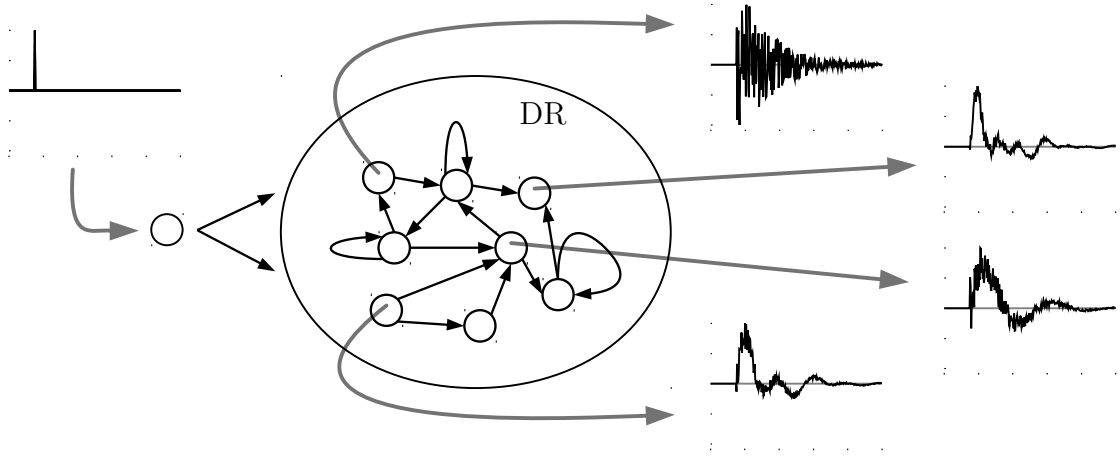


Figure 4.5 – Echo state property refers to the rich set of oscillating and decreasing (damped) responses of the unit of the DR when perturbed. Based on Jaeger (2002).

A simple measure of the stability of the reservoir is the *spectral radius* (ρ). The spectral radius (SP) is the magnitude of the largest eigenvalue of the reservoir weight matrix. From control theory it is known that if the spectral radius is smaller than 1, then the system is stable[2]. Additionally, the magnitude of the spectral radius can be used as a measure of the system's speed. A large SP implies less dumping and thus longer reverberations and 'memory'. However, it also brings the system closer to unstable states. On the contrary, a small SP implies fast dynamics that vanish quickly over time. Although useful, the spectral radius only offers a first and limited view to the reservoir's characteristics, for instance, a spectral radius smaller than 1 only ensures stability, but it does not say much about the echo state property nor is it an indication of performance (Ozturk et al., 2007).

A robust approach for designing reservoirs with the echo state property has been recently proposed by Yildiz et al. (2012). Yildiz et al. (2012) showed that reservoir matrices that satisfy *one* of the following conditions are proven to be diagonally Schur stable[3] and have the echo state property for any input:

[1]The oscillations need to be damped and vanish over time.

[2]Assuming that the reservoir represents the linearised model of the dynamical system.

[3]Yildiz et al. (2012, p. 6) define *diagonal Schur stability* as: "A matrix $\mathbf{w} \in \Re^{N \times N}$ is called Schur stable if there exists a positive definite symmetric matrix $P > 0$ such that $\mathbf{w}^T P \mathbf{w} - P$ is

- $\mathbf{w} = (w_{ij})$ so that $\rho(|\mathbf{w}|) < 1$ where $|\mathbf{w}| = (|w_{ij}|)$.
- $\mathbf{w} = (w_{ij})$ so that $w_{ij} \geq 0, \forall i, j$ and $\rho(\mathbf{w}) < 1$.
- $\mathbf{w}$ so that $\rho(\mathbf{w}) < 1$ and there exists a nonsingular diagonal matrix D so that $D^{-1}\mathbf{w}D$ is symmetric.
- $\mathbf{w}$ is a triangular matrix and $\rho(\mathbf{w}) < 1$.
- $\mathbf{w} \in \Re^{2\times 2}$, $|\det(\mathbf{w})| < 1$, $|w_{11}+w_{22}| < 1+\det(\mathbf{w})$ and $|w_{11}-w_{22}| < 1-\det(\mathbf{w})$,

where $\det(\mathbf{w})$ is the determinant of $\mathbf{w}$.

Based on the first listed condition, Yildiz et al. (2012) suggest a simple recipe to create reservoirs with the echo state property, guaranteed for any input:

1. Create a reservoir $\mathbf{w}$ consisting only of non-negative weights ($w_{ij} \geq 0$)
2. Scale $\mathbf{w}$ so that $\rho(\mathbf{w}) < 1$
3. Change the sign of a desired number of connection weights.

It is a more restrictive condition that $\rho(\mathbf{w}) < 1$ for stability, but it guarantees the echo state property. Although reservoirs with SP larger than 1 are possible, there is no principled way to determine their stability or echo state property.

It is clear that many possible reservoirs with the same spectral radius exist but not all may perform well or even similarly. As mentioned earlier, the SP is a measure of stability and speed of the system, but not performance. Thus a complementary metric is needed for this purpose. The performance of the ESN depends on the diversity of the dynamics of the reservoir. Therefore, one of the goals when designing a reservoir is to obtain a 'rich' set of dynamical responses. Ozturk et al. (2007) suggested the use of the average state entropy (ASE) of the instantaneous echo state as a measure of dynamics' richness. Specifically, they suggested a nonparametric estimator of Renyi's entropy[4] which does not need the probability density function (pdf) of the data and can be computed as follows (Principe et al., 2000; Xu and Erdogmuns, 2010):

$$\mathrm{H}_2(x) = -\log\left[\frac{1}{N^2}\sum_j\left(\sum_i K_\sigma\left(x_j - x_i\right)\right)\right], \tag{4.16}$$

where K_σ is a kernel function[5] of size σ, x are values from the reservoir state vector $\mathbf{x}$, and N is the size of the reservoir.

negative definite. If the matrix P can be chosen as a positive definite diagonal matrix, then $\mathbf{w}$ is called diagonally Schur stable. The positive definite and negative definite matrices are denoted by $P > 0$ and $P < 0$, respectively."

[4]Renyi's entropy is a global measure of randomness or diversity of information.

[5]A kernel is a transformation of data from a dimensional space into a higher dimensional space.

One way to increase the diversity of the dynamics of an ESN without redesigning the reservoir is to use the so-called 'augmented' network states. This is particularly useful when dealing with highly dynamical problems. Augmented network states make again use of a kernel approach for expanding the input. Specifically, to compute the network's output Eq. (4.15) is extended to include the squares of the activation vectors to obtain:

$$\mathbf{y}_t = \mathrm{f}_o\left(\mathbf{w}^o\left(\mathbf{u}_t + \mathbf{x}_t + \mathbf{y}_{t-1} + \mathbf{u}_t^2 + \mathbf{x}_t^2 + \mathbf{y}_{t-1}^2\right)\right). \tag{4.17}$$

As indicated earlier, the dynamics within the reservoir are directly driven by the input. Thus, the dynamical responses of the reservoir also change with it, e.g. if the inputs are close to 0, the neurons tend to operate within their linear range (tanh), while large amplitude inputs will tend to saturate the reservoir neurons and lead to non-linear responses (Jaeger, 2001, 2002). To compensate for these effects the input vectors to the reservoir can be scaled and shifted as described in Section 4.3.2.7. Complementarily, a bias unit can be added to the input vector to adjust the operation point of the reservoir[6]. However, contrary to the MLPs, here it is more convenient to change the bias value (Ozturk et al., 2007) rather than to modify multiple connection weights. The operation point can also be changed by the use of a noise term in the activation function of the reservoir, modifying Eq. (4.14) to obtain Eq. (4.18).

$$\mathbf{x}_t = \mathrm{f}\left(\mathbf{w}^i\mathbf{u}_t + \mathbf{w}\ \mathbf{x}_{t-1} + \mathbf{w}^b\mathbf{y}_{t-1}\right) + v_t, \tag{4.18}$$

where v_t is a small white noise term with an amplitude of 0.0001 to 0.01.

The input weight and the optional output feedback weight matrices can be as sparse or as dense as the reservoir. Similarly as MLPs, the inputs may be scaled and shifted to better match the output range of the activation function, see Section 4.3.2.7. Finally, all the network parameters mentioned above have to be chosen in conjunction. Because, similarly as for the design of hidden layers for MLPs, there is no principled way for choosing ESN parameters but they have to be empirically chosen for every particular application (Lukoševičius and Jaeger, 2009).

4.6.2. Concluding Remarks on Echo State Network

Although the reservoir is designed to have as rich as possible dynamics, ESN are not well suited to deal with different time scales simultaneously, because the dynamics within the reservoir are interdependent. A solution to this problem is to have multiple reservoirs or sub-reservoirs with inhibitory connections (Xue et al.,

[6]The operation point refers to a specific range of operation within the dynamical response of a system, in this case the reservoir.

2007). The successful implementation of these inhibitory connections is achieved by delaying the output from the readout network by one time step (Lukoševičius and Jaeger, 2009).

Other ESN architectures are also being explored such as hierarchical reservoirs (Jaeger, 2007), multiple single readout layers (Skowronski and Harris, 2007), more complex readout such as MLP (Lukoševičius, 2007) or SVM-style readout (Shi and Han, 2007), as well as, different training mechanisms for the readout weights including linear regression algorithm (Jaeger, 2001), and reinforcement learning (Szita et al., 2006).

Chapter 5
Energetic Autonomy and Reward-Seeking Behaviours

Although reward-seeking behaviours have the same evolutionary significance for self-preservation as danger avoidance, they are usually neglected when referring to self-preservative mechanisms (Martin-Soelch et al., 2007). These two types of behaviours, reward-seeking and avoidance, are closely associated with the release of dopamine (DA) in the brain (Schultz et al., 1997) and thus with movement control, reward prediction and motivation (Doya, 2002). For a neural description of the circuitry involved see Section 2.2, Section 2.4 and Section 2.5.

Reward-seeking behaviours in animals can be linked to energetic autonomy, one of the three core types of autonomy defined by McFarland (2009, p. 15): energy, motivation and mental autonomy. As discussed in Section 3.1, for artificial systems energetic autonomy depends on a number of subsystems, e.g. mechanical, electrical and behavioural, that should ideally be designed together to fit the robot's embodiment and task. However, this is not always feasible due to the poor documentation, license issues, etc. Therefore, in this research, we focus on behavioural aspects of reward-seeking behaviours and energetic autonomy for artificial systems. Additionally, the underlying learning mechanisms used here for reward-seeking behaviour enable a wide range of adaptive behaviours that may be further exploited for other applications as will also be shown in this chapter.

5.1. Introduction

Considering the importance of autonomous recharging behaviours, combined with our interest in studying domestic robot applications motivated us to tackle the issue of autonomous recharging for humanoid robots. Humanoid robots like the NAO[1] are used in a growing number of social, service and entertainment robot scenarios (KSERA; Heinrich et al., 2014; Louloudi et al., 2010).

[1]NAO is a small-sized humanoid robot produced by Aldebaran-Robotics.

The NAO's major limitations are its energetic autonomy, which typically does not surpass 45 min, and its lack of ability for precise navigation and positioning, due to slipping and skidding effects, positioning error, manoeuvring speed and jerk. NAO is a commercial platform which poses additional restrictions when developing third party solutions to the problem of autonomous recharging, particularly, when it comes to making modifications to hardware or simply mounting additional equipment on the robot. Therefore, we only provide a conceptual minimally invasive hardware solution for a NAO robot to test our proposed behavioural solution to the problem of NAO's autonomous recharging.

This section provides a description of the considerations and experience gained during the development of this solution with the hope that our design choices provide useful inspiration to more advanced and integral solutions. The main design restrictions are to keep hardware modifications at a minimum, and not to interfere with the robot's mobility or sensor capabilities.

5.1.1. Recharging Station First Prototype

A docking procedure relying solely on the walking capabilities of NAO is challenging, mainly due to its rather coarse manoeuvring capabilities. Thus, we first attempted to take advantage of the dexterity of NAO's arms. We conceptualized a two-stage procedure: firstly, the robot will navigate to a target area and subsequently, the hands will be placed on specially designed electrical contacts to start recharging. Figure 5.1 shows a prototype of this design.

The first prototype of the recharging station consists of a wide and short u-shaped entry bay. This shape prevents collisions during entry or exit[2]. This is required because during walking the robot's upper body oscillates sideways. The low height of the 'armrests' allow the robot to 'kneel' during recharging as shown in Figure 5.1(c). This kneeled pose is more stable than a standing pose. Additionally, in this kneeled pose, also called crouch pose, the motors can be safely turned off to reduce power consumption. Landmarks on the recharging station can be used to aid docking. A forward docking behaviour was chosen to use the robot foot bumper to identify the terminal docking position. To reduce risk of a short circuit, the cathode (positive) and the anode (negative) were installed on different hands. Cables going from the hands to the battery are placed along the arms.

After preliminary testing, all elements work as expected. However, the cables could only be installed on the outside and were required to be loose to cope with the arm movements, thus they are prone to get caught on surrounding obstacles. Therefore, we decided not to further develop this solution.

[2]The recharging station can be covered with foam or a similar material to prevent scratches on the robots and reduce the noise in case of collisions.

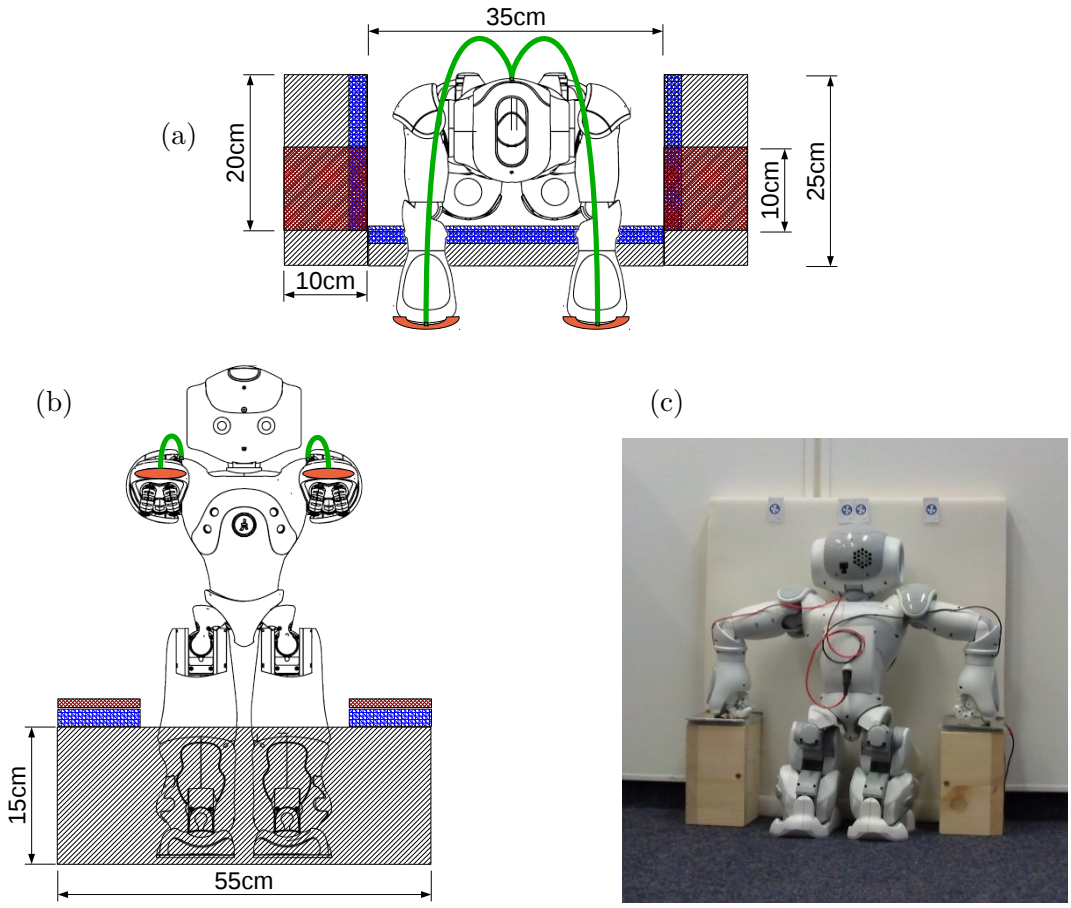


Figure 5.1 – Recharging station for NAO. Green lines represent cables, orange pads on the hand and red on the station are the cathode and the anode. Blue areas represent foam. Grey indicates wood. (a) Schematic of the recharging station top view. (b) Schematic of the recharging station front view. (c) Demonstration of recharging station with NAO.

5.1.2. Recharging Station Second Prototype

Due to the inconvenience caused by the long cables going from the battery to the robot's hand, we opted for a partial backward docking, despite the challenge to manoeuvre the robot backwards (Navarro, Weber, et al., 2011; Navarro-Guerrero, Weber, et al., 2012). This offers advantages such as easy mounting on the robot's back. It does not limit the robot's mobility, does not obstruct any sensor, nor does it require cables going to the robot's extremities. The partial backward docking allows a quick deployment after the recharging has finished or if the robot is asked to do some urgent tasks, because the robot is no longer facing a wall.

This second prototype built for the proposed autonomous recharging is shown

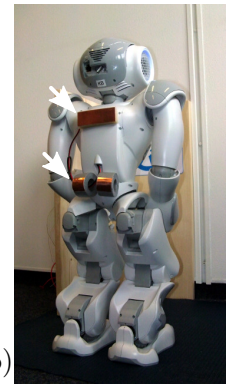
(a)  (b) 

Figure 5.2 – Backward docking station for NAO. (a) White arrows indicate the electrical contacts placed on the docking station and gray arrows indicate the landmarks' position. (b) the robot's electrical connections.

in Figure 5.2. The large landmark (naomark[3]) is used to localize the recharging station when the robot is more than 40 cm away from the station[4], while the two smaller landmarks are used for an accurate docking behaviour, see Section 5.3.

The overall autonomous recharging was split into four phases:

- During the first phase a coarse approach behaviour takes place, which navigates the robot to a distance of approximately 40 cm away from the landmarks; see Figure 5.3(a). This behaviour is temporarily a hard-coded algorithm that searches for the charging station via a scanning head rotation followed by a robot rotation. The robot estimates the charging station's relative position based on geometrical properties of the large landmark and moves towards the charging station. This phase can be replaced or combined with more sophisticated methods such as the one developed in the KSERA project (KSERA, Yan et al., 2012, 2013) where a ceiling camera is used to locate the robot anywhere within an indoor room and then navigate the robot to a desired target position.
- In the second phase the robot re-estimates its position and places itself so that its left shoulder as well as its face are oriented towards the landmark, as shown in Figure 5.3(b).
- In the third phase a reinforcement learning algorithm is used to navigate the robot backwards very close to the electric contacts as presented in Figure 5.3(c)[5].

[3] 2-dimensional landmark provided by Aldebaran-Robotics
[4] Distance measured from the landmark to the robot's camera
[5] In this docking phase, NAO's gaze direction is oriented towards the landmark.

- After reaching the final target position, in the fourth and final phase, a hard-coded algorithm moves the robot to a crouch pose; see Figure 5.3(d). Then, the motors are deactivated and the recharging process starts.

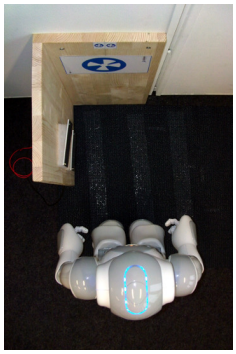
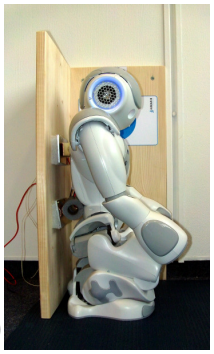
(a)  (b)  (c)  (d) 

Figure 5.3 – Top view of the autonomous robot behaviour in its four different phases: (a) Approaching, (b) Alignment, (c) Docking and (d) Recharging in crouch pose.

5.1.3. Forward Docking Station for Grasping

As mentioned earlier, behavioural aspects developed for autonomous recharging can be exploited to enable other reward-seeking behaviours. In particular, we applied the same principles described in Section 5.1.1 and 5.1.2 to an autonomous grasping scenario. A conceptual schema of the set-up for grasping is depicted in Figure 5.4. The forward docking consists of two phases. The first phase contains all the features of the coarse approach described for the backward docking. Since the coarse approach behaviour places the robot facing the landmarks at approx. 40 cm away, the transition from coarse docking to precise docking does not require an alignment phase. The second phase corresponds to the precise docking behaviour implemented using the RL algorithm described later in Section 5.3, which navigates the robot forward to the docking station and places it within 15 cm proximity of the landmark. Once the robot is in this position the grasping task takes place.

5.2. Motivation for the Learning Mechanism

Reinforcement learning (RL) (Sutton and Barto, 1998) is the leading framework behind learning goal-directed behaviour, see Section 4.4. RL models are able to learn complex action sequences in a trial and error fashion and have been successfully applied to a variety of problems such as navigation tasks (Conn and Peters, 2007; Muse and Wermter, 2009) and resource allocation (Sutton and Barto, 1998, pp. 274, 279).

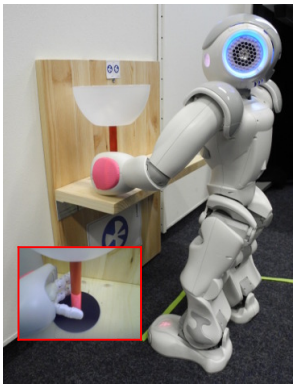
(a)  (b) 

Figure 5.4 – Scenario for grasping a cup from a shelf. (a) Shelf with landmarks for accurate docking behaviour and a graspable object. (b) NAO robot is in grasping position (the inset shows robot's view).

Under a reinforcement learning paradigm, an agent learns from the interactions with its environment while trying to maximize a reward signal (or minimize punishment). At the beginning, the agent does not know the relationship between its actions and the immediate or future reward, but by performing actions the agent is able to build associations between particular situations and actions that lead to high reward values. The trial-and-error search of the appropriate action for a given situation and the delivery of delayed reward values are the two main characteristics of, temporal-difference learning methods of, RL (Sutton and Barto, 1998, p. 4).

For tasks with delayed reward, methods based on temporal-difference (TD) learning have been broadly accepted because of their simplicity requiring minimal computational power, as indicated by Sutton and Barto (1998, p. 158) and supported by a vast body of research (Conn and Peters, 2007; Ghory, 2004; Ito et al., 2007; Kietzmann and Riedmiller, 2009; Provost et al., 2004; Zang et al., 2010). TD-based methods do not require detailed models of the environment and are fully incremental, i.e. are capable of learning based on the agent's accumulative experience (Sutton and Barto, 1998, p. 138), see Section 4.4.2 for an extended description.

In the literature, reinforcement learning is usually used within simulated environments or abstract problems (Ghory, 2004; Provost et al., 2004; Saeb et al., 2009; Weber and Triesch, 2009; Zang et al., 2010). Here, a model of the agent-environment dynamics is available, which is not always available or easy to infer in real-world problems. Moreover, a number of assumptions, which are not always realistic, have to be made, e.g. on the state-action transition model, the design of the reward criterion, and the magnitude and kind of noise if any, etc.

On the other hand, real-world RL approaches are scarce (Conn and Peters,

2007; Ito et al., 2007; Kietzmann and Riedmiller, 2009), mostly because RL is expensive in data or learning steps, the state space tends to be large and the turnaround times for results are long. Moreover, real-world problems present additional challenges, such as safety considerations, real-time action execution, changing sensor characteristics, actuators and environmental conditions, among many others.

Several techniques exist to improve real-world learning capabilities of RL algorithms. *Dense reward functions* (Conn and Peters, 2007) provide performance information for intermediate steps, thereby shaping the policy and restricting the emergence of novel unforeseen policies. *State space reduction* (Conn and Peters, 2007) is dependent on the particular problem and can be a very time-consuming design task. Another approach proposes modification of the agent's properties to fit the given problem (Ito et al., 2007), which relies on a smart definition of the state space that accounts for a reduction of dimensionality. *Batch reinforcement learning* (Kietzmann and Riedmiller, 2009) uses information from past state transitions, instead of only the last transition, to calculate the prediction error function based on storage and reuse of state-action pairs. *Supervised reinforcement learning* (Conn and Peters, 2007; Zang et al., 2010) is based on batch RL, but differs in the generation of training examples. In batch RL, the state-action pairs are generated autonomously through random exploration while supervised RL uses human-guided action sequences during initial learning stages avoiding the costly random exploration.

From these techniques, we opt for supervised reinforcement learning (Conn and Peters, 2007; Zang et al., 2010), because it offers the possibility of reducing the number of learning steps by avoiding the initial random exploration of the state space. This is achieved by providing the agent with a few correct training examples and using them for off-line training.

5.3. Realization of the Docking Behaviours

We create the training examples by tele-operating the robot from several random positions to the goal position, while saving state, action and reward information. The off-line training consists of the presentation of the saved action and state vectors (or action sequences) to the agent. Thus, the agent can learn the given action sequences without additional real-world execution of actions. Since the training examples represent only a reduced subset of possible solutions, we use additional reinforcement learning to safely operate the robot around the near-optimal solutions provided by the operator. Specifically, we use SARSA learning, which is a classical on-policy algorithm for TD-learning (Sutton and Barto, 1998, p. 145).

In order to limit even more random exploration and to achieve efficient real-world reinforcement learning, we introduce an additional modification that boosts the learning speed. Instead of using a single active state at a given time, as conventionally used in reinforcement learning techniques, we use a Gaussian activation of state units (Foster et al., 2000): a Gaussian is centred around the current robot state; see Figure 5.5.

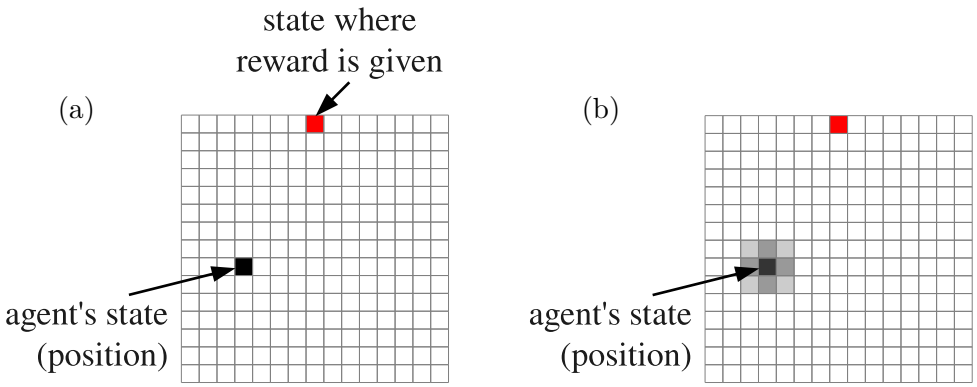


Figure 5.5 – 2-dimensional grid-world example with two forms of state representation. The goal position is indicated by a red cell. Gray-scale indicates state activation. (a) Single state activation at the agent's position. (b) Gaussian-distributed state activation. The spread of activation to neighbouring states speeds up learning.

One motivation for a Gaussian state activation is that neighbouring states to the current state should often generate the same action. Using this concept, we can extend and spread what we know about a state to *neighbouring* regions of the state space. This differs from eligibility traces that allow faster on-line learning by strengthening states *recently* visited, see Section 4.5. Batch learning or repetitive off-line training, though, incorporates the effect of eligibility traces.

The model has an input layer, which represents the agent's current state, and an output layer, which represents the chosen action. Both layers are fully connected (see Figure 5.6). The number of states, actions and the size of the actions are adjusted empirically as a trade-off between speed and accuracy for each of the tested docking behaviours. The algorithm implementation will be explained using a grid-world example, which offers an intuitive ground and facilitates graphical representation of the modifications that are being introduced.

The navigation problem is modelled as a Markov decision process (MDP). An MDP is defined by a set of states S, a set of actions A, a transition model $P(s'|s,a)$ that specifies the probability of reaching the next state s' by taking action a in state s, a reward model $R(s',s,a)$ that specifies the immediate reward received when taking action a in state s, and an exploration policy $\pi(s|a)$, which is a mapping from states to actions.

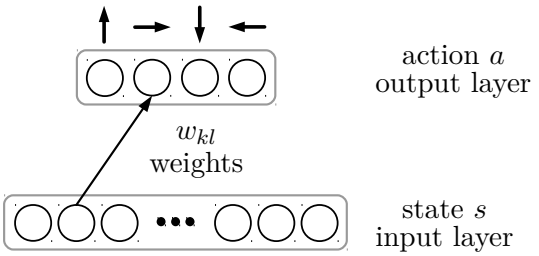


Figure 5.6 – Neural network schematic overview. For clarity, only one connection weight is shown (thin arrow between neuron layers).

Considering the 2-dimensional grid-world example shown in Figure 5.5(a), the state space S is formed by all cells. The goal position is indicated by a red cell and the current agent's position by a black cell. The agent's objective is to reach the rewarded goal position as quickly as possible.

The actions are moving `UP`, `DOWN`, `LEFT` and `RIGHT`. A move does not depend on the history but only on the policy $\pi(s|a)$, which depends on the learnt network weights $\mathbf{w}$. A binary reward r is used to indicate whether the agent has succeeded or not. The agent is given $r = 0$ as long as the desired position is not reached. Once the goal position is reached, the agent receives $r = 1$ and the "trial" is finished.

The learning algorithm is based on SARSA (Weber et al., 2008; Sutton and Barto, 1998, p. 145) and can be summarized as follows. For each trial the robot is placed at an initial random position within the defined workspace. The agent reads the cell's coordinates to obtain the internal state activation vector s, with all entries zero except for the entry that corresponds to the world position.

The net activation h_i of action unit i is computed as

$$h_i = \sum_l w_{il}\, s_l \ , \tag{5.1}$$

where w_{il} is the connection weight between action unit i and state unit l.

For the particular case that only one state unit l^* is activated, Eq. (5.1) becomes

$$h_i = w_{il^*}\, s_{l^*} = w_{il^*} \ . \tag{5.2}$$

Connection weights w_{il} are initially set to zero. Next, we used a softmax-based stochastic action selection policy

$$P_\beta(a_i = 1) = \frac{e^{\beta h_i}}{\sum_k e^{\beta h_k}} \ , \tag{5.3}$$

where β controls how deterministic the action selection is. Large β implies a more deterministic action selection or a greedy policy. Small β encourages the exploration

of new solutions. We use $\beta = 500$ to bias exploitation of known routes and to conservatively explore unforeseen policies.

Based on the active state (l^*) and on the current selected action (k^*), i.e. $a_{k^*} = 1$; $a_{i \neq k^*} = 0$, the current estimate value $Q(s, a)$ is computed:

$$Q(s, a) = \sum_{k,l} w_{kl}\, a_k\, s_l \ . \tag{5.4}$$

For the particular case of SARSA, where only one single state l^* and one action k^* can be active at a time, this becomes

$$Q(s, a) = w_{k^* l^*}\, a_{k^*}\, s_{l^*} = w_{k^* l^*} \ . \tag{5.5}$$

The old state-action value $Q(s, a)$ is subtracted from the time-discounted new value $\gamma Q(s', a')$ to yield the network prediction error δ. The time-discount factor $\gamma \in [0, 1]$ controls the importance of proximal rewards against distal rewards. Small values are used to prioritize proximal rewards. In contrast, values close to one are used to equally consider all rewards. Considering also the binary reward $r \in \{0, 1\}$, the prediction error is computed as

$$\delta = \begin{cases} \gamma Q(s', a') - Q(s, a), & \text{if } r = 0, \\ r - Q(s, a), & \text{if } r \neq 0. \end{cases} \tag{5.6}$$

Eq. (5.6) differs from $\delta = \gamma Q(s', a') + r - Q(s, a)$, as presented by Sutton and Barto (1998, p. 145), from which the name SARSA originates. Our modified version works with binary reward schemas and permits an intuitive implementation. We set $\gamma = 0.65$.

The weights are updated using a δ-modulated Hebbian rule with learning rate $\epsilon = 0.5$:

$$\Delta w_{il} = \epsilon\, \delta\, a_i\, s_l \ . \tag{5.7}$$

At this point, the two techniques to facilitate real-world RL, mentioned earlier in this section, come into play. First, to avoid random exploration, a set of training examples are recorded and used for off-line training. Within each trial, the learning algorithm was realized as described in Eq. (5.1)-(5.7). However, instead of using Eq. (5.3) for stochastic action selection, the selected action was provided by the tele-operation data. We refer to this procedure as *"supervised reinforcement learning"*. Second, instead of using single state activation as in Eq. (5.2), where only a single input neuron has maximal activation ($s_l = 1$) at a time, we use a Gaussian activation of state units (Foster et al., 2000): a Gaussian is centred at the current robot state (*"active state"*)

$$s_l = N \cdot e^{-\frac{(x_l - \mu_x)^2 + (y_l - \mu_y)^2}{2\sigma^2}} \ , \tag{5.8}$$

where N is a normalizing factor, i.e. the sum over the state space activations is 1. The different paradigms of state activation are shown in Figure 5.5.

We use $\sigma = 0.85$, which effectively "blurs" the activation around the *"active state"*. In this way, generalization to states that have not been directly visited is possible. The dimensionality of the Gaussian distribution will depend on the number of variables used to build the state space. In this grid-world example, we show schematically a 2-dimensional Gaussian distribution. μ_x represents the current x-cell coordinate and μ_y the y-cell coordinate of the agent.

5.4. Results from Simulations and Real Robot

5.4.1. Analysis of Results from Simulation (Grid-World)

In Table 5.1, we compare the performance of two supervised RL methods after off-line training, i.e. using *"single active state"* and using Gaussian distributed state activation. The training examples for supervised RL in both cases consist of 3 user-generated action sequences. The trajectories for the training examples include the borders and the central path and cover 15.2% of the state space. Testing was performed with 100 trials with random starting positions.

Results are shown after 300 off-line training trials, i.e. each of the 3 tele-operated example trials is repeated approximately 10 times. This number was sufficient for good performance. Training is governed by tele-operated policy π_{sup} without random exploration, i.e. without autonomous *"interaction"* with the environment. This would be appropriate to do with real-world hardware that must not run unattended. Testing is *"interactive"*, i.e. the agent action selection is governed by the learnt policy π_{sup^*}.

After training using single state activation, the agent's actions remain random in those states that have not yet been visited. This leads to a high STD of the number of steps required, see Table 5.1. After training with Gaussian state activation the agent generalizes to unvisited states and this way requires fewer steps, leading to a small STD.

We also verified the case of stochastic action selection following Eq. (5.3) for learning. The average number of steps required to solve a single trial using stochastic action selection without any prior learning is 3072. RL without any guidance or optimization would require many times this number of learning steps. In contrast, only 99 steps were performed by tele-operation, which would be required with real robot hardware. This advantage of several orders of magnitude enables real-world RL.

Table 5.1 – Performance of two supervised RL methods after 300 off-line training trials for a grid world of size 25×25. Average (Avg.) number (#) of steps, standard deviation (STD) and 95% confidence interval ($95\%CI$).

State activation	Avg. # steps	STD	95% CI
Single activation	86.74	107.59	21.09
Gaussian activation	**36.01**	38.53	7.55
Tele-operated	33.00	6.93	7.84

5.4.2. Real-World Docking Scenarios and Experimental Results

After demonstrating the effectiveness of the combination of supervised reinforcement learning and Gaussian state activation in simulation, we applied them to the two real-world docking scenarios described in Section 5.1. Results of both cases are presented in the following sections.

5.4.3. Backward Docking Station for Autonomous Recharging

Once the robot is 40 cm away from the docking position, see Figure 5.3(b), the two small landmarks placed on the docking station can reliably be detected and used for precise docking by the RL algorithm.

The state space is formed by the combination of three variables. These are the angular sizes of the two small naomarks and the yaw (pan) head angle. They encode the robot's distance and orientation relative to the docking station, respectively. The minimal allowed distance of the robot's camera to the landmark is approximately 13 cm, which corresponds to the robot's shoulder size plus a safety distance.

Those three values are discretized as follows. The angular size of each landmark within the visual field is discretized into 10 values for each landmark. These values represent distances in an interval of $[13, 40]$ cm in increments of 2.7 cm. We add one value for each landmark to indicate the absence of the corresponding landmark. This leads to a total of 11 values per landmark. The third variable is the head's pan angle. An internal routine permanently turns the robot's head to keep the interesting landmark centred in the visual field. The head movements are limited to $[70°, 120°[$ and the values are discretized with increments of 3.3° yielding 15 values. Hence, the total number of used states is obtained by the combination of all the values, i.e. $11 \times 11 \times 15 = 1815$.

The actions that the robot can perform are as follows: move forward and backward 2.5 cm, turn left or right 9° and move sidewards to the left or right 2.5 cm. The turn and sidewards movements are unfortunately very unreliable, which will be

discussed later in Section 5.5. These values are adjusted empirically as a trade-off between speed and accuracy.

We tele-operate the robot from several random positions to the goal position saving the action state vectors and reward value. This training set with near-optimal routes is used for off-line learning. Specifically, a total of 50 training examples with an average of 20 action steps are recorded. Then, using this training set, 300 trials are performed off-line, i.e. each of the 50 examples is presented 6 times. Table 5.2 summarizes the obtained results. We consider it as *success* when the robot successfully reaches the desired goal position; as *false positive* when the robot perceives to be in the goal position but fails to make electrical contact with the charging station; and as an *aborted* trial when the robot leaves from the working space or collides with the docking station. The Gaussian activation leads to more successful trials and a slightly reduced number of steps required during these trials.

Table 5.2 – Summary of 10 backward docking trials.

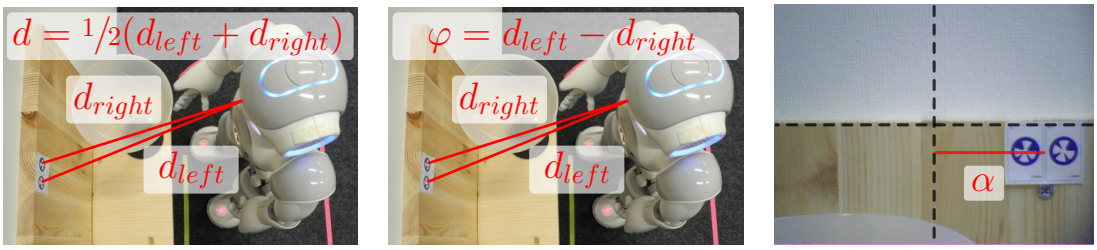
State activation	# of success	# of false positive	# of aborted	Avg. # of steps on success	STD
Single activation	6	1	3	23.80	8.23
Gaussian activation	8	1	1	**19.30**	8.35

5.4.4. Forward Docking Station for Grasping

Similar to the backward docking scenario, the state space is formed by the combination of three variables. In this case, the variables are as follows: the average distance d to the two small naomarks measured in cm (estimated from the perceived size of the landmarks), the difference φ of the perceived distance between both naomarks, and the horizontal position α measured in radians between the center of the visual field and the naomarks array; see Figure 5.7. They encode the robot's relative distance and orientation, respectively. Another difference to the backward docking is that the head remains fixed.

The three variables are discretized as follows: d is discretized into 14 values, representing distances within the interval $[15, 45]$ cm. φ is discretized into 14 values. α ranges from $[-0.35, 0.35]$ in radians and is reduced to 10 values. Hence, the total number of states is obtained by the combination of all the values, i.e. $14 \times 14 \times 10 = 1960$ states. We use the same actions as in the backward case.

We tele-operate the robot from 9 random positions to the goal position saving the action state vectors and reward value. The average number of steps required for tele-operated trials is 12. This training set with near-optimal routes is used for 300



(a) (b) (c)

Figure 5.7 – State space definition for the forward docking scenario.

off-line repetitions. We compare results obtained after 300 off-line learning trials with supervised single-state and supervised Gaussian activation. After the training phase using single state activation, the robot is able to reach the goal imitating the tele-operated routes, while the robot's actions remain random in those states that have not been visited. In contrast, after training with a Gaussian distributed state activation the robot is able to dock successfully from almost every starting point. Table 5.3 summarizes the obtained results.

Table 5.3 – Summary of 25 forward docking trials.

State activation	# of success	# of aborted	Avg. # of steps on success	STD
Single activation	13	12	71.08	59.08
Gaussian activation	23	2	**13.70**	11.19

Samples of obtained receptive fields (RFs) are presented in Figure 5.8. The goal position is shown centred in the left side of each picture. Colour intensity indicates weight strength, blue excitatory weights and red inhibitory weights. White pixels represent unlearned state-action pairs, which are the majority after training with single state activation. More intense coloured pixels represent a stronger state-action binding and thus the action is more likely to be selected when the robot is in this state. When using Gaussian activation all weights have a non-zero value, although it may be small.

5.5. Interpretation of Robot Behaviour

We notice significant performance differences in the tested real-world scenarios. Specifically, for the autonomous recharging case, side movements are used as main actions. Unfortunately, these movements were very unreliable, leading often to the

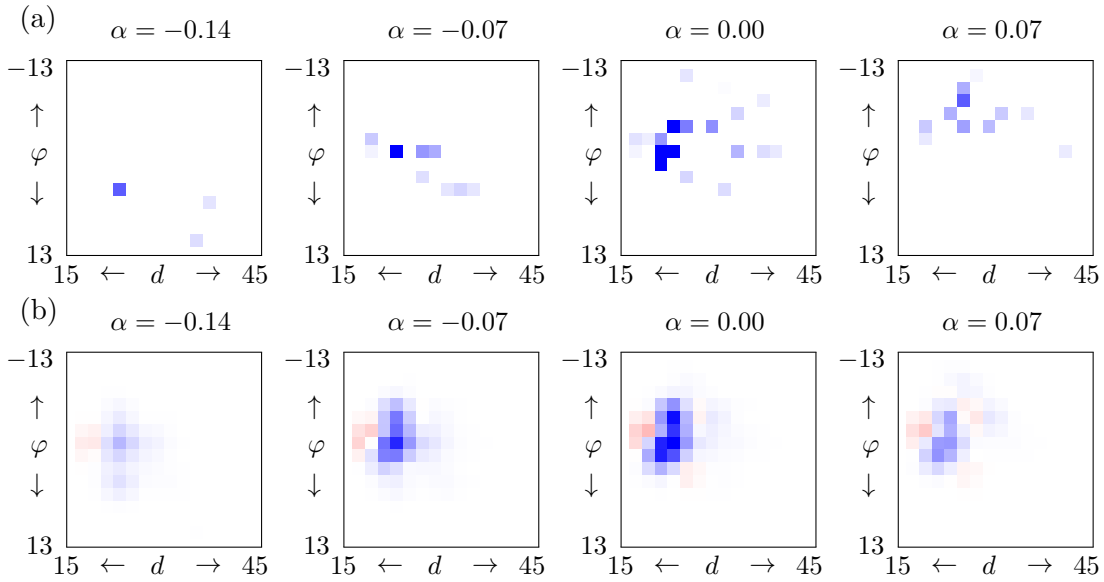


Figure 5.8 – Receptive field (RF) samples of one action unit (*Move forward*) after learning. Colour intensity represents the weight strength. Blue colors represent excitatory weights and red colors inhibitory weights. From left to right the RF samples for $\alpha \in \{-0.14, -0.07, 0, 0.07\}$ are presented. Units for d and φ are in centimetres. (a) Single state activation. (b) Gaussian state activation, $\sigma = 0.85$.

robot standing still, slight turns or even side movements to the opposite direction. Furthermore, for backward docking, we used an automatic head repositioning to keep the landmarks centred in the visual field, and we used the robot's yaw angle as one of the variables to encode in the state space, which includes motor errors. Therefore, the encoding of the state space is less precise than when keeping the head fixed and using the horizontal position of the landmark within the visual field, as done in the case of forward docking for grasping.

These two factors, inaccurate sidewards movements and less precise state space definition, contributed to a lower success rate of the autonomous recharging behaviour. This is why 50 tele-operated training examples were required to achieve acceptable results; see Table 5.2. However, the higher number of tele-operated examples implies that a larger portion of the state space has been covered, approx. 5%. This was not necessary in the case of forward docking and acceptable results were obtained using only 9 tele-operated examples, equivalent to approx. 1% coverage of the state space. This, of course, has an impact on the overall performance of supervised SARSA, but not much on supervised SARSA with Gaussian state activation. Note that we did not use any state space reduction technique.

Figure 5.9 presents a common problem in both scenarios, i.e. the effect of noisy sensory input and action execution. It shows 7 actions of a successful forward docking trial after off-line training using Gaussian state activation. The blue curve

represents a reconstruction of NAO's perception of its position and orientation, and the red curve shows the real NAO position and orientation, as obtained from a ceiling camera. The letters inside the head-like shape denote the selected action. Of special interest are the cases of wrong perception. For example in the particular case shown in Figure 5.9, when the NAO performs a step to the right from location 1 to location 2, it perceives a backward-directed movement, or, when turning left at location 6, it perceives a larger translation.

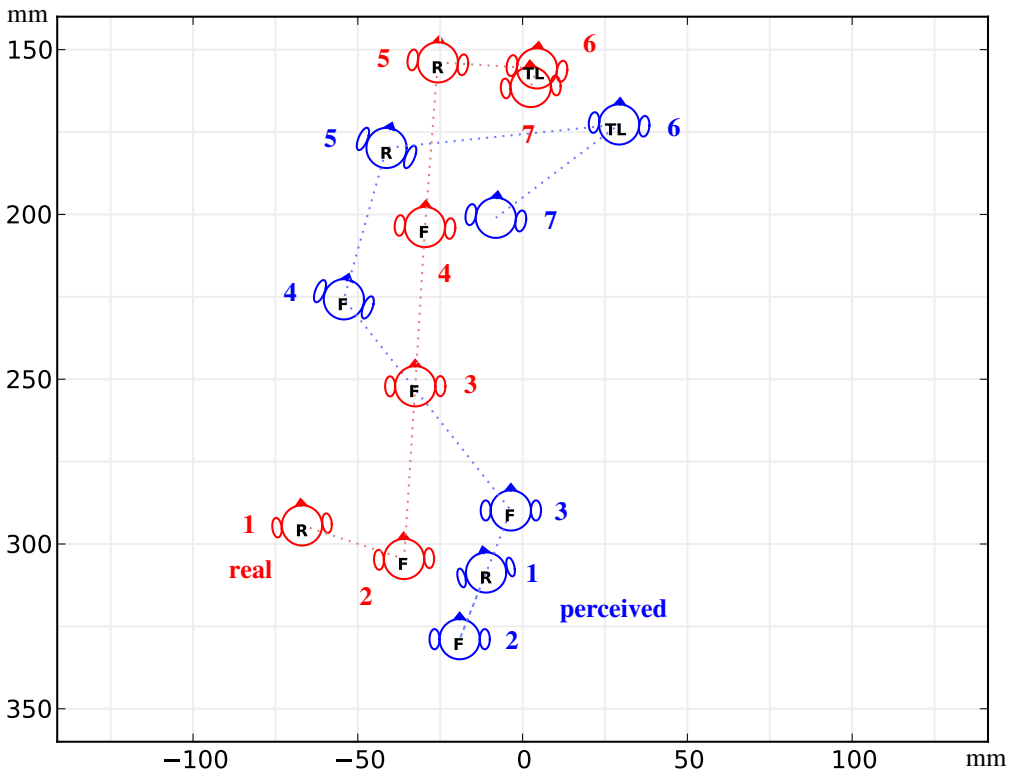


Figure 5.9 – NAO's trajectory during forward docking after being trained using Gaussian state activation. The head-like shape represents NAO's position and orientation. Letters F, B, R, L, TR and TL denote forward, backward, move right, move left, turn right and turn left respectively, starting at the shown position. Blue represents NAO's perceived positions. Red represents NAO's real positions captured by a ceiling camera.

Tele-operation creates a few representative training examples to cover substantial parts of the state space and to speed up initial learning. Insufficient sampling from the state space during training would lead to insufficient initial performance in unexplored regions. A representative training set should consist not only of the most frequent trajectories but it should particularly cover less frequently visited regions of the state space. A practical way to build a representative training set is in an incremental fashion, i.e. generate a training set, train the network and test the output placing the robot in a random position within the workspace. If the result is unsatisfactory, generate additional training examples by tele-operation containing

the difficult cases and re-train the network. These steps should be repeated until the results are satisfactory.

5.6. Discussion

Motivated by the need for autonomous recharging behaviours for humanoid robots for service robot application and the suitability of RL techniques for navigation, we developed a real-world learning algorithm based on SARSA and supervised RL. We achieved a considerable reduction in the required learning steps from several thousand to a few hundred. The use of appropriate training examples proved to be a key factor for real-world learning scenarios, i.e. a representative sampling from the state space during tele-operation will contribute to the performance of the running system.

Additionally, Gaussian distributed state activation was demonstrated to be useful for generalization and eliciting a state space reduction effect while not losing performance when applied to large state spaces. This technique reduces failures that may be induced by ambiguous or insufficiently sampled state spaces. Furthermore, the use of a memory of successful action sequences may be of considerable value in other applications. This memory could be generated independently by tele-operation or fully automated operation. Then these examples could be used for automatic off-line training, while the robot is executing less demanding tasks.

Other well-established methods for speeding up learning exist. For instance, TD(λ) accelerates learning by maintaining an eligibility trace of recently used states (in Actor-Critic learning), or state-action pairs (in SARSA), controlled by a trace decay parameter $0 \leq \lambda \leq 1$ (Sutton and Barto, 1998, p. 163). Thereby, when δ becomes large at time t, not only the current, but also more recently visited state-action pairs, prior to t, of the current trial will be affected by the update. The number of trials required for learning can thereby decrease by an order of magnitude. In our model, however, we distinguish real-world trials of the robot from the repetition of stored sequences in an off-line mode. The repetitions have an effect similar to TD(λ) reducing the number of necessary real-world trials, which is the important quantity in terms of costs. Moreover, while an eligibility trace only affects the most recent trial, repetitions affect all trials stored in memory, so dynamic programming can be performed on all stored real-world trials until convergence. Finally, the use of Gaussian activated states affects not only visited states but also neighbouring states.

The proposed method was tested in two real-world scenarios; a partially backward docking used for autonomous recharging, which the robot can perform successfully, and a forward docking for a grasping task. During the experimental phase, we

noticed that the 2-dimensional naomarks can be detected only from within a small angle range, i.e. when the robot sees them without much distortion, and detection is very noise-susceptible. Additionally, for the particular case of NAO, forward, backward and turn movements have to be preferred, because of the limited effectiveness of sidewards movements due to slippage of the NAO.

Chapter 6
Punishment and Nociception in Robot Motor Learning

The ability to learn from mistakes and successful behaviours is central to self-preservative mechanisms. On the one hand, reward such as that received from the consumption of nutrient and energy-rich food is crucial to reach energetic autonomy in an ever changing environment, as discussed in Chapter 5. On the other hand, punishment and nociception are vital for detecting and avoiding harm. However, the relative and combined effects of reward (appetitive behaviour) and punishment (aversive behaviour) on learning are not yet known (Abe et al., 2011).

6.1. Introduction

A wealth of research has identified the key brain regions involved in different aspects of reward- and punishment-driven learning, including the midbrain, the striatum, the amygdala, the orbitofrontal cortex, and the medial prefrontal cortex. Most findings shed light on the neural pathways involved in reward seeking behaviours only, however, less is known about punishment-driven learning (Kim et al., 2015; Wächter et al., 2009) and the combined effects of both types of reinforcer on behaviour learning (Dayan and Niv, 2008; Wächter et al., 2009). Evidence suggests that there are substantial neurobiological differences (Kim et al., 2015; Wächter et al., 2009). For instance, the striatum, the amygdala, and the medial OFC seem to be more involved in reward-driven learning, while for punishment-driven learning the insula or the lateral OFC play a greater role (Kim et al., 2015; Wächter et al., 2009). Moreover, Kim et al. (2015) and Wächter et al. (2009) show the differential involvement of the striatum in reinforcement learning tasks that require action execution. Specifically, the ventral striatum was said to be linked to reward anticipation, while the dorsal striatum was said to be associated with both reward and punishment anticipation and thus valence-free action value representations. These findings support existing evidence that the ventral striatum is involved in reward-driven learning, whereas the dorsal striatum is associated with the formation

of habits during reward learning (Kim et al., 2015).

During feedback receipt, punishment-driven, in comparison to reward-driven, learning, elicits a greater engagement of the prefrontal cortex and thus more attentional and cognitive resources are recruited (Kim et al., 2015). This could indicate that punishment-driven learning is a more complex and cognitively demanding process. Dayan and Niv (2008) agree with this view and argue that this may be due to the heterogeneous range of effects that aversive predictions elicit, which greatly depend on contextual information.

A behavioural study, on a procedural learning task by Wächter et al. (2009), found that reward-driven learning can lead to significantly higher performance than punishment-driven learning, where performance is measured with respect to reaction time and error rate. On the contrary, punishment did not have an effect on learning, but did have an effect on behavioural aspects, e.g. lead to an immediate reduction of the reaction time, when tested on a sequence-less version of the task (Wächter et al., 2009). These results could be considered contradictory to other studies (Hester et al., 2010), where punishment-driven learning has been found to improve learning performance, however, the nature of the tested tasks are different, i.e. procedural learning vs. associative learning, respectively. Besides, the recognized existence of differential pathways for reward- and punishment-driven learning reconciles both results.

The above-mentioned evidence motivates the research presented in this Chapter. Here, we study to what extent reward in combination with punishment and nociceptive input affects agent behavioural performance and motor skill learning capabilities during an inverse kinematics learning scenario. We chose a version of the well-established TD-learning algorithm to evaluate their suitability for capturing the differential dynamics of reward- and punishment-driven learning.

6.2. Computational Models of Learning by Feedback

Reinforcement learning and particularly temporal-difference (TD) learning are the preferred algorithms to model learning by feedback, whether this is in the form of reward or punishment. Reinforcement learning algorithms follow a trial-and-error learning paradigm and are formalized around the idea of an agent who learns from its experience, and whose task is to maximize the cumulative reward in the long-term (Sutton and Barto, 1998, p. 56). However, no special treatment is given to punishment which is simply modelled as a negative reward, without further implications (Lowe and Ziemke, 2013; Seymour et al., 2005). Yet, recent evidence indicates that this may not be the case, and that in fact punishment is processed by a different neural pathway than reward at least on procedural or skill motor

learning (Galea et al., 2015). Moreover, it may be a more demanding cognitive task than reward-driven learning (Kim et al., 2015; Wächter et al., 2009).

The use of punishment in applications of TD-learning algorithms is widespread and often used: 1) with the intention to limit the time spent in certain states or to avoid them altogether (Weber et al., 2004), and 2) with the expectation to obtain solutions with shorter action sequences (van der Wal, 2012). However, most approaches do not take into account that punishment may not be appropriately modelled by TD-learning algorithms.

Attempts to fill the gap of TD-learning models with respect to punishment-driven learning and the combined effect with reward are still scarce. A few rare examples consider different multiplicative combinations of reward and punishment. For instance, Lowe and Ziemke (2013) use independent representations of reward and punishment in a two-armed bandit navigation task. These representations are updated independently and combined into an action value function (Q_{rp}) used for action selection. Here, the value representation of reward (Q_r) is linearly modulated by the value representation of punishment (Q_p), so that Q_r is inhibited as Q_p increases. Additionally, two meta-parameters are used to influence the probability of exploration. Similarly as for the action value function, an internal representation of reward $\mathfrak{R}$ is linearly modulated by an internal representation of punishment $\mathfrak{P}$, where strong punishment inhibits behaviours associated with reward. This method encompasses many reinforcement contingencies modulated by the expected reward and punishment that are observable in context-dependent levels of exploration versus exploitation.

From the reported results it is not clear if this alternative method of combining reward and punishment, or their expected values, performs better than a naïve additive combination in the same two-armed bandit navigation task. In other words, it would be useful to know if the additional complexity added by Lowe and Ziemke (2013)'s model is needed and enough to solve the two-armed bandit navigation task or if it is better than a naïve additive combination. For many problems the additive combination of both types of feedback into a single value function is enough or at least not detrimental, inferred from the wealth of applications of TD-learning algorithms. However, in other cases, as the one presented in this Chapter, the additive combination of reward and punishment has undesirable consequences and it is useful to have additional information about the specific reinforcement contingencies, i.e. dimensions of value/valence.

There are a number of options to inform a model of specific reinforcement contingencies to aid decision making. The multiplicative combination of reward and punishment suggested by Lowe and Ziemke (2013) could well be one of them, as it resembles the behavioural dynamic produced by the differential pathway for reward- and punishment-driven learning described earlier in this Chapter.

Alternatively, Damasio (1996)'s somatic marker hypothesis (SMH) also provides a potential solution. Here, somatosensory signals or processes (somatic states) focus and influence the individual cognitive processes on a subset of relevant contingent behaviours, which may save time processing alternatives and reduce risk of carrying out stochastically selected inappropriate behaviour.

6.3. Task Description and Methodology

Despite advances in humanoid robot control there are still challenges regarding generalization and autonomous learning of new tasks. One of these tasks is robust object reaching. Although this task is actively studied, e.g. Stahlhut et al. (2015a), due to its number of applications for industrial and domestic robots, it is still challenging and many aspects remain to be studied such as self-calibration, adaptation, speed and force control.

Evidence from both child development research (Thelen et al., 1993) and adult novel sensorimotor task learning (Franklin et al., 2007) suggests that learning to reach does not require visual feedback, but seems to be useful for fine correction at the end of a reaching movement. Moreover, in early infancy, pre-planned motor programs for reaching are not explicitly planned ahead of a movement (trajectory planning), which points at a trial-and-error learning paradigm. Reinforcement learning methods are particularly suitable for this type of learning.

Actor-Critic architectures are powerful TD-learning methods that model phasic changes in dopamine neuron activity (Suri, 2002). The Critic guides the learning of action sequences generated by the Actor in order to maximize the accumulated reward. The dual memory structure, one for the Critic and one for the Actor, allows storing the learned policy explicitly, which significantly reduces computation of action selection of large state and action spaces, when compared to other TD-learning methods (Sutton and Barto, 1998, p. 153). Moreover, Actor-Critic methods are thought to be consistent with biological evidence (Suri, 2002). This is due to the fact that the reward prediction signal of TD-learning resembles the dopamine neuron activity in the striatum. Also, connection-wise the Actor typically goes from a high-dimensional sensory input to a smaller action space, which resembles its neural equivalent, i.e. projections from the striatum to the basal ganglia output nuclei (Suri, 2002).

6.3.1. Experimental Set Up

Here the problem of autonomous learning or inverse kinematics of a single robot arm is addressed. The robot's objective is to move the geometrical centre of its end-

effector towards a target as precisely and as quickly as possible. Arm movements are controlled using motor commands relative to the current joint position, but no inverse or direct model of the arm dynamics is provided to the agent.

Because our main interest lies in the effects of punishment and nociception on the learning of motor skills, a number of simplifications are made. A simplified 2-degrees-of-freedom model of a NAO robot is used, i.e. restricted to only one shoulder and one elbow joint. The link lengths are 105 mm for the upper arm, and 113.7 mm in total for the lower arm and hand[1]. The shoulder joint is limited to the range $[-18, 76]$ degrees and the elbow joint is limited to the range $[-88.5, 0]$ degrees[2]. The robot is able to precisely perceive the target's position in an egocentric reference frame, i.e. exteroception. It can also precisely perceive the absolute angular position of its joints, i.e. proprioception. It can perceive when the joints are at or close to their upper or lower limits, i.e. nociception. Nociceptive input is maximal when a joint is at the mechanical limit and decreases exponentially as the joint moves away from the limit. Nociception is perceived only when the current joint position is within the upper or the lower 10% of its mechanical range. Reaching is considered successful when the robot's hand is at least 25.4 mm away from the target. This distance approximately corresponds to half the length and the width of the robot's hand and such an error still allows for a successful grasping.

To compare different learning conditions a unique training and test set for all conditions was used. The training and test sets are 1000 and 100 samples large, respectively. Each sample consists of a target in Cartesian coordinates and an initial joints configuration in degrees. Samples are randomly generated and the resulting end-effector positions are at least twice the reaching threshold of 25.4 mm apart. Training and test samples are always presented in the same order. The complete presentation of the training set is termed training session. Before any learning is performed, the agent is tested on the test set, and after each training session afterwards. Figure 6.1 shows both the training and test set used.

6.3.2. Continuous Actor Critic Learning Automaton (CACLA)

CACLA (van Hasselt and Wiering, 2007) is a model-free reinforcement learning algorithm with Actor-Critic architecture. This algorithm was designed to work with large and continuous state and actions space, thus an excellent alternative to learn the problem described in Section 6.3.1. These characteristics are obtained through the use of function approximation techniques such as feed-forward multilayer perceptron neural networks (MLP) that allow generalization, see Section 4.1.

[1] `http://doc.aldebaran.com/2-1/family/nao_h25/links_h25.html`
[2] `http://doc.aldebaran.com/2-1/family/nao_h25/joints_h25.html`

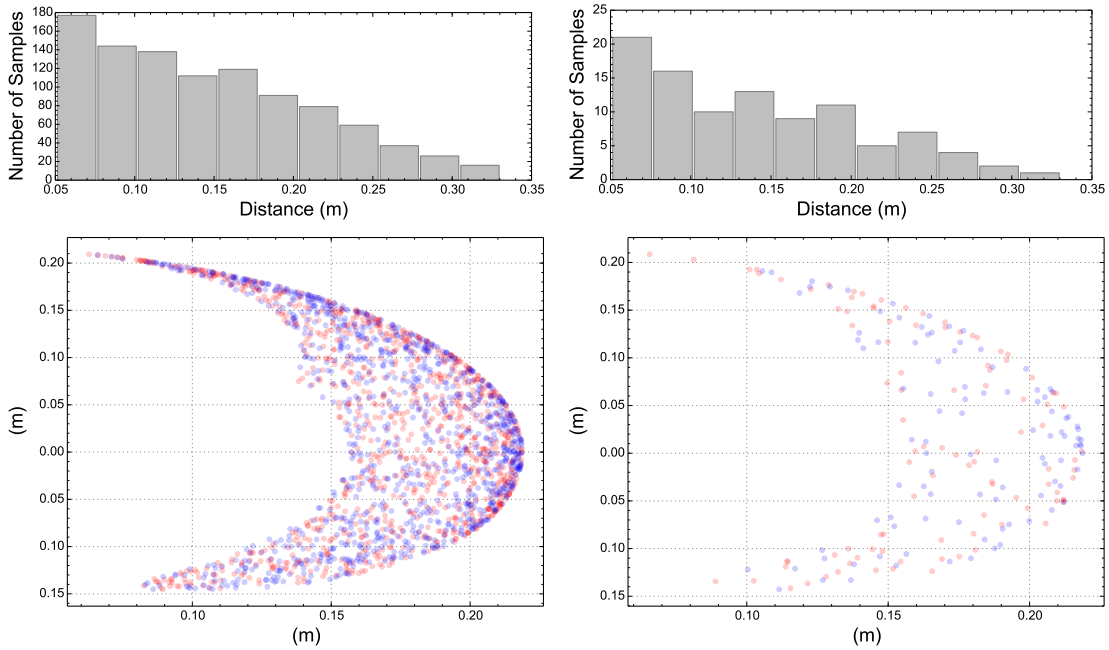


Figure 6.1 – Depiction of target and end-effector coordinates of the randomly generated training and test sets. Blue dots represent targets, whereas red dots represent end-effector initial positions. The histograms show the initial distance between the end-effector and the corresponding target. Left hand side, training set. Right hand side, test set.

Actor-Critic methods are on-policy temporal-difference (TD)-learning methods that have two memory structures, i.e. a dedicated memory for policies and another for value functions, see Section 4.4.3. The Actor represents the policy and this is denoted as $A(s)$. The Critic provides a state-value function $V(s)$. The Critic evaluates the outcome of the selected action against its existing value estimate (expectation) and generates a TD error to the extent that it differs, see Eq. (6.1). The TD-error is then used to update both the Actor and the Critic. If the error is positive, the selected action should be strengthened, whereas a negative error suggests the opposite (Sutton and Barto, 1998, p. 152). The TD-error is defined as:

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{6.1}$$

where r_t is the reward received at time t, γ is the discount factor of future rewards, $V(s_{t+1})$ is the expected reward at the state s_{t+1} and $V(s_t)$ is the expected reward for state s_t.

Action selection is based on the current policy but in order to discover new and better policies, i.e. to learn, exploration is required. We use Gaussian exploration, where the performed action is sampled from a Gaussian distribution centred at the

Actor's output $A(s_t)$. So the probability of selecting action a in time t is:

$$p_t(s_t, a) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(a - A(s_t))^2 / 2\sigma^2} \tag{6.2}$$

where π denotes the mathematical constant and σ denotes the standard deviation and is here also called exploration rate.

CACLA differs from conventional Actor-Critic systems (Sutton and Barto, 1998, p. 152) in that the magnitude of the Actor's update is independent of the size of the TD-error. The Actor is instead updated towards the explored action only when the sign of the TD-error is positive. This idea originates from the fact that punishing or moving away from an action that does not lead to a higher reward does not guarantee a better solution (van Hasselt and Wiering, 2007). Thus the Actor is only updated towards actions that improve agent performance instead of pulling the weights around without a destination. To control how strongly actions will be reinforced a derived algorithm called CACLA+var is used (van Hasselt and Wiering, 2007). CACLA+var keeps a running average of the TD-error's variance, so actions leading to unusually higher rewards are reinforced more:

$$var_{t+1} = (1 - \beta) var_t + \beta \delta_t^2 \tag{6.3a}$$

$$\text{number of updates} = \lceil \delta_t / \sqrt{var_t} \rceil \tag{6.3b}$$

CACLA+var requires two additional parameters to be tuned, i.e. var_0 which should be comparable to the typical value of δ, this is important to avoid high reinforcement rates early in learning when the agent behaviours are mostly random, and β.

Then the Actor's policy update can be expressed in pseudo-code as:

Algorithm 6.1 Actor's update

```
if δ_t > 0 then
    for i := 1 to ⌈δ_t/√var_t⌉ step 1 do
        θ^A_{i,t+1} = θ^A_{i,t} + α (a_t − A(s_t)) ∂A(s_t)/∂θ^A_{i,t}
    end for
end if
```

where $\theta_{i,t}^A$ is the i^{th} item of the parameter vector of the Actor at time t, s_t is the state vector at time t and α is the learning rate for the Actor's function approximator. Unlike the Actor, the Critic is updated every time step as follows:

$$\theta_{i,t+1}^V = \theta_{i,t}^V + \eta \delta_t \frac{\partial V(s_t)}{\partial \theta_{i,t}^V} \tag{6.4}$$

where $\theta_{i,t}^V$ is the i^{th} item of the parameter vector of the Critic at time t, and η is the learning rate for the Critic's function approximator.

6.3.3. Reward Function

The reward function consists of two parts, i.e. a rewarding component depending on the end-effector position and a punishing component depending on the joints' position, which are additively combined into a single scalar value after every step. The rewarding component is computed as follows:

$$r_t^+ = \begin{cases} R & \text{, if } d_t \leq 2.54\,\text{cm} \\ 0 & \text{, otherwise} \end{cases} \tag{6.5}$$

where R is the highest reward value, and d_t the distance from the end-effector to the target at time t.

Joint positions close to the lower or upper limit are considered harmful and a punishment signal is used to signal this. The amount each joint contributes to the total punishment per time step is computed as follows:

$$r_t^- = -\frac{P}{dof} \times \begin{cases} 0 & \text{, if } J_i^{min} + m_i < j_i < J_i^{max} - m_i \\ e^{-0.5\left(\frac{j_i - \left(J_i^{min} + m_i\right)}{m_i}\right)^2} & \text{, if } J_i^{min} + m_i \leq j_i \\ e^{-0.5\left(\frac{j_i - \left(J_i^{max} - m_i\right)}{m_i}\right)^2} & \text{, if } j_i \leq J_i^{max} - m_i \end{cases} \tag{6.6}$$

where P is the maximum magnitude of punishment, dof the total number of degrees of freedom, j_i the absolute angular position of the i-th joint at time t. J_i^{min} and J_i^{max} are the minimum and maximum possible angular position of the i-th joint, and m_i is the margin of safety for safety factor of 0.1 for the i-th joint.

6.3.4. Neural Architecture

We use two MLPs one for the Actor and one the Critic, see Figure 6.2, both share the same input layer, the output layer for the Actor has as many units as degrees of freedom, whereas the Critic has a single output unit, the rest of the layout is determined separately. The input layer is divided into three perceptual modalities. Firstly, there are two exteroceptive units which encode the Cartesian coordinates of the target in a 2-dimensional task space relative to the robot. Secondly, there are two proprioceptive units that encode the angular position of each of the joints of the robot's arm, i.e. absolute joint value of the shoulder and elbow joint. Finally, two nociceptive units associated with each robot joint, with an activation almost identical to function of punishment, see Eq. (6.6), however, here punishment triggered by movements towards the lower limit of a joint are negative and positive for movements towards the upper limit.

All input values are scaled to the range [-1, 1]. The squashing function for the output units is linear, and for all other units a custom hyperbolic tangent as defined by LeCun et al. (1998, 2012) is used, see Section 4.3.2.2. Weight initialization is also performed as defined by LeCun et al. (1998, 2012), see Section 4.3.2.3. Bias units with value -1 are always used. Momentum and learning rate decay are not used. Both networks, the one for the Actor and the one for the Critic, are trained using back-propagation, see Section 4.3.

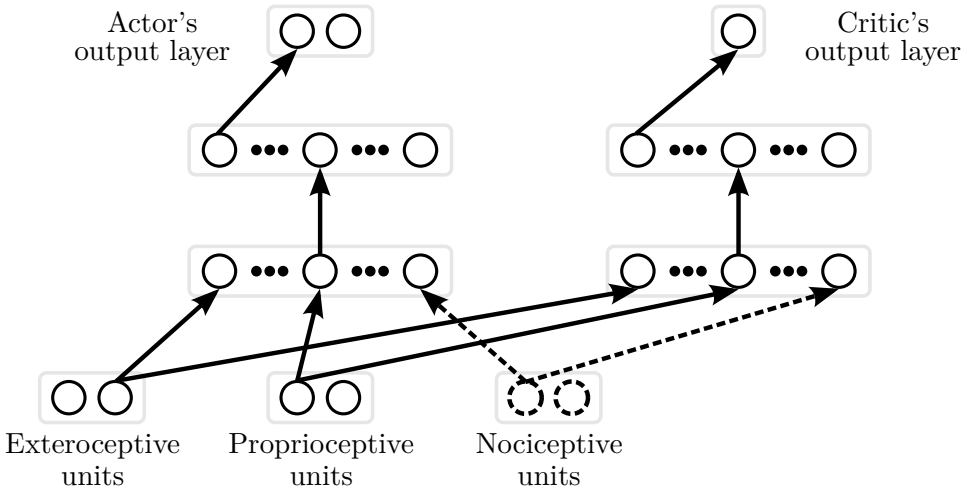


Figure 6.2 – Neural architecture used for inverse kinematics learning. For clarity, only one connection weight is shown (arrow between neuron layers). The hidden layers for both the Actor and the Critic are independently tuned. Dashed units and connections weights are only considered under the *reward and nociceptive input*, and *reward, punishment and nociceptive input* conditions.

6.3.5. Hyperparameter Optimization

Due to the large number of possible combinations of hyperparameters the systematic and exhaustive testing of them is impractical. Thus, we decided to use a genetic algorithm (GA) to explore the hyperparameter space, which helps to discover novel solutions and to determine what hyperparameters have the greatest influence on performance. The hyperparameters subject to evolutionary optimization and the search space for each of them are detailed in Table 6.1. Because small changes in the hyperparameters are likely to produce little change in performance we decided to discretize their values and thus the search space as indicated in Table 6.1.

Regarding the GA, a small randomly initialized population is used due to the computational cost of large populations, for practical reasons we chose 32 individuals per generation which corresponded to the number of cores we had available for parallel computation. Also because the exploration achieved with

Table 6.1 – List of hyperparameters for CACLA and MLP subject to evolutionary search.

Parameter name	Symbol	Search space
CACLA+var beta	β	$\{k : k + 0.0001, 0.0001 \leq k \leq 0.01\}$
Initial variance	var_0	$\{k : k + 0.1, 0.1 \leq k \leq 10\}$
Initial iterations	$\lceil \delta_0 / \sqrt{var_0} \rceil$	$\{k : k + 1, 1 \leq k \leq 20\}$
Discount factor	γ	$\{k : k + 0.001, 0.75 \leq k \leq 1.0\}$
Exploration rate	σ	$\{k : k + 0.1, 0.2 \leq k \leq 2.0\}$
Exploration rate decay	κ	$\{k : k + 0.001, 0.90 \leq k \leq 1.0\}$
Learning rate Critic	α	$\{k : k + 0.01, 0.01 \leq k \leq 0.2\}$
Critic MLP in→h1st	$C_{in \to h1st}$	$\{k : k + 5, 10 \leq k \leq 50\}$
Critic MLP h2nd→out	$C_{h2nd \to out}$	$\{k : k + 5, 0 \leq k \leq 25\}$
Learning rate Actor	η	$\{k : k + 0.01, 0.01 \leq k \leq 0.2\}$
Actor MLP in→h1st	$A_{in \to h1st}$	$\{k : k + 5, 10 \leq k \leq 50\}$
Actor MLP h2nd→out	$A_{h2nd \to out}$	$\{k : k + 5, 0 \leq k \leq 25\}$
Reward	R	$\{1, 10, 100\}$
Punishment	P	$\{-1, -0.1, 0\}$

larger populations in the initial generation can also be obtained by using operators such as crossover and mutation (Schaffer et al., 1989). Elitism is used to preserve the best four solutions, in addition, a mutated copy of the best solutions are added to the next generation to explore promising parameter combinations more effectively. To foster exploration and reduce the likelihood or premature convergence, two new randomly generated individuals are introduced every generation. The remaining 22 individuals are selected using *Tournament selection*, recombined using a single-point crossover, and finally mutated. Tournament selection is a simple selection method with an adaptable selection pressure, i.e. low when fitness distribution is high and vice versa, which also helps prevent premature convergence (Mitchell, 1998). Single-point crossover is also chosen due to its simplicity and efficacy with short genome encoding (Mitchell, 1998). A normally distributed mutation was used to explore the neighbourhood of tested solutions but also allowed a certain degree of exploration. Each gene is mutated with a 10% probability and a sigma of 6.25% of the corresponding hyperparameter range, both percentage values were manually tuned.

The fitness function for the GA consists of the total distance between the robot's end-effector and target on the testing set after learning, i.e. after the last training session. Thus, here, the lower the fitness values the better. Eq. (6.7) shows the

mathematical formulation of the fitness function:

$$D = \sum_{i=1}^{p} d(h_i, t_i) \tag{6.7}$$

where p represents the total number of testing pairs, h_i corresponds to the initial joint positions of the arm for testing pair i, t_i corresponds to the coordinates of the target for testing pair i and d is the final Euclidean distance between the arm's end-effector and the corresponding target.

6.4. Experimental Results

Figure 6.3 and 6.4 show detailed maps of fitness values over time for the four conditions tested in this work, i.e. agent with *reward only*, agent with *reward and nociceptive input*, agent with *reward and punishment*, and agent with *reward, punishment and nociceptive input*. The population per condition and per generations is 32 individuals. Evolution is carried out for 50 generations after which convergence was observed for all tested conditions. The best solution of each condition is verified by training it 10 times with different initializations. All solutions show a robust behaviour regardless of the initializations.

When no punishment is used, i.e. the *reward only* and the *reward and nociceptive input* conditions, the fitness score remains high over a large number of generations, visualized according to the extent of darker 'pixels' in Figure 6.3, but finally, in all four conditions the best individuals reach a similarly good fitness score after convergence, see Table 6.2. However, when observing the solutions more closely a different picture arises.

At a macroscopic level, i.e. at the evolutionary level, solutions for conditions with punishment, when compared to conditions without punishment, are more stable with respect to changes in fitness scores and the best solution spread rapidly over the population as observed in Figure 6.4, overall a larger number of lighter 'pixels' in Figure 6.4 than in Figure 6.3. This could be interpreted as an advantage for the conditions that use punishment, because a fewer number of generations is needed to reach a low fitness score. However, at the single solution level, i.e. at the individual level after learning, agents trained in the conditions without punishment, when compared to agents trained in conditions with punishment, reach a lower reaching error early on during training and require a smaller number of steps, a more detailed discussion of these differences will be given in the coming sections.

Figure 6.5 shows the 30 best unique solutions over all generations for the tested conditions. The hyperparameter values from Table 6.1 are normalized with respect to their corresponding range allowing us to use a common y-axis in Figure 6.5.

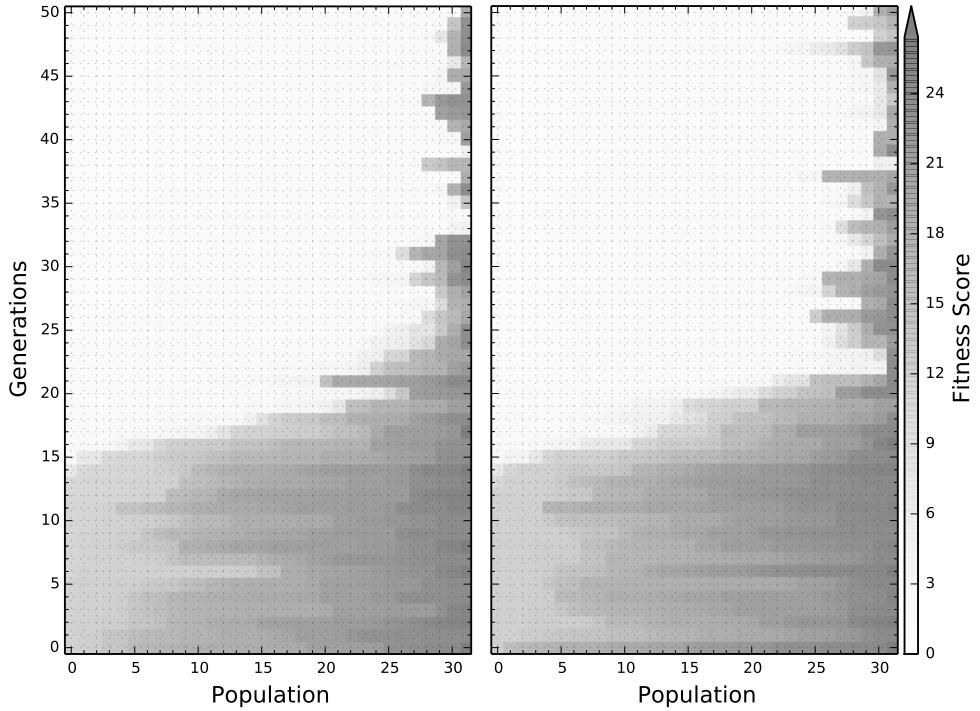


Figure 6.3 – Fitness distribution in populations trained without punishment. The fitness is directly computed from the total distance to the target, thus the lower the value the better. Left: evolution for agents trained only using reward. Right: evolution for agents trained with *reward and nociceptive input*.

Table 6.2 – Summary of fitness scores of generation number 50. The fitness is the total reaching distance, in meters, on the testing set, thus the smaller the better.

Condition	Best Fitness	Avg. Fitness	STD
reward only	1.3681706429	2.2435908394	3.0012820775
reward and nociceptive input	1.4012037080	2.8870844357	4.3544947841
reward and punishment	1.7929313888	4.4355310490	5.5794189903
reward, punishment and nociceptive input	1.3944763185	3.2573050047	4.4638217090

This means that the value 0.0 in Figure 6.5 represents the minimum allowed value for a given hyperparameter, e.g. the minimum allowed value for Reward is 1 which will be represented as 0.0 in Figure 6.5, whereas the maximum value of 100 will be represented as 1.0.

Optimal hyperparameter values for agents trained with *reward only*, *reward and nociceptive input*, and *reward, punishment and nociceptive input* follow comparable

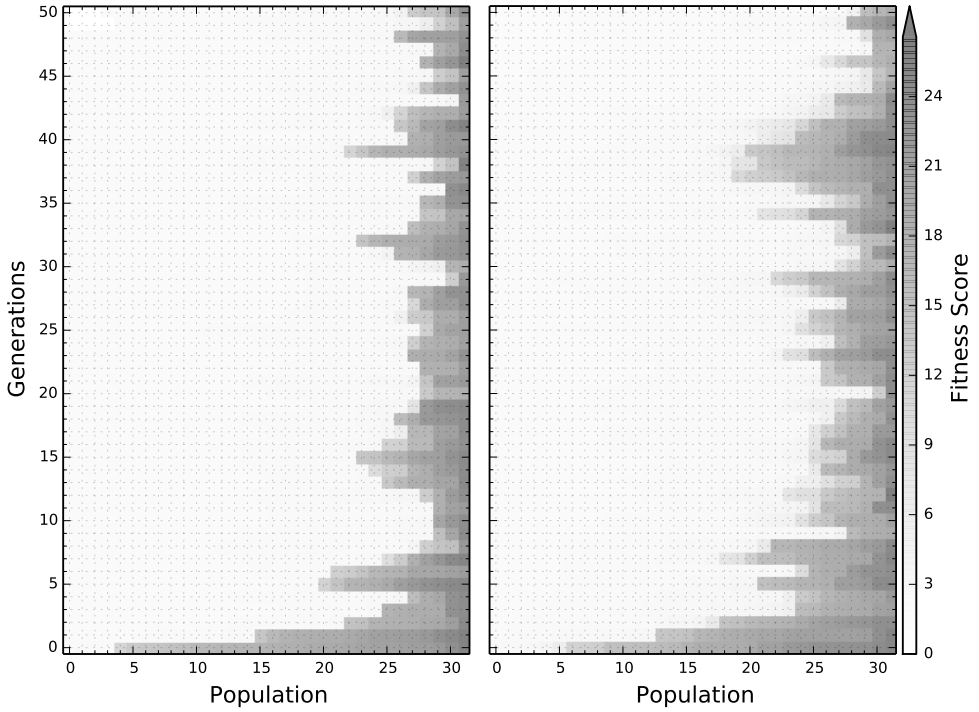


Figure 6.4 – Fitness distribution in populations trained with punishment. The fitness is directly computed from the total distance to the target, thus the lower the value the better. Left: evolution for agents trained using *reward and punishment*. Right: evolution for agents trained with *reward, punishment and nociceptive input*.

patterns, see Figure 6.5. On the other hand, hyperparameter values for agents trained with *reward and punishment* tend to cluster around the mid value of the corresponding range. The best solutions shown in Figure 6.5 were explored by the GA in all conditions, but in all cases this leads to higher reaching error, meaning that the set of hyperparameters that is better suited for a condition does not guarantee good results in another condition.

The exploration rate decay κ is only used in the *reward and punishment* condition as a value of 1.0 in Figure 6.5 translate to no decay. For the other three, conditions CACLA+var β, the discount factor γ, and learning rate for the Critic α can take any value within the defined range. The learning rate for the Actor η, on the other hand, seems to be more crucial and requires small learning rates.

6.4.1. Effect of Reward on Learning

Figure 6.6 shows the typical performance of the best individual over 100 training sessions only using reward as described in Section 6.3.3 and without nociceptive

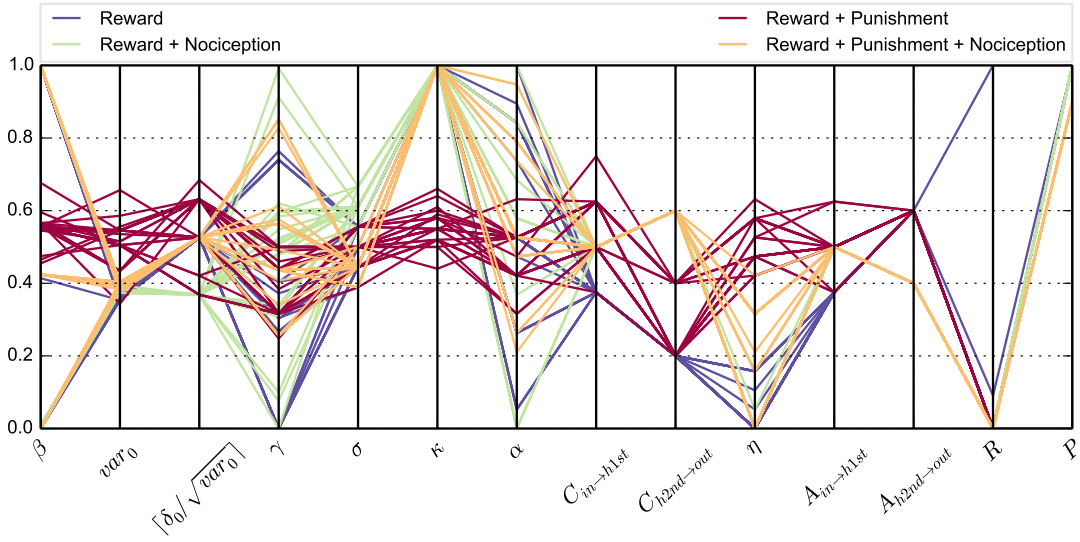


Figure 6.5 – Parallel coordinates plot of the best solutions for all tested conditions. Contrary to all other tested conditions, solutions for agents trained with *reward and punishment* are set with parameter values around the centre of the corresponding range. Hyperparameter values are normalized with respect to their corresponding range allowing the use of a common y-axis.

input. The average distance to targets in the testing set falls within the desired maximum error just after the 3rd training session requiring on average only 3 steps. A few targets in the testing set cannot systematically be reached with the required precision, but there are multiple instances where the distance is barely about the desired precision. It is possible that these hard-to-reach targets may be reached if the maximum number of steps is increased. Figure 6.6 also shows that reinforcement learning, unlike supervised learning, is not predisposed to overfitting and behaves very well under long term training sessions.

6.4.2. Effect of Punishment on Learning

When *reward and punishment* feedback are used to train the agent, a much slower reduction of average distance to targets in the testing set is observed, see Figure 6.7. At least 29 training sessions are needed for the average to fall within the maximum desired error, however, the standard deviation stays comparably higher than when using *reward alone* and the outliers are typically beyond twice the maximum desired error. Also the number of steps needed to reach the target is considerably higher than when using *reward alone* and remains high for over three quarters of the training sessions.

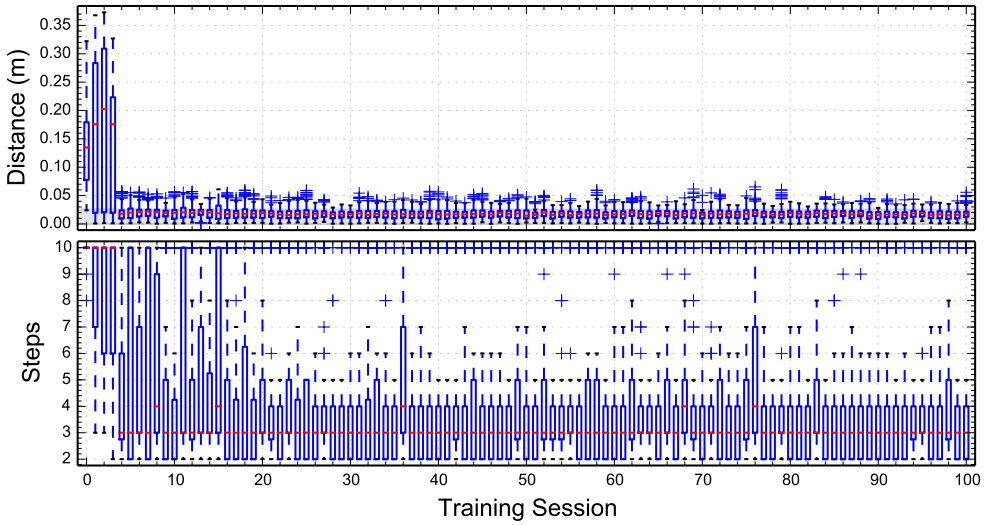


Figure 6.6 – Performance of the best individual trained only with reward. The average distance to targets in the testing set falls and stays within the maximum accepted error just after the 3rd training session. Few testing targets cannot systematically be reached with the expected precision. Most targets can be reached with only 3 steps and some targets can be reached with as little as 2 steps.

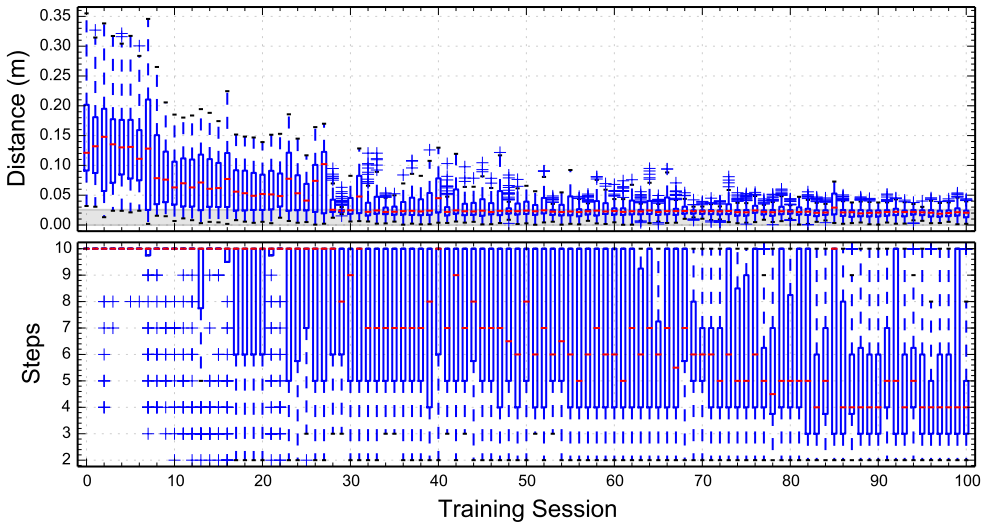


Figure 6.7 – Performance of the best individual trained with *reward and punishment*. The average distance to targets in the testing set drops below the maximum accepted error after the 29th training session. Even after 100 training sessions many targets cannot be reached with the desired precision. Also the number of steps needed to reach the target has a high dispersion.

6.4.3. Effect of Nociception on Learning

When observing the best fitness information, see Table 6.2, agents trained with *reward and nociceptive inputs* perform slightly worse than agents trained under the *reward only* condition. However, these results consider only the performance after the last training session. Yet, when observing the development of the reaching error during training, the results seem to be slightly better for *reward and nociceptive inputs* than for *reward only*.

The average distance to targets in the testing set falls and stays within the maximum accepted distance just after the 3rd training session just as for the case of *reward only*, see Figure 6.8. The number of steps needed is as low as for agents trained under the *reward only* condition. However, what sets both conditions apart is the fact that the best agent trained with *reward and nociceptive inputs*, unlike the best agent of the *reward only* condition, reached all targets of the testing set with the desired precision at least a few times, see training session 66 and 90 in Figure 6.8.

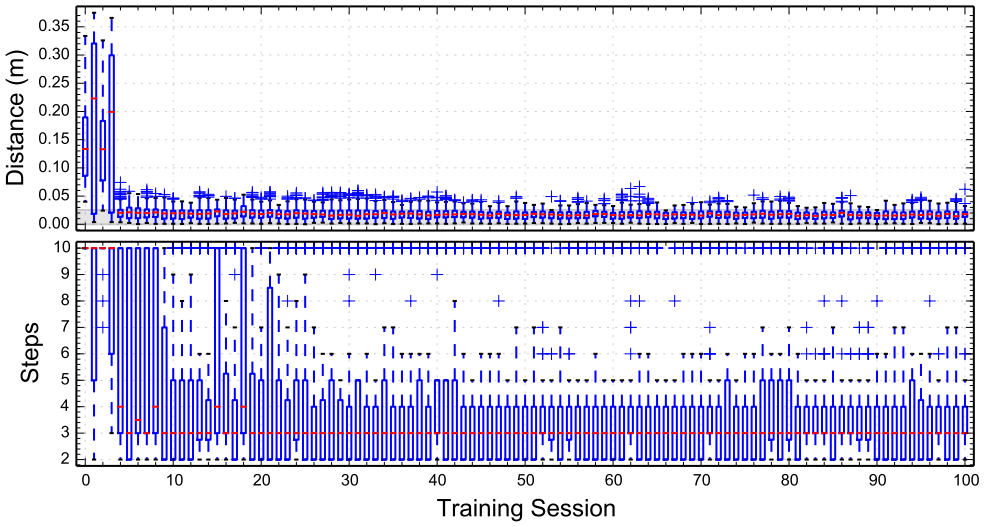


Figure 6.8 – Performance of the best individual trained with *reward and nociceptive input*. The average distance to targets falls within the maximum accepted error just after the 3rd training session. The distance to all targets in the testing set reached the desired precision at training session 66 and 90. The average number of steps needed is as low as for agents trained under the *reward only* condition but with a slightly smaller distribution here.

6.4.4. Combined Effect of Punishment and Nociception

Agents trained with *reward, punishment and nociceptive input* have a similar performance to agents trained with *reward only* or *reward and nociceptive input*, see Figure 6.9. Although the average distance to testing targets fall the fastest of all tested conditions, the distance distribution over testing set is often up to three times the maximum desired distance. There are a few instances where all distances are very close to the desired precision. The average number of steps is low, but the overall performance is in between the *reward only* and the *reward and punishment* conditions.

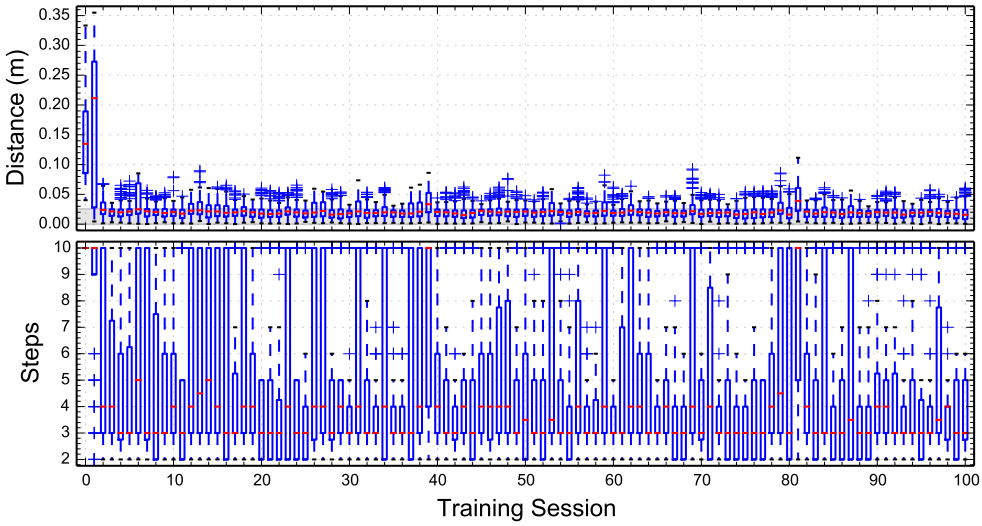


Figure 6.9 – Performance of the best individual trained with *reward, punishment and nociceptive input*. The average distance to targets drops below the maximum accepted error just after the 2nd training session. The average number of steps is also low but with a high dispersion.

6.5. Discussion

Although, the difference in performance for the best solutions across conditions is not large. The *reward only* condition performs better than any other condition based on the results for the best fitness score, the mean fitness and the standard deviation. Yet, these results only consider the performance after the last training session.

A more detailed comparison among the best solutions for each condition shows a clearer picture. The best results are for the *reward only* and the *reward and*

nociceptive input conditions. However, the *reward and nociceptive input* condition performs slightly better, as it is the only condition that reaches all targets in the testing set with the desired minimal precision at least 2 times during training.

Overall, the presented results suggest that the nociceptive input has a positive effect on learning, which is highlighted when used alongside punishment. The positive influence manifested as improved behavioural performance and learning speed in motor learning tasks, which can be observed with respect to two aspects: 1) the average distance and required number of steps to reach the testing targets decreases at least as rapidly as for agents trained with reward only, and 2) the distribution of both distance and number of steps for all samples in the test set are smaller, hence leading to reduced training and execution time.

On the contrary, punishment makes learning slower and reduces performance. Specifically, more training sessions are required for the average distance and number of steps to reach comparable levels to those observed in conditions where no punishment was used. Moreover, the distribution of distance and number of steps remains large during all training sessions. These results are in agreement with results presented by Wächter et al. (2009).

Computationally, the negative effect of punishment may be attributed to a loss of predictive power of both reward and punishment when combined into a single scalar value representation (Lowe and Ziemke, 2013). Moreover, it would be beneficial to have a separated representation of reward and punishment, or expectations thereof, which could be used for sophisticated context-specific interactions between appetitive and aversive predictions (Lowe and Ziemke, 2013). This is consistent with the evidence presented at Section 6.1, where we described the existence of two differential pathways to process reward and punishment signals.

Although, we did not model this differential pathway, we proposed an alternative and simple way of preserving or enhancing the predictive power of reinforcement signals via the use of nociceptive inputs. This nociceptive input consistently and significantly improves the overall learned behaviour particularly when the reward and punishment are additively conflated into a single value.

Taking into account the design consideration for TD-learning algorithms, of maximizing the cumulative reward (Sutton and Barto, 1998, p. 56), and growing evidence from biology of the existence of differential processing mechanisms for reward and punishment (Galea et al., 2015; Kim et al., 2015; Wächter et al., 2009), we believe that the use of punishment and nociceptive input in TD-learning applications related to procedural or motor skill learning should be reconsidered and further studied. We believe that further study of the combined effect of different reinforcement signals and sensorimotor input can help us refine biological models of learning by reinforcement, and has the potential of greater computational effectiveness for TD-learning methods, as shown is this work.

Others aspects that may be affected by punishment and nociceptive input such as robot pose and the amount of punishment received along a given trajectory could be studied. It would be also interesting to see how well does this effect scale to more complex problems such as on a real redundant humanoid robot arm. Although we did contribute with a reliable and easy to implement method to provide negative feedback into motor skill learning tasks, we did not explore ways to implement the differential pathway for reward- and punishment-driven learning, which would likely shed light on the interaction of both types of feedback in the striatum.

Chapter 7

A Neurocomputational Model for Event Anticipation

The last mechanism of adaptive self-preservation studied in this thesis is the anticipation of aversive events. Anticipation of aversive events is well studied in the field of psychology, and neuroscience under the name of (fear) conditioning, see Section 2.5.

Computational fear conditioning has experienced a growing interest over the last few years, on the one hand, because it is a robust, and quick learning paradigm that can contribute to the development of more versatile robots, and on the other hand, because it can help in the understanding of fear conditioning, and related dysfunctions in animals. Fear learning involves sensory, and motor aspects (Pape and Pare, 2010), and it is essential for adaptive self-preservative systems. We argue that a deeper study of the mechanisms underlying fear circuits in the brain will contribute not only to the development of safer robots but eventually also to a better conceptual understanding of neural fear processing in general. Towards the development of a robotic adaptive self-preservative system, we have designed a neural model of fear conditioning based on LeDoux's dual-route hypothesis of fear (LeDoux, 1992), and also dopamine modulated Pavlovian conditioning (Lowe et al., 2011). Our hybrid approach is capable of learning the temporal relationship between auditory sensory cues, and an aversive or appetitive stimulus such as pain or food. The model was tested as a neural network simulation but it was designed to be used with minor modifications on a robotic platform.

7.1. Introduction

Pavlovian conditioning is a special case of conditioning described in Section 2.5. Pavlovian fear conditioning is a form of emotional learning in which a neutral or innocuous stimulus (conditioned stimulus or CS) such as a sound or light, is paired with an aversive stimulus (unconditioned stimulus or US) such as an electric shock. Animals are evolutionarily hard-wired to rapidly acquire, consolidate, and generalize

fear memories. After only a few trials, animals quickly learn to anticipate the aversive US using the CS information, and to elicit behavioural defence responses, and associated autonomic and endocrine adjustments (LeDoux, 2007).

The amygdala (AMG) plays a crucial role in self-preservative systems by assigning biological significance to relevant neutral cues, see Section 2.4. The AMG represents the affective/emotional valence of a situation, a "state value" necessary for coordinating physiological, behavioural, and cognitive responses. Furthermore, recent evidence suggests that the human amygdala, in addition to its important role in cue fear conditioning, contributes to many reward-based decision-making tasks (Gupta et al., 2011). Understanding, and modelling these mechanisms is of great interest not only for the interpretation of neuropsychological findings, but also for computational modelling, and building safer robot assistants.

7.1.1. Related Work

A detailed literature review reveals that the most meaningful, and closely related publications in recent years are based on mathematical models of the cue-dependent fear conditioning dynamics of *acquisition*, and *extinction*, see Section 3.3. Many of these models are based on the dual-route hypothesis (cortical, and subcortical) proposed by LeDoux (Armony et al., 1995; Balkenius and Morén, 2001; den Dulk et al., 2003; Krasne et al., 2011; LeDoux, 1992; Li et al., 2009; Pavlou and Casey, 2009; Vlachos et al., 2011), which explains parallel processing of stimuli at different degrees, and temporal response improvements. Often, they use simplified binary or abstract numerical inputs (Balkenius and Morén, 2001; Krasne et al., 2011; Li et al., 2009; Lowe et al., 2011; Mannella et al., 2008; Vlachos et al., 2011) neglecting unforeseen sensory, and temporal relationships that may be relevant for fear learning dynamics.

In general, there is a lack of research on amygdala modelling with realistic sensory input taken from a realistic physical environment. In one rare example, however, Mannella et al. (2008) addressed cue conditioning in a simulated robot experiment. The model is able to reproduce, and demonstrate, with a simulated rat, experiments of first, and second order conditioning, and devaluation. Alexander and Sporns (2002) and Zhou and Coggins (2002) conducted research on prediction learning, and conditioning with real Khepera robots, but only from a normative, rather than a neurocomputationally realistic, viewpoint. These models consist of feed-forward networks with a very abstract timing model, only coarsely mapped to neurobiological circuits, and do not capture as rich a variety of dynamics as other works, see Section 3.3, but the embodied approach makes them attractive, and their relative success encourages the development of more sophisticated, and biologically plausible embodied models.

7.1.2. Suggested Approach

In this chapter, we present a biologically motivated model of auditory-cue fear conditioning. The model neurocomputationally describes the known thalamic, and auditory cortex routes (LeDoux, 2007) plus reward learning based on phasic dynamics of dopamine, previously described by Lowe et al. (2011). Here, we study fear conditioning taking into consideration bio-plausible sensory pathways, and interpretable real-world sensory input. Applications of this learning mechanism may be used in artificial self-protective systems to predict both appetitive, and aversive behavioural outcomes, or in the modulation of complex behaviours such as autonomous battery recharging, see Chapter 5, or reaching movements, see Chapter 6. This could represent an important step towards more biologically plausible computational models of embodied autonomous systems.

7.2. Biological Inspiration

The amygdala is a brain region in the medial temporal lobe composed of diverse nuclei. Since the amygdala is not a single brain structure or region it has historically been defined on the basis of connection density, chemical signature, and configuration, see Section 2.4. An initial coarse division may consist of the basolateral complex (BLA), and the central nucleus (CeA) (LeDoux, 2007). The BLA is the main input structure in the amygdala, and receives sensory information from many cortical and subcortical regions. The BLA consists of three nuclei: the lateral (LA), basolateral (BL), and basomedial (BM) also known as accessory basal (AB). Almost 80% of BLA neurons are glutamatergic cells (GLU) having multiple projections to neighbouring cells, amygdala nuclei, and other brain structures. The remaining 20% are GABAergic cells (GABA) of short axons regarded as local-circuit neurons (Pape and Pare, 2010). In contrast, the CeA is recognized as the main output component from the amygdala, modulating both cortical and subcortical structures, and controlling the selection of passive and active fear reactions (Pape, 2010). The CeA mainly GABAergic in nature can be divided into a lateral (CeL), and a medial (CeM) part (Pape and Pare, 2010). Many comprehensive reviews on the structure, connectivity, and influences of amygdaloid and fear conditioning dynamics can be found, for instance Davis (1992), LeDoux (2007), and Pape and Pare (2010).

Although fear conditioning is ubiquitous to all sensory modalities, most progress has been made on auditory-cue fear conditioning, which is why we based our model on this paradigm, see Figure 7.1.

The standard dual-route hypothesis suggested by LeDoux (1992) identifies the medial geniculate body (MGB) of the thalamus as the subcortical auditory

pathway to the amygdala. Specifically, the medial division of the MGB (MGm), and the posterior intralaminar nucleus (PIN) project to the primary and association areas of the auditory cortex, and also to the lateral nucleus of the amygdala. The MGm/PIN complex is considered as an auditory and somatosensory relay to the LA (Weinberger, 2011). The MGm is highly multimodal responding to auditory, tactile, thermal and nociceptive stimulation. With respect to auditory input, MGm lacks tonotopic organization. PIN is also multimodal. In contrast, the ventral division of the MGB (MGv) specializes in auditory stimuli, has a tonotopic organization and is identified as the main subcortical route to the primary auditory cortex (Weinberger, 2011).

More precise information about the auditory CS seems to indirectly reach the LA via the primary auditory and the associative cortex (L. R. Johnson et al., 2008). It is likely that this information includes fine frequency tuning, abstraction of pitch, and pattern discrimination, among other possible functions (Bakin and Weinberger, 1990).

The medial prefrontal cortex (mPFC) has also substantial projections to the amygdala. Although the mPFC projects to all amygdaloid nuclei, the connections to LA seem to be the more abundant ones (Pape and Pare, 2010). In turn, the BA projects back to the mPFC. Moreover, the mPFC plays a key role in extinction of fear conditioning affecting ITCm cells blocking the excitation of CeM neurons through the BA (Pape and Pare, 2010).

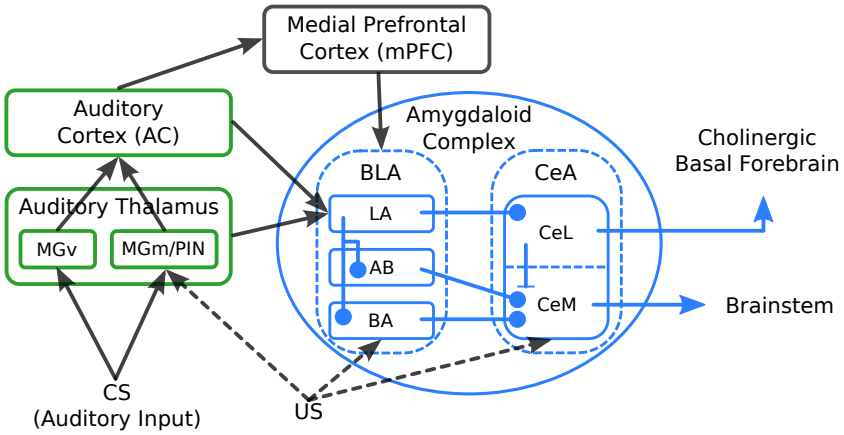


Figure 7.1 – Main inputs to the amygdala and intranuclear pathways of the amygdala involved in auditory-cue fear conditioning. MGv, ventral division or lemniscal component of the medial geniculate body; MGm, medial division of the medial geniculate body; PIN: posterior intralaminar nucleus; LA, lateral nucleus; BL, basolateral nucleus; AB, accessory basal nucleus; CeL, lateral division of the central nucleus; CeM, medial division of the central nucleus; US, unconditioned stimulus. Adapted from LeDoux (2007), Pape (2010), and Pape and Pare (2010).

Dopamine dynamics (DA) are thought to be involved in the coordination of

different stress responses. Stress-induced dopamine release allows animals to relocate attention, prioritize perceptual processing and is involved in appropriate action selection (Stevenson and Gratton, 2003). A broad body of research links stress-responsive dopamine projections from the ventral tegmental area (VTA) to the basolateral complex of the amygdala (BLA) with fear conditioning (Koob and Volkow, 2009; Stevenson and Gratton, 2003). In turn, glutamatergic projections from the amygdala to the nucleus accumbens (NAc) and medial prefrontal cortex (mPFC) regulates dopamine stress responses in NAc, mPFC and VTA. The amygdala also mediates further autonomic, endocrine and behavioural responses to emotionally significant stimuli (Davis, 1992). Figure 7.2 shows the interplay between mPFC, NAc, VTA, and BLA during stress-responsive dopamine release.

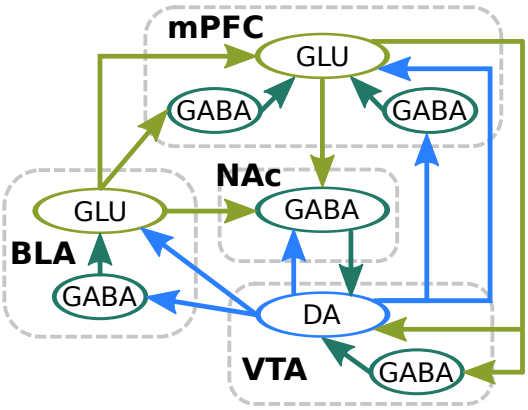


Figure 7.2 – Schematic illustration of stress-responsive projections between NAc, mPFC, and BLA that are involved in fear conditioning. NAc, nucleus accumbens; mPFC, medial prefrontal cortex; VTA, ventral tegmental area; BLA, basolateral complex of the amygdala; GLU, glutamatergic cells; GABA, GABAergic cells; DA, dopaminergic cells. Adapted from Stevenson and Gratton (2003).

There are two types of dopamine dynamics that influence the coordination of different stress responses and reward-driven learning. Slow changes of dopamine concentration are known as tonic dynamics of dopamine and are essential for the regulation of synaptic plasticity (Atcherley et al., 2015). Whereas rapid changes of dopamine concentration are known as phasic dynamics of dopamine and are associated with salient stimuli (Atcherley et al., 2015).

7.3. Methodology and Realization

The overall architecture, shown in Figure 7.3, is intended to capture both phasic dynamics of dopamine, and input and output pathways underlying fear conditioning learning. This architecture combines Hebbian components (blue

modules) for association, and a recurrent component (green modules) for reward prediction. It was designed to be portable to a real NAO robot[1] working in a home-like environment.

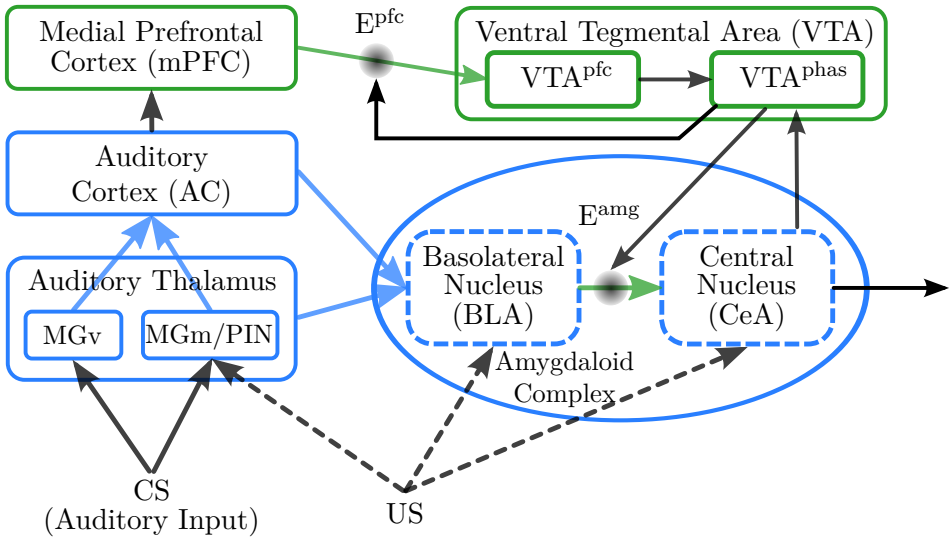


Figure 7.3 – System's architecture overview, based on Armony et al. (1995) auditory fear conditioning model and Lowe et al. (2011) dopamine modulated Pavlovian conditioning. Black arrows represent fixed weight values. Blue arrows represent weights updated using the Stent-Hebb rule (Stent, 1973). Weights values represented by green arrows are updated using the Hebbian learning rule and an eligibility trace.

7.3.1. Sensory Inputs and Preprocessing

In order to emulate auditory-cue fear conditioning experiments, the model is fed with synthetic audio signals which consist of single tones plus a very low amplitude (about 2.5%) noise floor from measurements in our home lab, which are sampled at 48 KHz by the NAO robot.

We process the incoming signal in frames of approx. 21 ms (1024 samples), which corresponds to the physiological construction of receptive fields (Armony et al., 1995). In this way, the system response to a given input may be interpreted as the time-averaged response of a cell to a tone presented in this 21 ms window or time step. For each window, we compute the spectral amplitude of the signal using a short-time Fourier transform (STFT).

The entire available frequency range of 20 Hz to 20 KHz, which corresponds to the microphones specifications of NAO, is divided into 24 intervals. Each interval

[1] NAO is a small-sized humanoid robot produced by Aldebaran-Robotics.

is represented by one neuron in an auditory input layer. The neural activation corresponds to the sum of all spectral amplitudes in the corresponding interval and is then normalized to $[0, 1]$. We implemented a simple signal detector module that detects the onset and the ending of signals based on the root mean square (RMS) value of the incoming signal and RMS value of the noise floor. This information is used to generate the unconditioned stimulus (US) just for the desired neutral conditioned stimulus (CS).

7.3.2. Neural Architecture and Learning

The auditory thalamus is modelled based on the model suggested by Armony et al. (1995) of the medial geniculate body (MGB), which is also supported by Weinberger (2011). The ventral division (MGv) of the MGB with a tonotopic organization feeds the auditory cortex module. The medial division (MGm) and the posterior intralaminar nucleus (PIN) are merged in a single module, which makes an initial association between CS and US, and then forwards its output to the amygdala's basolateral complex (BLA) and the auditory cortex modules as shown in Figure 7.3.

The "auditory thalamus" (MGv and MGm/PIN), "auditory cortex" (AC) and "basolateral complex" (BLA) modules are based on the architecture described by Armony et al. (1995). Each structure is modelled by a single-layer neural network and the modules' connectivity is done in a feed-forward manner, see Figure 7.4. The output of each of these modules is proportional to the output of the sending layer and normalized through both a squashing function and a winner-takes-all algorithm that serves to laterally inhibit the activation of less active or "loser" neurons.

From this point on we use f and g to denote a linear squashing function that trims neural activation to the interval $[0, 1]$ and $[-1, 1]$ respectively. For all equations, time dependence (t) is omitted and just indicated when it is different from the current time step.

The activation of the winning unit a_{win} in the receiving module is computed as follows:

$$a_{win} = \mathrm{f}\left(\sum_{j \in S} a_j w_{ji}\right), \tag{7.1}$$

where S are all units in the sending layer(s) and w_{ij} is the weight between the sending unit j and the current unit i. Connection weights between modules are randomly initialized.

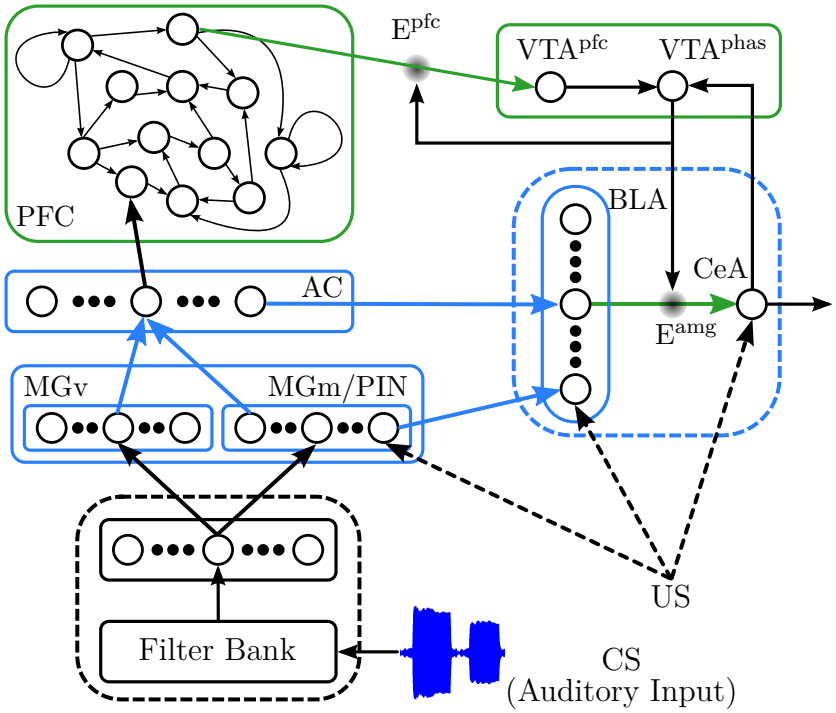


Figure 7.4 – Neural network architecture overview. Based on an echo state network (PFC), online learning algorithm for echo state network readout layer and amygdala internal connections indicated with E^{pfc} and E^{amg}, and single-layer feed-forward neural networks. For simplicity only one connection per layer in shown. Black lines represent fixed weight values.

The activation for each unit i in the receiving module r is calculated as follows:

$$a_i = \mathrm{f}\left(\sum_{j \in S} a_j w_{ji} - \mu_r a_{win}\right), \tag{7.2}$$

where μ_r is the strength of the lateral inhibition in module r.

Connection weights are updated after each cycle or epoch using the Stent-Hebb rule (Stent, 1973, see also Section 4.2), which prevents weights saturation:

$$w'_{ji} = \begin{cases} w_{ji} + \epsilon \cdot a_i \cdot a_j, & \text{if } a_j > \overline{a} \\ w_{ji}, & \text{otherwise,} \end{cases} \tag{7.3}$$

and

$$w_{ji} = \frac{w'_{ji}}{\sum_{j \in S} w'_{ji}}, \tag{7.4}$$

where $\overline{a}$ is the mean activation of the sending layer and ϵ is the learning rate.

Table 7.1 summarizes the parameters used for the auditory thalamus, the auditory cortex and the basolateral complex modules (Armony, Servan-Schreiber, Romanski, et al., 1997), which were determined with empirical trials and based on Armony et al. (1995) results.

Table 7.1 – Summary of parameters used in MGv, MGm/PIN, AC, and BLA modules.

Variable name	Value	Description
ϵ	0.2	common learning rate value
w^{us}	0.4	fixed weight value for US connections
μ for MGv	0.1	lateral inhibition in MGv module
μ for MGm/PIN	0.3	lateral inhibition in MGm/PIN module
μ for AC	0.6	lateral inhibition in AC module
μ for BLA	0.1	lateral inhibition BLA module
MGv size	10	number of units in module
MGm/PIN size	10	number of units in module
AC size	10	number of units in module
BLA size	10	number of units in module

The modules representing the "prefrontal cortex" (PFC), "ventral tegmental area" (VTA), and amygdala's "central nucleus" (CeA) are based on the model described by Lowe et al. (2011). The interactions of these three modules capture the basic functionality of biological reward prediction learning. This part of the model is based on an echo state network (ESN) approach (Jaeger, 2002; Lukoševičius and Jaeger, 2009, see also Section 4.6). ESNs are three-layered recurrent neural architectures that have demonstrated to be particularly effective at processing temporal stimuli. Their main characteristic is that only the *readout* weights are trained, which, in this case, are the weights connecting PFC units with a VTA^{pfc} unit. PFC is the reservoir of our ESN. This reservoir is sparsely connected with randomly generated weights. The reservoir has to satisfy the so-called echo state property that guarantees damping reverberations of the input signals, for details see Section 4.6 or Jaeger's report Jaeger (2002). The input layer of our ESN corresponds to the auditory cortex units, which are connected to the PFC using fixed weights, randomly and sparsely generated. Table 7.2 summarizes ESN parameters.

The VTA^{pfc} neuron corresponds to the readout layer of the ESN. To improve biological plausibility Lowe et al. (2011) introduced two features in the use of ESN. Firstly, they only allow non-negative activation within the reservoir. Secondly, they use a *"phasic dynamics of dopamine"* (DA)-based online learning rule to update the readout weights, see Lowe et al. (2011) for details.

Table 7.2 – Summary of parameters used in PFC, VTA, and CeA modules.

Variable name	Value	Description
Reservoir size	40	number of units in module
Reservoir connectivity	25%	random weights w_{dr} in $[-1, 1]$
Spectral radius	0.95	reservoir spectral radius
Input connectivity	25%	random weights w_{in} in $[0, 1]$
κ	0.1	learning rate
η	0.075	learning rate

Weights of the ESN readout layer and amygdala's CeA neuron are updated using the Hebbian learning rule and an eligibility trace $\mathrm{E}^{\mathrm{pfc}}$, and $\mathrm{E}^{\mathrm{amg}}$ respectively. $\mathrm{E}^{\mathrm{pfc}}$, and $\mathrm{E}^{\mathrm{amg}}$ are computed as follows:

$$\mathrm{E}^{*}_{t} = \max\left[\text{incoming signal}, \Omega \cdot \mathrm{E}^{*}_{t-1}\right], \tag{7.5}$$

where $*$ substitutes for pfc and amg, and Ω $(= 0.9)$ is a decay constant.

The PFC readout weights w^{pfc_i} connecting the reservoir unit i to the $\mathrm{VTA}^{\mathrm{pfc}}$ unit and the weights w^{bla_i} connecting BLA units to the CeA unit are updated as follows:

$$w^{pfc_i} = \begin{cases} \mathrm{f}\left(w^{pfc_i}_{t-1} + \kappa\, \mathrm{VTA}^{\mathrm{phas}}\, \mathrm{E}^{\mathrm{pfc}}_{t-1}\, PFC^{i}\right), & \text{if } \mathrm{VTA}^{\mathrm{phas}} \geq 0 \\ \mathrm{f}\left(w^{pfc_i}_{t-1} + \kappa\, \mathrm{VTA}^{\mathrm{phas}}\, PFC^{i}\right), & \text{if } \mathrm{VTA}^{\mathrm{phas}} < 0 \end{cases} \tag{7.6}$$

$$w^{bla_i} = \begin{cases} \mathrm{f}\left(w^{bla_i}_{t-1} + \eta\, \mathrm{VTA}^{\mathrm{phas}}\, \mathrm{E}^{\mathrm{amg}}\, \mathrm{CeA}\right), & \text{if } \mathrm{VTA}^{\mathrm{phas}} \geq 0 \\ \mathrm{f}\left(w^{bla_i}_{t-1} + \eta\, \mathrm{VTA}^{\mathrm{phas}}\, \mathrm{E}^{\mathrm{amg}}\right), & \text{if } \mathrm{VTA}^{\mathrm{phas}} < 0 \text{ and } \mathrm{US} = 0 \end{cases} \tag{7.7}$$

PFC^{i} is the current activation of reservoir unit i. $\mathrm{VTA}^{\mathrm{phas}}$ is the output value of VTA module. CeA is the output value of the amygdala module. The current activation value of PFC^{i}, $\mathrm{VTA}^{\mathrm{phas}}$ and CeA are computed as follows:

$$\mathrm{VTA}^{\mathrm{phas}} = \mathrm{g}\left(\mathrm{CeA} - \mathrm{VTA}^{\mathrm{pfc}}\right), \tag{7.8}$$

$$PFC^{i} = \tanh\left(\sum_{i,j} w^{dr_{ij}}\, PFC^{j}_{t-1} + \sum_{ik} w^{in_{ik}}\, AC^{out_k}\right), \tag{7.9}$$

$$\mathrm{CeA} = \mathrm{f}\left(\mathrm{US}\, w^{us} + \sum_{i} BLA^{i}\, w^{bla_i}\right). \tag{7.10}$$

Eq. (7.9) shows the recursive nature of ESN. This property provides a short-term memory that is used for updating estimates of the value of the stimulus. This spatial-temporal relationship between input signals differs from the classical temporal difference learning rule but the system as a whole allows for temporal dynamics between stimuli to be captured.

7.4. Experimental Procedure

The weights of Armony-based modules (MGv, MGm/PIN, BLA, and AC) are randomly generated with values ranging within $[0, 1]$. The weights connecting the BLA units and the CeA unit are set to 1 divided by the number of BLA units. PFC (ESN) readout weights are randomly initialized with values from $[0, 1]$.

Although we try to process sensory input within bio-plausible time windows, we do not consider latencies in signal processing nor transmission. Instead, we only consider coincident convergence of subcortical and cortical information to the BLA, which seems to be around 15 ms as reported by L. R. Johnson et al. (2008). This is translated to the following data flow in our current implementation: the auditory input is first preprocessed by the filter bank, then by the auditory thalamus followed by the auditory cortex. Both modules (AC and MGm/PIN) feed simultaneously the BLA module and an affective state is generated at the CeA. The amygdala output in conjunction with the auditory cortex activation is then used to trigger a dopamine modulation in the amygdala via the PFC and VTA modules.

The experimental part was divided into two phases. The first phase, called *"development"*, allows the modules to define initial receptive fields for the frequencies (system's response to a determined frequency), facilitates conditioning and reduces transient effects – that may emerge due to the weights' random initialization – during conditioning (Armony et al., 1995). A number of randomly generated tones not paired with the US was presented to the Armony-based modules (Armony et al., 1995). We added white noise to the generated signals to emulate a real robot's recordings. The dopamine circuit modules (PFC and VTA) were switched off. We repeated this procedure varying the frequency ranges, number of frequencies, and signal lengths not detecting major changes. Based on Lowe et al. (2009) findings we decided to use 300 randomly generated tones ranging from 100 Hz to 12 KHz for 5 time steps (approx. 100 ms per tone). In the second phase, termed *"conditioning"*, we selected a tone (CS), which was then paired with a binary US signal. Contrary to the *"development"* phase, during *"conditioning"* the dopamine circuit modules (PFC and VTA) were active. The training phase lasted 300 trials and each trial lasted 4 time steps. The selected CS and the US were presented during four trials only, i.e. 75, 150, 225 and 300. Taken into account that the CS (after learning)

acquires an anticipatory value and allows for more timely behavioural responses, we delayed the onset of the US with respect to the onset of the CS in two time steps. Finally, for biological plausibility, the remaining *"conditioning"* trials, those where the CS-US pairing did not took place, were completed with randomly generated tones emulating environmental noise.

7.5. Results

7.5.1. Receptive Fields Development

Figure 7.5 shows an example of a receptive field obtained after development. In this phase, the CeA output is characterized by a weak activation ($< 3\,e^{-4}$) with similar activation profiles at all frequencies.

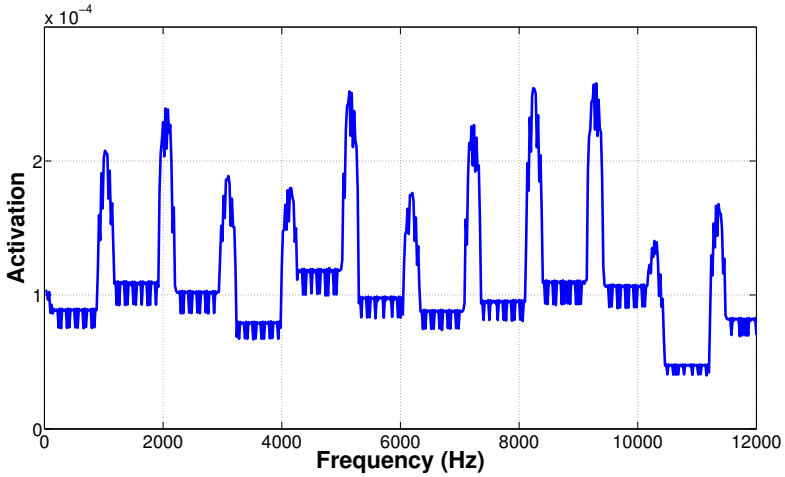


Figure 7.5 – Amygdala's (CeA) receptive fields after development phase, i.e. no frequency has yet been paired with any US signal. The figure shows the CeA activation after presenting single tones in the range $[20, 12000]$ Hz in a 20 Hz interval.

7.5.2. Conditioning

An example of a typical receptive field after *conditioning* is presented in Figure 7.6, where an enhancement of the system's response to the conditioned and neighbouring frequencies can be seen. Similar results were observed in animal experiments, producing what is known as stimulus generalization (Desiderato, 1964; Hoffman and Fleshler, 1961). Stimulus generalization has been interpreted as crucial for survival since it can elicit fast defensive responses under ambiguous sensory stimulation (Resnik et al., 2011).

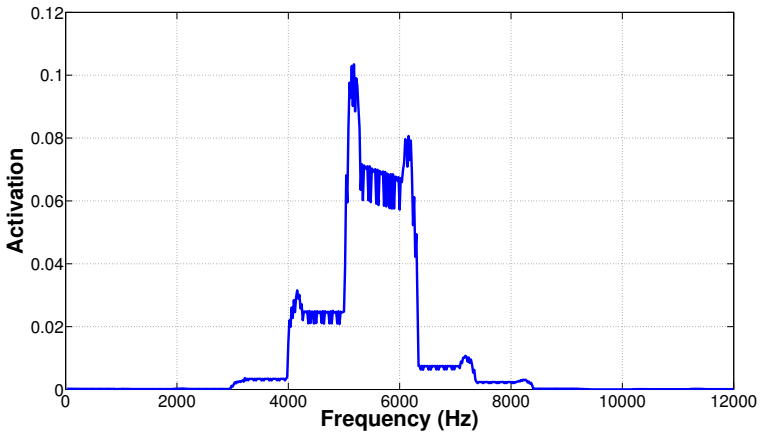


Figure 7.6 – Amygdala's (CeA) receptive fields after *conditioning* phase, without US signal and CS of 6 KHz. Consistently with animal experiments (Desiderato, 1964; Hoffman and Fleshler, 1961), amygdala activation decreases inversely with the distance to the CS. The figure shows the CeA activation after presenting single tones in the range $[20, 12000]$ Hz in a 20 Hz interval.

We observed that the system's response is higher for frequencies in the range $[5.1, 5.5]$ KHz than for the CS frequency. This phenomenon is due to resolution lost when converting spectral intensity to neural activation. Because 6 KHz divides two intervals, the spectral magnitude contributes to the activation of two neurons, i.e. intervals $[5, 6]$ KHz and $[6, 7]$ KHz respectively. This sort of ambiguous activation is encountered only for frequencies that divide two intervals. We believe that using a different discretization procedure, such as a gammatone filter along with a greater number of intervals, may help to address this issue in future implementations.

Figure 7.7 shows the system's receptive fields when both the CS and US signals are presented. The CeA activation is a combination of both the direct influence of the US signal (40%) and the dynamic magnification of the BLA response (60%).

7.5.3. Anticipation

Figure 7.8 shows the temporal changes of the system activation after *conditioning*. The maximal system activation is reached when both the CS and the US are presented at the same time. When only the CS is presented, the system's response is the result of the combination of the receptive field and the VTA modulation and a weak but consistent anticipatory system response is obtained, which is around 10% of the maximal possible activation. This output could easily be used to trigger a conditioned behaviour.

The feed-forward nature of the amygdala module allows the system to trigger a conditioned response independent of the US delay. We also observed that the

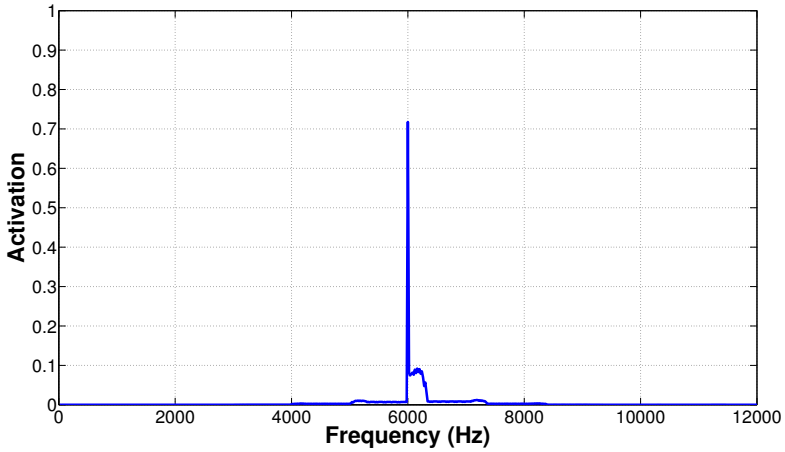


Figure 7.7 – Amygdala's (CeA) receptive fields after *conditioning* phase, with US signal. The figure shows the CeA activation after presenting single tones in the range $[20, 12,000]$ Hz in a 20 Hz interval.

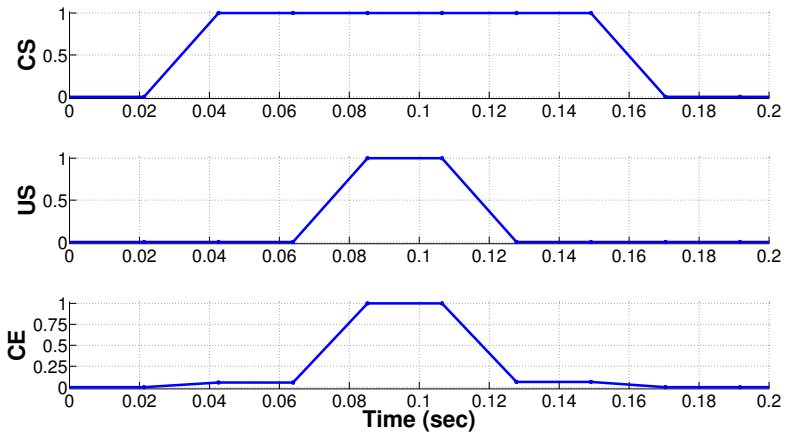


Figure 7.8 – Amygdala (CeA) activation profile after *conditioning* when presenting no signal (first and last 2 time steps), CS (6 time steps) and US (2 time steps). 40% of total activation corresponds to a direct contribution of US signal. The activation related to reward prediction before and after US signal is about 0.1.

number of trials does not have a major impact on the system activation. As few as one CS paired with the US and 200 trials suffice for conditioning, but a greater number of positive examples improve the overall response. The quick acquisition of fear memories is consistent with animal and human studies, where few trials account for a wider stimulus generalization (Resnik et al., 2011). Animal studies also support the fact that a greater number of trials increase stimulus discrimination, which improves inversely with the distance to the CS (Desiderato, 1964; Hoffman and Fleshler, 1961).

7.6. Discussion

A reservoir system approach for auditory-cue fear conditioning was presented. The hybrid architecture is able to quickly associate a CS with a US and to perform frequency discrimination and long-lasting fear memories. Our implementation can support acquisition reliably and it is consistent with animal and human studies in terms of stimulus generalization and discrimination (Desiderato, 1964; Hoffman and Fleshler, 1961; Resnik et al., 2011). The weak but consistent anticipatory response after conditioning can be used after amplification for triggering conditioned behaviours.

A difference between our implementation and Lowe et al. (2011) implementation is the origin of the CS signal. Lowe et al. (2011) model use abstract CS signals that are connected directly to both the PFC and to the CeA modules. Instead, we fed the PFC module with the output generated by the auditory cortex module and the CeA with the output generated by the BLA module. The US signal is connected to the MGm/PIN, BLA and CeA modules (LeDoux, 2007; Pape and Pare, 2010).

Another difference to most models on fear conditioning, as explained in Section 7.1, is the auditory input dimensionality. Since we considered noisy input signals and a preprocessing layer, the number of active units and the amplitude of the activation vary between trials, which represent an important step towards more bio-plausible computational models on fear conditioning. Our model does not model detailed temporal contingencies and only convergent cortical and subcortical signals are considered.

One of the limitations of the system is the simple modulation made by the PFC module through the VTA. This pathway can be used to support not only acquisition, but also inhibition of ambiguous responses like that observed in Figure 7.6, which originates when converting spectral amplitude to neural activation. The single output of the system limits the possible conditioned behaviour that the model may trigger in a real-world scenario.

An application of the bottom-up/top-down temporal behavioural control of this architecture could be to freeze or to make way for a human, thus avoiding collisions or blockages, and the robot previous behaviour would be resumed when appropriate. Here, the freezing or avoidance behaviour would be elicited by the amydgala based on environmental cues, and maintained or inhibited by the PFC based on the temporal sensorimotor input and neural dynamics. The learned reactive response from the amygdala supervised by the top-down control from the PFC represents a training-driven shift from reactive to proactive behavioural control. The advantages of using this type of architecture are least two-fold: firstly, the one-shot learning properties of the conditioned response via the amygdala, and the secondly, the self-sustaining/self-correcting response via the cortical loop.

8 Chapter Discussion

In this thesis biological adaptive self-preservation mechanisms were studied while exploring the benefits of models of different levels of abstraction for humanoid robot applications. Particularly in this research, we studied and modelled neural mechanisms involved in goal-directed behaviour, proprioceptive and nociceptive perceptual experience, and conditioning. The functional computational models developed were motivated by the cognitive architecture for the development of artificial autonomous systems suggested by Ziemke and Lowe (2009), introduced in Chapter 3. Although no detailed biological models were developed, a neurocomputational approach was chosen, because we believe that computational models have the potential to test neurophysiological hypotheses, contribute to the interpretation of unexplained phenomena and to the integration of different isolated observations into a single framework providing clues on how they may work alongside, compete or collectively contribute to more complex behaviours and cognitive capabilities.

In the following sections a summary of the three different neural architectures, their novelty and contributions will be discussed. Then, a general description of the contributions of the thesis will be given. Lastly, the possible extensions will be elucidated.

8.1. Reward-Seeking Behaviours

The first experiment addressed the problem of energetic autonomy from the perspective of goal-directed behaviour, see Chapter 5. Here a NAO humanoid robot was trained to seek for appetitive stimuli that let it autonomously recharge its battery. We used a classical reinforcement learning algorithm called SARSA (Sutton and Barto, 1998, p. 145). However, we optimized it to learn in a real-world scenario and manoeuvre a humanoid robot towards a docking station.

The introduced modifications can be summarized as follows:

Firstly, to avoid the initial random exploration characteristic of reinforcement

learning algorithms, we tele-operated the robot from random positions within the workspace to the goal. We recorded the resulting action sequences, states and received reward and used them as training examples for off-line training. During off-line training we by-passed the stochastic action selection policy and instead used the recorded actions. We refer to this procedure as supervised reinforcement learning. Off-line training considerably reduces the running time of the real-world.

Finally, instead of having a single state unit active at a time as prescribed for discrete state reinforcement learning algorithms, we used a normally distributed activation of state units (Foster et al., 2000) centred at the current robot state. We demonstrated that a Gaussian distributed state activation produced a state space reduction effect by spreading the knowledge about the current state to neighbouring states. This further reduces the required online learning steps.

8.2. Punishment and Nociception in Learning

The second experiment focuses on the differential effect of punishment and nociceptive inputs on motor skill learning, see Chapter 6. Here we use the continuous actor-critic automaton (CACLA) algorithm (van Hasselt and Wiering, 2007) to learn the inverse kinematics of a 2-dimensional model of the NAO's arm from scratch.

We demonstrated that CACLA, even more we believe than TD-learning algorithms in general, does not take into consideration the different neural pathways involved in reward- and punishment-driven learning. Consequently, TD-algorithms perform worst when using both reward and punishment than when using reward only. This results is agreement with current evidence from procedural and skill motor learning as shown in Chapter 6.

We believe that the detrimental effect of punishment can be explained by the loss of predictive power of reward and punishment when both feedback signals are conflated into a single scalar value (Lowe and Ziemke, 2013). To circumvent this problem we suggest the use of nociceptive input signals as a simple way of preserving the predictive power of both positive and negative feedback. Although this approach does not directly model the differential pathways for reward- and punishment-driven learning, it consistently and significantly improves the overall learned behaviour when both reward and punishment are used. In the case when only reward was used, the nociceptive input had a positive but negligible effect.

8.3. Conditioning for Event Anticipation

Finally, the third experiment focuses on the role of noxious stimuli in the formation of anticipatory behaviour, see Chapter 7. Here we use a novel architecture based on the vast literature on Pavlovian and instrumental conditioning, and a hybrid approach using an echo state network (ESN) and a dopamine modulated Pavlovian conditioning to anticipate noxious stimuli based on auditory cues. In the presented simulations, this architecture can quickly and reliably associate a conditioned stimuli (CS) with an unconditioned stimuli (US), perform frequency discrimination, and create long-lasting fear memories.

Consistent with animal and human studies (Desiderato, 1964; Hoffman and Fleshler, 1961; Resnik et al., 2011), our implementation establishes a relationship between the CS and the US with as few as one example of the CS paired with the US. The receptive field is initially large. However, the sole observation of stimuli similar to the CS but not paired to the US drives the size of the receptive field down. Moreover, the observation of further positive examples increases stimulus discrimination.

In addition to capturing the dynamic of acquisition of auditory-cue fear conditioning, our model differentiates from most existing models on fear conditioning in the sophistication of the auditory input. We use embodied noisy input and a preprocessing layer to convert audio signals to neural activation, this lead to varying neural activation between trials. Although, we still use single tones in our experiments, the use of noise inputs already represents an important step towards more biologically plausible computational models of fear conditioning.

8.4. Conclusion

We presented three different neuro-inspired experiments and discussed their relevance as adaptive self-preservative mechanisms for robot behaviour. We showed how neuro-inspired embodied perception at different abstraction levels could enhance and enable learning of self-preservative behaviour while solving artificial intelligence problems. All three experiments were motivated by neurocomputational learning mechanisms including goal-directed behaviour, proprioceptive and nociceptive perceptual experience, and conditioning. We also developed novel extensions to the learning algorithms used and applied them to the neglected niche of artificial self-preservative behaviours.

The scope of this thesis was not to provide algorithms or architectures with direct application in industrial or commercial settings. Instead, we used principles

underlying self-preservative behaviours to motivate the design of our neural architecture and experiments, and to show that these principles can also be successfully applied to neuro-robotic architectures while solving different robotic tasks.

We do not claim that the presented design principles are superior in performance or simplicity to other approaches, but we do support the hypothesis of a bottom-up development of autonomy and cognitive development, as presented in Chapter 3. The intertwined nature of neural and non-neural autonomic regulatory mechanisms and sensorimotor activity strengthen this hypothesis. Consequently, we advocate a more mindful inclusion of self-preservative mechanisms into the design of robotic systems if the goal is to develop truly autonomous robots. However, if the goal is to develop cost-effective and specialised robotic solutions conventional machine learning and automation approaches may be better suited.

8.5. Future Research

Finally, several experiments and open questions that could be conducted based on the theoretical and methodological framework established within this thesis will be suggested in this section.

8.5.1. Reward-Seeking Behaviours

In addition to further improvements of the behavioural aspects of docking for recharging and grasping, there is a more complex open research question regarding energetic autonomy and reward-seeking behaviours, namely how to develop homeostatic and metabolic energy management system. Furthermore, how can the *feeling* of being hungry be grounded on sensorimotor experience?

We believe that whatever a homeostatic and metabolic energy management system may look like, it will still require a set of reactive and learned mechanisms such as those developed in Chapter 5.

8.5.2. Punishment and Nociception in Learning

Aspects other than the success rate on a reaching task could be studied, for instance, it would be interesting to know how nociceptive input does alter the robot pose, and the amount of punishment experienced along a given trajectory.

It would be also interesting to know how well does the positive effect of nociceptive input perform in more complex reaching scenarios such as highly redundant humanoid robot arms or in problems outside the niche of motor skill learning.

Another problem not explored in this research was the computational modelling of the differential pathway for reward- and punishment-driven learning in the striatum. Although we provide an effective alternative it would be desirable to explore alternatives for the modelling of the two pathways within the striatum which would help to shed light onto the interaction of both types of feedback in different learning tasks.

8.5.3. Event Anticipation via Conditioning

For our auditory-cue fear conditioning model, a biologically constrained preprocessing of the auditory signal would be required, i.e. incorporating different degrees of processing and latencies for subcortical and cortical areas. An appropriate filter bank such as a gammatone filter may contribute positively to improve frequency discrimination and to develop a more biologically plausible thalamus and auditory cortex modules.

We believe that keeping a coarse division of the amygdala into two sub-modules may facilitate the use of a modulating signal coming from the PFC module. However, our amygdala model needs improvement at different levels before being use in a fully embodied agent. For instance, the recurrent nature of the main input nuclei in the amygdala (L. R. Johnson et al., 2008) encourages exploring a reservoir approach for the future implementation of the BLA module. In order to drive or modulate different conditioned behaviours a CeA module with multiple output units may be necessary.

Finally, the model could be further improved to create the first fully embodied fear conditioning model on a humanoid robot. A fully embodied model may serve not only to improve robot assistance but also to contribute to a better understanding of fear circuits. Embodied models can provide a framework to better understand otherwise hard to examine dynamics. For instance, embodied models could be used to study the effects of timing between sensory input, internal state and action on different aspects of conditioning dynamics and animal behaviour. Similarly, it could inform about the sensory complexity necessary or sufficient for forming robust fear memories.

A Appendix

References

A

Abe, M., Schambra, H., Wassermann, E. M., Luckenbaugh, D., Schweighofer, N., and Cohen, L. G. (2011). « Reward Improves Long-Term Retention of a Motor Memory Through Induction of Offline Memory Gains ». *Current Biology* 21(7), pp. 557–562. DOI: 10.1016/j.cub.2011.02.030 (cit. on p. 90).

Adolphs, R. (2010). « Conceptual Challenges and Directions for Social Neuroscience ». *Neuron* 65(6), pp. 752–767. DOI: 10.1016/j.neuron.2010.03.006 (cit. on p. 31).

Akayama, S., Matsunaga, N., and Kawaji, S. (2006). « Experimental Analysis and Modeling of Superficial Pain on Upper Limb ». In: *International Joint Conference SICE-ICASE.* Busan, South Korea: IEEE, pp. 2891–2894. DOI: 10.1109/sice.2006.314861 (cit. on pp. 39, 40).

Alexander, W. H. and Sporns, O. (2002). « An Embodied Model of Learning, Plasticity, and Reward ». *Adaptive Behavior* 10(3-4), pp. 143–159 (cit. on pp. 46, 111).

Anderson, A. K. and Phelps, E. A. (2001). « Lesions of the Human Amygdala Impair Enhanced Perception of Emotionally Salient Events ». *Nature.* Letters to Nature 411(6835), pp. 305–309. DOI: 10.1038/35077083 (cit. on pp. 24, 27).

Arbib, M. A. and Fellous, J.-M. (2004). « Emotions: From Brain to Robot ». *Trends in Cognitive Sciences* 8(12), pp. 554–561. DOI: 10.1016/j.tics.2004.10.004 (cit. on pp. 2, 30).

Armony, J. L., Servan-Schreiber, D., Cohen, J. D., and LeDoux, J. E. (1995). « An Anatomically Constrained Neural Network Model of Fear Conditioning ». *Behavioral Neuroscience* 109(2), pp. 246–257. DOI: 10.1037/0735-7044.109.2.246 (cit. on pp. 44, 51, 111, 115, 116, 118, 120).

Armony, J. L., Servan-Schreiber, D., Cohen, J. D., and LeDoux, J. E. (1997). « Computational Modeling of Emotion: Explorations Through the Anatomy and Physiology of Fear Conditioning ». *Trends in Cognitive Sciences.* Review 1(1), pp. 28–34. DOI: 10.1016/s1364-6613(97)01007-3 (cit. on p. 44).

Armony, J. L., Servan-Schreiber, D., Romanski, L. M., Cohen, J. D., and LeDoux, J. E. (1997). « Stimulus Generalization of Fear Responses: Effects of Auditory Cortex Lesions in a Computational Model and in Rats ». *Cerebral Cortex* 7(2), pp. 157–165. DOI: 10.1093/cercor/7.2.157 (cit. on pp. 44, 118).

Arredondo, T., Freund, W., Navarro-Guerrero, N., and Castillo, P. (2013). « Fuzzy Motivations in a Multiple Agent Behaviour-Based Architecture ». *International Journal of Advanced Robotic Systems* 10(313), pp. 1–13. DOI: 10.5772/56578 (cit. on p. 153).

Asada, M., MacDorman, K. F., Ishiguro, H., and Kuniyoshi, Y. (2001). « Cognitive Developmental Robotics as a new Paradigm for the Design of Humanoid Robots ». *Robotics and Autonomous Systems.* Humanoid Robots 37(2-3), pp. 185–193. DOI: 10.1016/s0921-8890(01)00157-9 (cit. on p. 31).

Atcherley, C. W., Wood, K. M., Parent, K. L., Hashemi, P., and Heien, M. L. (2015). « The Coaction of Tonic and Phasic Dopamine Dynamics ». *Chemical Communications* 51(12), pp. 2235–2238. DOI: 10.1039/C4CC06165A (cit. on p. 114).

B

Bakin, J. S. and Weinberger, N. M. (1990). « Classical Conditioning Induces CS-Specific Receptive Field Plasticity in the Auditory Cortex of the Guinea Pig ». *Brain Research* 536(1-2), pp. 271–286. DOI: 10.1016/0006-8993(90)90035-a (cit. on p. 113).

Balkenius, C. and Morén, J. (2001). « Emotional Learning: A Computational Model of the Amygdala ». *Cybernetics and Systems: An International Journal* 32(6), pp. 611–636. DOI: 10.1080/01969720118947 (cit. on pp. 44, 45, 111).

Balleine, B. W. (2005). « Neural Bases of Food-Seeking: Affect, Arousal and Reward in Corticostriatolimbic Circuits ». *Physiology & Behavior.* Purdue University Ingestive Behavior Research Center Symposium. Dietary Influences on Obesity: Environment, Behavior and Biology 86(5), pp. 717–730. DOI: 10.1016/j.physbeh.2005.08.061 (cit. on p. 62).

Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983). « Neuronlike Adaptive Elements that can Solve Difficult Learning Control Problems ». *IEEE Transactions on Systems, Man and Cybernetics* SMC-13(5), pp. 834–846. DOI: 10.1109/tsmc.1983.6313077 (cit. on p. 62).

Bechara, A., Damasio, H., Damasio, A. R., and Lee, G. P. (1999). « Different Contributions of the Human Amygdala and Ventromedial Prefrontal Cortex to Decision-Making ». *The Journal of Neuroscience* 19(13), pp. 5473–5481 (cit. on p. 23).

Belova, M. A., Paton, J. J., and Salzman, C. D. (2008). « Moment-to-Moment Tracking of State Value in the Amygdala ». *The Journal of Neuroscience.* Behavioral/Systems/Cognitive 28(40), pp. 10023–10030. DOI: 10.1523/jneurosci.1400-08.2008 (cit. on pp. 23, 24).

Blessing, W. W. (1997). « Inadequate Frameworks for Understanding Bodily Homeostasis ». *Trends in Neurosciences* 20(6), pp. 235–239. DOI: 10.1016/s0166-2236(96)01029-6 (cit. on pp. 21, 22, 26).

Blessing, W. W. and Benarroch, E. E. (2012). « Lower Brainstem Regulation of Visceral, Cardiovascular, and Respiratory Function ». In: *The Human Nervous System.* Third ed. Academic Press/Elsevier. Chap. 29, pp. 1058–1073. DOI: 10.1016/b978-0-12-374236-0.10029-x (cit. on pp. 12, 14, 15, 21, 30).

Bonica, J. J. (1979). « The Need of a Taxonomy ». *PAIN* 6(3), pp. 247–252. DOI: 10.1016/0304-3959(79)90046-0 (cit. on p. 11).

Boureau, Y.-L. and Dayan, P. (2011). « Opponency Revisited: Competition and Cooperation Between Dopamine and Serotonin ». *Neuropsychopharmacology* 36(1), pp. 74–97. DOI: 10.1038/npp.2010.151 (cit. on p. 16).

Bovet, S. I. (2007). « Robots with Self-Developing Brains ». PhD thesis. Mathematisch-naturwissenschaftlichen Fakultät der Universität Zürich, pp. 201+ (cit. on p. 32).

Brooks, J. and Tracey, I. (2005). « REVIEW: From Nociception to Pain Perception: Imaging the Spinal and Supraspinal Pathways ». *Journal of Anatomy* 207(1), pp. 19–33. DOI: 10.1111/j.1469-7580.2005.00428.x (cit. on p. 11).

Buonomano, D. V. and Maass, W. (2009). « State-Dependent Computations: Spatiotemporal Processing in Cortical Networks ». *Nature Reviews Neuroscience* 10(2), pp. 113–125. DOI: 10.1038/nrn2558 (cit. on p. 65).

C

Canteras, N. S. (2002). « The Medial Hypothalamic Defensive System: Hodological Organization and Functional Implications ». *Pharmacology Biochemistry and Behavior.* Functional Role of Specific Systems within the Extended Amygdala and Hypothalamus 71(3), pp. 481–491. DOI: 10.1016/s0091-3057(01)00685-2 (cit. on pp. 8, 12, 19, 20, 30).

Cardinal, R. N., Parkinson, J. A., Hall, J., and Everitt, B. J. (2002). « Emotion and Motivation: The Role of the Amygdala, Ventral Striatum, and Prefrontal Cortex ». *Neuroscience & Biobehavioral Reviews* 26(3), pp. 321–352. DOI: 10.1016/s0149-7634(02)00007-6 (cit. on p. 46).

Carrive, P. and Morgan, M. M. (2012). « Periaqueductal Gray ». In: *The Human Nervous System.* Third ed. Elsevier. Chap. 10, pp. 367–400. DOI: 10.1016/b978-0-12-374236-0.10010-0 (cit. on pp. 17, 18).

Cellier, L., Dauchez, P., Zapata, R., and Uchiyama, M. (1995). « Collision Avoidance for a Two-Arm Robot by Reflex Actions: Simulations and Experimentations ». 14(2), pp. 219–238. DOI: 10.1007/bf01559613 (cit. on p. 3).

Conn, K. and Peters, R. A. (2007). « Reinforcement Learning with a Supervisor for a Mobile Robot in a Real-World Environment ». In: *International Symposium on Computational Intelligence in Robotics and Automation (CIRA).* Jacksonville, FL, USA: IEEE, pp. 73–78. DOI: 10.1109/cira.2007.382878 (cit. on pp. 76–78).

Cummins, J. (2012). « Do the Origins of Biological General Intelligence Lie in an Adaptation of the Stress Response? » In: *Integral Biomathics: Tracing the Road to Reality.* Springer Berlin Heidelberg, pp. 155–168. DOI: 10.1007/978-3-642-28111-2_15 (cit. on p. 36).

D

Damasio, A. R. (1996). « The Somatic Marker Hypothesis and the Possible Functions of the Prefrontal Cortex [and Discussion] ». *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 351(1346), pp. 1413–1420. DOI: 10.1098/rstb.1996.0125 (cit. on p. 93).

Darken, C. and Moody, J. E. (1991). « Note on Learning Rate Schedules for Stochastic Optimization ». In: *Advances in Neural Information Processing Systems (NIPS).* Learning Systems No. 3. Denver, CO, USA: Morgan Kaufmann, pp. 832–838 (cit. on p. 56).

Davis, M. (1992). « The Role of the Amygdala in Fear and Anxiety ». *Annual Review of Neuroscience* 15(1), pp. 353–375. DOI: 10.1146/annurev.ne.15.030192.002033 (cit. on pp. 112, 114).

Daw, N. D., Niv, Y., and Dayan, P. (2005). « Uncertainty-Based Competition between Prefrontal and Dorsolateral Striatal Systems for Behavioral Control ». *Nature Neuroscience* 8(12), pp. 1704–1711. DOI: 10.1038/nn1560 (cit. on pp. 61, 62).

Dayan, P. and Niv, Y. (2008). « Reinforcement Learning: The Good, The Bad and The Ugly ». *Current Opinion in Neurobiology.* Cognitive Neuroscience 18(2), pp. 185–196. DOI: 10.1016/j.conb.2008.08.003 (cit. on pp. 58, 60, 61, 90, 91).

den Dulk, P., Heerebout, B. T., and Phaf, R. H. (2003). « A Computational Study into the Evolution of Dual-Route Dynamics for Affective Processing ». *Journal of Cognitive Neuroscience* 15(2), pp. 194–208. DOI: 10.1162/089892903321208132 (cit. on p. 111).

Desiderato, O. (1964). « Generalization of Conditioned Suppression ». *Journal of Comparative and Physiological Psychology* 57(3), pp. 434–437 (cit. on pp. 121–124, 128).

Diamantaras, K. I. and Kung, S. Y. (1996). *Principal Component Neural Networks: Theory and Applications.* First ed. Adaptive and Learning Systems for Signal Processing, Communications and Control. Wiley-Interscience, pp. 272+ (cit. on p. 58).

Doya, K. (2002). « Metalearning and Neuromodulation ». *Neural Networks* 15(4-6), pp. 495–506. DOI: 10.1016/s0893-6080(02)00044-8 (cit. on pp. 62, 72).

F

Fellous, J.-M., Armony, J. L., and LeDoux, J. E. (2002). « Emotional Circuits and Computational Neuroscience ». In: *The Handbook of Brain Theory and Neural Networks.* Second ed. Computer Science and Intelligent Systems. A Bradford Book/The MIT Press, pp. 398–401 (cit. on p. 23).

Foster, D. J., Morris, R. G. M., and Dayan, P. (2000). « A Model of Hippocampally Dependent Navigation, using the Temporal Difference Learning Rule ». *Hippocampus* 10(1), pp. 1–16 (cit. on pp. 79, 81, 127).

Franklin, D. W., So, U., Burdet, E., and Kawato, M. (2007). « Visual Feedback is not Necessary for the Learning of Novel Dynamics ». *PLoS ONE* 2(12), e1336+. DOI: 10.1371/journal.pone.0001336 (cit. on p. 93).

G

Galea, J. M., Mallia, E., Rothwell, J., and Diedrichsen, J. (2015). « The Dissociable Effects of Punishment and Reward on Motor Learning ». *Nature Neuroscience* 18(4), pp. 597–602. DOI: 10.1038/nn.3956 (cit. on pp. 92, 107).

Gebhart, G. F. and Schmidt, R. F., eds. (2013). *Encyclopedia of Pain.* Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 352+. DOI: 10.1007/978-3-642-28753-4 (cit. on pp. 12, 13).

Germana, J. (1969). « Central Efferent Processes and Autonomic-Behavioral Integration ». *Psychophysiology.* Brain & Behavior: Physiological Psychology 6(1), pp. 78–90. DOI: 10.1111/j.1469-8986.1969.tb02886.x (cit. on pp. 23, 24).

Geula, C. and Mesulam, M. M. (2012). « Brainstem Cholinergic Systems ». In: *The Human Nervous System.* Third ed. Elsevier. Chap. 14, pp. 456–470. DOI: 10.1016/b978-0-12-374236-0.10014-8 (cit. on p. 17).

Ghory, I. (2004). *Reinforcement Learning in Board Games.* Tech. rep. Department of Computer Science, University of Bristol, pp. 57+ (cit. on p. 77).

Gupta, R., Koscik, T. R., Bechara, A., and Tranel, D. (2011). « The Amygdala and Decision-Making ». *Neuropsychologia.* The Human Amygdala and Emotional Function 49(4), pp. 760–766. DOI: 10.1016/j.neuropsychologia.2010.09.029 (cit. on pp. 43, 111).

H

Ha, S. and Liu, C. K. (2015). « Multiple Contact Planning for Minimizing Damage of Humanoid Falls ». In: *IEEE/RSJ International Conference on Intelligent Robots and*

Systems (IROS). Hamburg, Germany: IEEE, pp. 2761–2767. DOI: 10.1109/IROS.2015.7353756 (cit. on p. 3).

Halliday, G., Reyes, S., and Double, K. (2012). « Substantia Nigra, Ventral Tegmental Area, and Retrorubral Fields ». In: *The Human Nervous System.* Third ed. Academic Press/Elsevier. Chap. 13, pp. 439–455. DOI: 10.1016/b978-0-12-374236-0.10013-6 (cit. on pp. 16, 17).

Harper, C. and Virk, G. (2010). « Towards the Development of International Safety Standards for Human Robot Interaction ». *International Journal of Social Robotics.* Special Issue: Towards Safety in Human Robot Interaction 2(3), pp. 229–234. DOI: 10.1007/s12369-010-0051-1 (cit. on pp. 2, 39).

Haugeland, J. (1989). *Artificial Intelligence: The Very Idea.* Computer Science and Intelligent Systems. Cambridge, MA, USA: A Bradford Book/The MIT Press, pp. 302+ (cit. on p. 31).

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory.* 2002 new ed. by Psychology Press. John Wiley & Sons Inc., pp. 378+ (cit. on p. 50).

Heinrich, S., Folleher, P., Springstübe, P., Strahl, E., Twiefel, J., Weber, C., and Wermter, S. (2014). « Object Learning with Natural Language in a Distributed Intelligent System: A Case Study of Human-Robot Interaction ». In: *Foundations and Practical Applications of Cognitive Systems and Information Processing: International Conference on Cognitive Systems and Information Processing (CSIP2012).* Vol. 215. Advances in Intelligent Systems and Computing. Beijing, China: Springer Berlin Heidelberg, pp. 811–819. DOI: 10.1007/978-3-642-37835-5_70 (cit. on p. 72).

Hester, R., Murphy, K., Brown, F. L., and Skilleter, A. J. (2010). « Punishing an Error Improves Learning: The Influence of Punishment Magnitude on Error-Related Neural Activity and Subsequent Learning ». *The Journal of Neuroscience* 30(46), pp. 15600–15607. DOI: 10.1523/JNEUROSCI.2565-10.2010 (cit. on p. 91).

Hoffman, H. S. and Fleshler, M. (1961). « Stimulus Factors in Aversive Controls: The Generalization of Conditioned Suppression ». *Journal of the Experimental Analysis of Behavior* 4(4), pp. 371–378. DOI: 10.1901/jeab.1961.4-371 (cit. on pp. 121–124, 128).

Hornung, J.-P. (2012). « Raphe Nuclei ». In: *The Human Nervous System.* Third ed. Elsevier. Chap. 11, pp. 401–424. DOI: 10.1016/b978-0-12-374236-0.10011-2 (cit. on pp. 14, 16).

I

Ieropoulos, I., Greenman, J., Melhuish, C., and Horsfield, I. (2010). « EcoBot-III - A Robot with Guts ». In: *Artificial Life XII: International Conference on the Synthesis and Simulation of Living Systems (ALIFE).* Robotics. Odense, Denmark: The MIT Press, pp. 733–740 (cit. on p. 35).

Ieropoulos, I., Melhuish, C., and Greenman, J. (2003). « Artificial Metabolism: Towards True Energetic Autonomy in Artificial Life ». In: *European Conference on Advances in Artificial Life (ECAL).* Vol. 2801. LNCS. Dortmund, Germany: Springer Berlin Heidelberg, pp. 792–799. DOI: 10.1007/978-3-540-39432-7_85 (cit. on pp. 35, 36).

Ieropoulos, I., Melhuish, C., Greenman, J., and Horsfield, I. (2005). « EcoBot-II: An Artificial Agent with a Natural Metabolism ». *International Journal of Advanced Robotic Systems,* pp. 295–300. DOI: 10.5772/5777 (cit. on p. 35).

Ito, K., Fukumori, Y., and Takayama, A. (2007). « Autonomous Control of Real Snake-Like Robot using Reinforcement Learning; Abstraction of State-Action Space using Properties of Real World ». In: *International Conference on Intelligent Sensors, Sensor Networks and Information (ISSNIP).* Melbourne, VIC, Australia: IEEE, pp. 389–394. DOI: 10.1109/issnip.2007.4496875 (cit. on pp. 77, 78).

J

Jaeger, H. (2001). *The 'Echo State' Approach to Analysing and Training Recurrent Neural Networks.* Tech. rep. 148. Bremen, Germany: Fraunhofer Institute for Autonomous Intelligent Systems (AIS), pp. 43+ (cit. on pp. 65, 66, 69, 70).

Jaeger, H. (2002). *A Tutorial on Training Recurrent Neural Networks, Covering BPPT, RTRL, EKF and the 'Echo State Network' Approach.* Tech. rep. 159. Bremen, Germany: Fraunhofer Institute for Autonomous Intelligent Systems (AIS), pp. 48+ (cit. on pp. 65, 67, 69, 118).

Jaeger, H. (2007). *Discovering Multiscale Dynamical Features with Hierarchical Echo State Networks.* Tech. rep. 10. Bremen, Germany: School of Engineering and Science. Jacobs University, pp. 30+ (cit. on p. 70).

Johnson, L. R., Hou, M., Ponce-Alvarez, A., Gribelyuk, L. M., Alphs, H. H., Albert, L., Brown, B. L., LeDoux, J. E., and Doyére, V. (2008). « A Recurrent Network in the Lateral Amygdala: A Mechanism for Coincidence Detection ». *Frontiers in Neural Circuits* 2(3). DOI: 10.3389/neuro.04.003.2008 (cit. on pp. 113, 120, 130).

Johnson, M. H. (2005). « Subcortical Face Processing ». *Nature Reviews Neuroscience* 6(10), pp. 766–774. DOI: 10.1038/nrn1766 (cit. on p. 8).

K

Kietzmann, T. C. and Riedmiller, M. (2009). « The Neuro Slot Car Racer: Reinforcement Learning in a Real World Setting ». In: *International Conference on Machine Learning and Applications (ICMLA).* Miami, FL, USA: IEEE, pp. 311–316. DOI: 10.1109/icmla.2009.15 (cit. on pp. 77, 78).

Kim, S. H., Yoon, H., Kim, H., and Hamann, S. (2015). « Individual Differences in Sensitivity to Reward and Punishment and Neural Activity During Reward and Avoidance Learning ». *Social Cognitive and Affective Neuroscience* 10(9), pp. 1219–1227. DOI: 10.1093/scan/nsv007 (cit. on pp. 90–92, 107).

Kohonen, T. K. (1990). « The Self-Organizing Map ». *Proceedings of the IEEE* 78(9), pp. 1464–1480. DOI: 10.1109/5.58325 (cit. on p. 50).

Kohonen, T. K. (2001). *Self-Organizing Maps.* Third ed. Springer Series in Information Sciences. Springer Berlin Heidelberg, pp. 521+. DOI: 10.1007/978-3-642-56927-2 (cit. on p. 50).

Koob, G. F. and Volkow, N. D. (2009). « Neurocircuitry of Addiction ». *Neuropsychopharmacology* 35(1), pp. 217–238. DOI: 10.1038/npp.2009.110 (cit. on p. 114).

Kötter, R. and Meyer, N. (1992). « The Limbic System: A Review of its Empirical Foundation ». *Behavioural Brain Research* 52(2), pp. 105–127. DOI: 10.1016/s0166-4328(05)80221-9 (cit. on pp. 21, 22).

Krasne, F. B., Fanselow, M. S., and Zelikowsky, M. (2011). « Design of a Neurally Plausible Model of Fear Learning ». *Frontiers in Behavioral Neuroscience* 5(41). DOI: 10.3389/fnbeh.2011.00041 (cit. on pp. 44, 46, 111).

Krichmar, J. L. (2008). « The Neuromodulatory System: A Framework for Survival and Adaptive Behavior in a Challenging World ». *Adaptive Behavior* 16(6), pp. 385–399. DOI: 10.1177/1059712308095775 (cit. on pp. 8, 10, 30).

KSERA (2010-2013). *The KSERA project (Knowledgeable SErvice Robots for Aging).* URL: http://ksera.ieis.tue.nl/ (cit. on pp. 3, 38, 72, 75).

Kuremoto, T., Ohta, T., Kobayashi, K., and Obayashi, M. (2009). « A Dynamic Associative Memory System by Adopting an Amygdala Model ». *Artificial Life and Robotics* 13(2), pp. 478–482. DOI: 10.1007/s10015-008-0590-9 (cit. on p. 45).

L

LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R. (1998). « Efficient BackProp ». In: *Neural Networks: Tricks of the Trade.* Vol. 1524. LNCS. Berlin, Heidelberg: Springer Berlin Heidelberg. Chap. 1, pp. 9–50. DOI: 10.1007/3-540-49430-8_2 (cit. on pp. 49, 51, 53–57, 98).

LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R. (2012). « Efficient BackProp ». In: *Neural Networks: Tricks of the Trade.* Second ed. Vol. 7700. LNCS. Springer Berlin Heidelberg. Chap. 1, pp. 9–48. DOI: 10.1007/978-3-642-35289-8_3 (cit. on pp. 49, 51, 53–57, 98).

LeDoux, J. E. (1992). « Brain Mechanisms of Emotion and Emotional Learning ». *Current Opinion in Neurobiology* 2(2), pp. 191–197. DOI: 10.1016/0959-4388(92)90011-9 (cit. on pp. 44, 110–112).

LeDoux, J. E. (2000). « Emotion Circuits in the Brain ». *Annual Review of Neuroscience* 23(1), pp. 155–184. DOI: 10.1146/annurev.neuro.23.1.155 (cit. on pp. 23, 27).

LeDoux, J. E. (2007). « The Amygdala ». *Current Biology* 17(20), R868–R874. DOI: 10.1016/j.cub.2007.08.005 (cit. on pp. 23, 25, 26, 28, 29, 111–113, 124).

LeDoux, J. E. (2012). « Rethinking the Emotional Brain ». *Neuron* 73(4), pp. 653–676. DOI: 10.1016/j.neuron.2012.02.004 (cit. on pp. 8–10, 24, 26, 29, 30).

Li, G., Nair, S. S., and Quirk, G. J. (2009). « A Biologically Realistic Network Model of Acquisition and Extinction of Conditioned Fear Associations in Lateral Amygdala Neurons ». *Journal of Neurophysiology* 101(3), pp. 1629–1646. DOI: 10.1152/jn.90765.2008 (cit. on p. 111).

López-González, R. (2008). « Neural Networks for Variational Problems in Engineering ». PhD thesis. Department of Computer Languages and Systems, Technical University of Catalonia, pp. 237+ (cit. on p. 48).

Louloudi, A., Mosallam, A., Marturi, N., Janse, P., and Hernandez, V. (2010). « Integration of the Humanoid Robot Nao inside a Smart Home: A Case Study ». In: *Swedish AI Society Workshop (SAIS).* Vol. 48. Linköping Electronic Conference Proceedings. Uppsala University. Uppsala, Sweden: Linköping University Electronic Press, pp. 35–44 (cit. on p. 72).

Lowe, R., Humphries, M., and Ziemke, T. (2009). « The Dual-Route Hypothesis: Evaluating a Neurocomputational Model of Fear Conditioning in Rats ». *Connection Science* 21(1), pp. 15–37. DOI: 10.1080/09540090802414085 (cit. on pp. 44, 45, 120).

Lowe, R., Mannella, F., Ziemke, T., and Baldassarre, G. (2011). « Modelling Coordination of Learning Systems: A Reservoir Systems Approach to Dopamine Modulated Pavlovian Conditioning ». In: *Advances in Artificial Life. Darwin Meets von Neumann. European Conference on Artificial Life (ECAL).* Vol. 5777. LNCS. Budapest, Hungary: Springer-Verlag. Chap. 51, pp. 410–417. DOI: 10.1007/978-3-642-21283-3_51 (cit. on pp. 45, 110–112, 115, 118, 124).

Lowe, R. and Ziemke, T. (2013). « Exploring the Relationship of Reward and Punishment in Reinforcement Learning ». In: *IEEE Symposium on Adaptive Dynamic Programming And Reinforcement Learning (ADPRL).* Singapore: IEEE, pp. 140–147. DOI: 10.1109/ADPRL.2013.6615000 (cit. on pp. 91, 92, 107, 127).

Lukoševičius, M. (2007). *Echo State Networks with Trained Feedbacks.* Tech. rep. 4. Bremen, Germany: School of Engineering and Science, Jacobs University, pp. 38+ (cit. on p. 70).

Lukoševičius, M. and Jaeger, H. (2009). « Reservoir Computing Approaches to Recurrent Neural Network Training ». *Computer Science Review.* Survey 3(3), pp. 127–149. DOI: 10.1016/j.cosrev.2009.03.005 (cit. on pp. 65, 67, 69, 70, 118).

M

Maass, W., Natschläger, T., and Markram, H. (2002). « Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations ». *Neural Computation* 14(11), pp. 2531–2560. DOI: 10.1162/089976602760407955 (cit. on p. 65).

Macnab, R. M. and Koshland, D. E. (1972). « The Gradient-Sensing Mechanism in Bacterial Chemotaxis ». *Proceedings of the National Academy of Sciences* 69(9), pp. 2509–2512 (cit. on p. 8).

Mai, J. K. and Forutan, F. (2012). « Thalamus ». In: *The Human Nervous System.* Third ed. Academic Press/Elsevier. Chap. 19, pp. 618–677. DOI: 10.1016/b978-0-12-374236-0.10019-7 (cit. on pp. 20, 21).

Mannella, F., Zappacosta, S., Mirolli, M., and Baldassarre, G. (2008). « A Computational Model of the Amygdala Nuclei's Role in Second Order Conditioning ». In: *From Animals to Animats: International Conference on the Simulation of Adaptive Behaviour (SAB).* Vol. 5040. LNCS. Osaka, Japan: Springer-Verlag. Chap. 32, pp. 321–330. DOI: 10.1007/978-3-540-69134-1_32 (cit. on pp. 44, 46, 111).

Marsland, S. (2009). *Machine Learning: An Algorithmic Perspective.* First ed. Machine Learning and Pattern Recognition. New Jersey, USA: Chapman and Hall Book / CRC Press, pp. 406+ (cit. on pp. 49, 51, 53, 54, 57).

Martin-Soelch, C., Linthicum, J., and Ernst, M. (2007). « Appetitive Conditioning: Neural Bases and Implications for Psychopathology ». *Neuroscience & Biobehavioral Reviews.* Review 31(3), pp. 426–440. DOI: 10.1016/j.neubiorev.2006.11.002 (cit. on p. 72).

Matsunaga, N., Akayama, S., and Kawaji, S. (2008). « Pain Generation Model on Upper Limb Considering the Laminated Structure of Skin ». In: *IEEE International Conference on Systems, Man and Cybernetics (SMC).* Singapore: IEEE, pp. 1097–1102. DOI: 10.1109/icsmc.2008.4811428 (cit. on pp. 39, 40).

Matsunaga, N., Akayama, S., and Kawaji, S. (2012). « Slow Pain Generation Model Caused by Mechanical Stimulus Based on the Laminated Structure of Skin ». *Electrical Engineering in Japan* 178(3), pp. 31–41. DOI: 10.1002/eej.21161 (cit. on pp. 39, 40).

Matsunaga, N., Kuroki, A., and Kawaji, S. (2005). « Superficial Pain Model using ANNs and its Application to Robot Control ». In: *IEEE/ASME International Conference on Advanced Intelligent Mechatronics.* Monterey, CA, Mexico: IEEE, pp. 664–669. DOI: 10.1109/aim.2005.1511058 (cit. on p. 40).

McCarthy, J. and Hayes, P. J. (1969). « Some Philosophical Problems from the Standpoint of Artificial Intelligence ». In: *Machine Intelligence 4.* Edinburgh University Press, pp. 463–502 (cit. on p. 31).

McFarland, D. (2009). *Guilty Robots, Happy Dogs: The Question of Alien Minds.* Artificial Intelligence. Oxford University Press, pp. 272+ (cit. on pp. 33, 34, 72).

McGrath, P. A. (1994). « Psychological Aspects of Pain Perception ». *Archives of Oral Biology* 39(Supplement), S55–S62. DOI: 10.1016/0003-9969(94)90189-9 (cit. on p. 11).

Mcintyre, C. K., Power, A. E., Roozendaal, B., and McGaugh, J. L. (2003). « Role of the Basolateral Amygdala in Memory Consolidation ». *Annals of the New York Academy of Sciences.* The Amygdala in Brain Function: Basic and Clinical Approaches 985(1), pp. 273–293. DOI: 10.1111/j.1749-6632.2003.tb07088.x (cit. on p. 26).

Mega, M. S., Cummings, J. L., Salloway, S., and Malloy, P. (1997). « The Limbic System: An Anatomic, Phylogenetic, and Clinical Perspective ». *The Journal of Neuropsychiatry and Clinical Neurosciences.* Anatomy and Neurochemistry 9(3), pp. 315–330 (cit. on p. 22).

Melhuish, C., Ieropoulos, I., Greenman, J., and Horsfield, I. (2006). « Energetically Autonomous Robots: Food for Thought ». *Autonomous Robots* 21(3), pp. 187–198. DOI: 10.1007/s10514-006-6574-5 (cit. on p. 35).

Miller, K. D. and MacKay, D. J. C. (1994). « The Role of Constraints in Hebbian Learning ». *Neural Computation* 6(1), pp. 100–126. DOI: 10.1162/neco.1994.6.1.100 (cit. on p. 50).

Minsky, M. (1961). « Steps Toward Artificial Intelligence ». *Proceedings of the IRE* 49(1), pp. 8–30. DOI: 10.1109/JRPROC.1961.287775 (cit. on p. 59).

Mirolli, M., Mannella, F., and Baldassarre, G. (2010). « The Roles of the Amygdala in the Affective Regulation of Body, Brain, and Behaviour ». *Connection Science.* Special Issue: Affective Robotics 22(3), pp. 215–245. DOI: 10.1080/09540091003682553 (cit. on pp. 30, 62).

Mitchell, M. (1998). *An Introduction to Genetic Algorithms.* First ed. Complex Adaptive Systems. A Bradford Book/The MIT Press (cit. on p. 99).

Mitsuhashi, N., Fujieda, K., Tamura, T., Kawamoto, S., Takagi, T., and Okubo, K. (2009). « BodyParts3D: 3D Structure Database for Anatomical Concepts ». *Nucleic Acids Research* 37(suppl 1), pp. D782–D785. DOI: 10.1093/nar/gkn613 (cit. on pp. 15, 19, 23).

Modha, D. S. and Singh, R. (2010). « Network Architecture of the Long-Distance Pathways in the Macaque Brain ». *Proceedings of the National Academy of Sciences* 107(30), pp. 13485–13490. DOI: 10.1073/pnas.1008054107 (cit. on p. 34).

Moore, A. W. and Atkeson, C. G. (1993). « Prioritized Sweeping: Reinforcement Learning with Less Data and Less Time ». *Machine Learning* 13(1), pp. 103–130. DOI: 10.1007/BF00993104 (cit. on p. 61).

Morén, J. (2002). « Emotion and Learning: A Computational Model of the Amygdala ». PhD thesis. Lund, Sweden: Lund University Cognitive Studies, pp. 160+ (cit. on pp. 27–29, 32, 58).

Morrison, S. E. and Salzman, C. D. (2010). « Re-Valuing the Amygdala ». *Current Opinion in Neurobiology.* Cognitive Neuroscience 20(2), pp. 221–230. DOI: 10.1016/j.conb.2010.02.007 (cit. on p. 24).

Murata, N., Müller, K.-R., Ziehe, A., and Amari, S.-i. (1996). « Adaptive On-line Learning in Changing Environments ». In: *Advances in Neural Information Processing Systems (NIPS).* Algorithms and Architecture No. 9. Denver, CO, USA: The MIT Press, pp. 599–605 (cit. on pp. 56, 57).

Murphy, R. and Woods, D. D. (2009). « Beyond Asimov: The Three Laws of Responsible Robotics ». *IEEE Intelligent Systems* 24(4), pp. 14–20. DOI: 10.1109/mis.2009.69 (cit. on pp. 2, 39).

Murray, E. A. (2007). « The Amygdala, Reward and Emotion ». *Trends in Cognitive Sciences.* Cognitive-Emotional Interactions 11(11), pp. 489–497. DOI: 10.1016/j.tics.2007.08.013 (cit. on p. 26).

Muse, D. and Wermter, S. (2009). « Actor-Critic Learning for Platform-Independent Robot Navigation ». *Cognitive Computation* 1(3), pp. 203–220. DOI: 10.1007/s12559-009-9021-z (cit. on p. 76).

N

Navarro, N., Lowe, R., Weber, C., and Wermter, S. (2011). « Many-Routes Hypothesis of Fear Conditioning: A Dynamical Reservoir Based Approach ». In: *Marie-Curie Researchers Symposium Poster.* Warsaw, Poland (cit. on p. 152).

Navarro, N., Weber, C., and Wermter, S. (2011). « Real-World Reinforcement Learning for Autonomous Humanoid Robot Charging in a Home Environment ». In: *Annual Conference Towards Autonomous Robotic Systems (TAROS).* Vol. 6856. LNCS. Sheffield, UK: Springer Berlin Heidelberg, pp. 231–240. DOI: 10.1007/978-3-642-23232-9_21 (cit. on pp. 74, 152).

Navarro-Guerrero, N., Lowe, R., and Wermter, S. (2012). « A Neurocomputational Amygdala Model of Auditory Fear Conditioning: A Hybrid System Approach ». In: *International Joint Conference on Neural Networks (IJCNN).* Brisbane, QLD, Australia: IEEE, pp. 214–221. DOI: 10.1109/IJCNN.2012.6252392 (cit. on p. 152).

Navarro-Guerrero, N., Weber, C., Schroeter, P., and Wermter, S. (2012). « Real-World Reinforcement Learning for Autonomous Humanoid Robot Docking ». *Robotics and Autonomous Systems* 60(11), pp. 1400–1407. DOI: 10.1016/j.robot.2012.05.019 (cit. on pp. 74, 152).

Neal, M., Feyereisl, J., Rascunà, R., and Wang, X. (2006). « Don't Touch Me, I'm Fine: Robot Autonomy Using an Artificial Innate Immune System ». In: *International Conference on Artificial Immune Systems (ICARIS).* Vol. 4163. LNCS. Oeiras, Portugal: Springer Berlin Heidelberg, pp. 349–361. DOI: 10.1007/11823940_27 (cit. on p. 36).

O

OpenStax College (2013). *Anatomy & Physiology.* Connexions Web site (cit. on pp. 13, 15, 18, 19, 22).

Ozturk, M. C., Xu, D., and Príncipe, J. C. (2007). « Analysis and Design of Echo State Networks ». *Neural Computation* 19(1), pp. 111–138. DOI: 10.1162/neco.2007.19.1.111 (cit. on pp. 65, 67–69).

P

Pan, W.-X., Schmidt, R., Wickens, J. R., and Hyland, B. I. (2005). « Dopamine Cells Respond to Predicted Events during Classical Conditioning: Evidence for Eligibility Traces in the Reward-Learning Network ». *The Journal of Neuroscience* 25(26), pp. 6235–6242. DOI: 10.1523/jneurosci.1478-05.2005 (cit. on p. 64).

Pape, H.-C. (2010). « Petrified or Aroused with Fear: The Central Amygdala Takes the Lead ». *Neuron* 67(4), pp. 527–529. DOI: 10.1016/j.neuron.2010.08.009 (cit. on pp. 24–28, 112, 113).

Pape, H.-C. and Pare, D. (2010). « Plastic Synaptic Networks of the Amygdala for the Acquisition, Expression, and Extinction of Conditioned Fear ». *Physiological Reviews* 90(2), pp. 419–463. DOI: 10.1152/physrev.00037.2009 (cit. on pp. 23–27, 110, 112, 113, 124).

Paré, D. (2003). « Role of the Basolateral Amygdala in Memory Consolidation ». *Progress in Neurobiology* 70(5), pp. 409–420. DOI: 10.1016/s0301-0082(03)00104-7 (cit. on p. 26).

Parisi, D. (2004). « Internal Robotics ». *Connection Science* 16(4), pp. 325–338. DOI: 10.1080/09540090412331314768 (cit. on pp. 30, 32).

Pavlou, A. and Casey, M. (2009). « A Computational Platform for Visual Fear Conditioning ». In: *International Joint Conference on Neural Networks (IJCNN).* Atlanta, GA, USA: IEEE, pp. 2451–2458. DOI: 10.1109/ijcnn.2009.5178718 (cit. on p. 111).

Paxinos, G., Xu-Feng, H., Sengul, G., and Watson, C. (2012). « Organization of Brainstem Nuclei ». In: *The Human Nervous System.* Third ed. Academic Press/Elsevier. Chap. 8, pp. 260–327. DOI: 10.1016/b978-0-12-374236-0.10008-2 (cit. on pp. 14, 15, 17).

Pessoa, L. (2010). « Emotion and Cognition and the Amygdala: From 'What is it?' to 'What's to be Done?' » *Neuropsychologia* 48(12), pp. 3416–3429. DOI: 10.1016/j.neuropsychologia.2010.06.038 (cit. on pp. 23, 24, 26–29).

Pessoa, L. and Adolphs, R. (2010). « Emotion Processing and the Amygdala: From a 'Low Road' to 'Many Roads' of Evaluating Biological Significance ». *Nature Reviews. Neuroscience* 11(11), pp. 773–783. DOI: 10.1038/nrn2920 (cit. on pp. 20, 25).

Pfeifer, R. and Bongard, J. (2006). *How the Body Shapes the Way We Think: A New View of Intelligence.* Computer Science and Intelligent Systems. Cambridge, MA, USA: The MIT Press, pp. 418+ (cit. on pp. 31, 49).

Prescott, T. J., Redgrave, P., and Gurney, K. (1999). « Layered Control Architectures in Robots and Vertebrates ». *Adaptive Behavior* 7(1), pp. 99–127. DOI: 10.1177/105971239900700105 (cit. on p. 30).

Principe, J. C., Xu, D., and Fisher, J. W. (2000). « Information-Theoretic Learning ». In: *Unsupervised Adaptive Filtering: Blind Source Separation.* Vol. 1. Adaptive and Learning Systems for Signal Processing, Communications and Control. John Wiley & Sons, Inc. Chap. 7, pp. 265–319 (cit. on p. 68).

Provost, J., Kuipers, B. J., and Miikkulainen, R. (2004). « Self-Organizing Perceptual and Temporal Abstraction for Robot Reinforcement Learning ». In: *AAAI Workshop on Learning and Planning in Markov Processes - Advances and Challenges.* San Jose, CA, USA: The AAAI Press, pp. 79–84 (cit. on p. 77).

Purves, D., Augustine, G. J., Fitzpatrick, D., Hall, W. C., LaMantia, A.-S., and White, L. E. (2012). *Neuroscience.* Fifth ed. Sinauer Associates, Inc. (cit. on pp. 14, 15, 22).

R

Ren, Y., Zou, W., Fan, H., Ye, A., Yuan, K., and Ma, Y. (2012). « A Docking Control Method in Narrow Space for Intelligent Wheelchair ». In: *International Conference on Mechatronics and Automation (ICMA).* Chengdu, China: IEEE, pp. 1615–1620. DOI: 10.1109/icma.2012.6284378 (cit. on p. 38).

Rescorla, R. A. and Wagner, A. W. (1972). « A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement ». In: *Classical Conditioning II: Current Research and Theory.* New York, NY, USA: Appleton-Century-Crofts. Chap. 3, pp. 64–99 (cit. on p. 45).

Resnik, J., Sobel, N., and Paz, R. (2011). « Auditory Aversive Learning Increases Discrimination Thresholds ». *Nature Neuroscience* 14(6), pp. 791–796. DOI: 10.1038/nn.2802 (cit. on pp. 8, 44, 121, 123, 124, 128).

Riedmiller, M. and Braun, H. (1993). « A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm ». In: *IEEE International Conference on Neural Networks.* Vol. 1. San Francisco, CA, USA: IEEE, pp. 586–591. DOI: 10.1109/icnn.1993.298623 (cit. on p. 54).

RobotDoC (2009-2014). *The RobotDoC Collegium: The Marie Curie doctoral training network in developmental robotics.* URL: http://robotdoc.org/ (cit. on p. 154).

Rojas, R. (1996). *Neural Networks - A Systematic Introduction.* Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 502+. DOI: 10.1007/978-3-642-61068-4 (cit. on p. 50).

Ruiz-del-Solar, J., Moya, J., and Parra-Tsunekawa, I. (2010). « Fall Detection and Management in Biped Humanoid Robots ». In: *IEEE International Conference on Robotics and Automation (ICRA).* Anchorage, AK, USA: IEEE, pp. 3323–3328. DOI: 10.1109/robot.2010.5509550 (cit. on pp. 3, 4, 41, 43).

Ruiz-del-Solar, J., Palma-Amestoy, R., Marchant, R., Parra-Tsunekawa, I., and Zegers, P. (2009). « Learning to Fall: Designing Low Damage Fall Sequences for Humanoid Soccer Robots ». *Robotics and Autonomous Systems.* Humanoid Soccer Robots 57(8), pp. 796–807. DOI: 10.1016/j.robot.2009.03.011 (cit. on p. 43).

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). « Learning Representations by Back-Propagating Errors ». *Nature* 323(6088), pp. 533–536. DOI: 10.1038/323533a0 (cit. on pp. 50, 51).

S

Saeb, S., Weber, C., and Triesch, J. (2009). « Goal-Directed Learning of Features and Forward Models ». *Neural Networks.* Special Issue: Advances in Neural Networks Research: International Joint Conference on Neural Networks (IJCNN2009) 22(5-6), pp. 586–592. DOI: 10.1016/j.neunet.2009.06.049 (cit. on p. 77).

Schaffer, J. D., Caruana, R. A., Eshelman, L. J., and Das, R. (1989). « A Study of Control Parameters Affecting Online Performance of Genetic Algorithms for Function Optimization ». In: *International Conference on Genetic Algorithms.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 51–60 (cit. on p. 99).

Schultz, W. (1998). « Predictive Reward Signal of Dopamine Neurons ». *Journal of Neurophysiology* 80(1), pp. 1–27 (cit. on p. 16).

Schultz, W., Dayan, P., and Montague, P. R. (1997). « A Neural Substrate of Prediction and Reward ». *Science* 275(5306), pp. 1593–1599. DOI: 10.1126/science.275.5306.1593 (cit. on pp. 62, 72).

Seymour, B., O'Doherty, J. P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K., and Dolan, R. (2005). « Opponent Appetitive-Aversive Neural Processes Underlie Predictive Learning of Pain Relief ». *Nature Neuroscience* 8(9), pp. 1234–1240. DOI: 10.1038/nn1527 (cit. on p. 91).

Shi, Z. and Han, M. (2007). « Support Vector Echo-State Machine for Chaotic Time-Series Prediction ». *IEEE Transactions on Neural Networks* 18(2), pp. 359–372. DOI: 10.1109/tnn.2006.885113 (cit. on p. 70).

Shimizu, T., Saegusa, R., Ikemoto, S., Ishiguro, H., and Metta, G. (2011). « Adaptive Self-Protective Motion Based on Reflex Control ». In: *International Joint Conference on Neural Networks (IJCNN).* San Jose, CA, USA: IEEE, pp. 2860–2864. DOI: 10.1109/ijcnn.2011.6033596 (cit. on pp. 42, 43).

Shimizu, T., Saegusa, R., Ikemoto, S., Ishiguro, H., and Metta, G. (2012). « Self-Protective Whole Body Motion for Humanoid Robots Based on Synergy of Global Reaction and Local Reflex ». *Neural Networks.* Special Issue: Selected Papers from IJCNN 2011 32, pp. 109–118. DOI: 10.1016/j.neunet.2012.02.011 (cit. on pp. 3, 4, 42, 43).

Skowronski, M. D. and Harris, J. G. (2007). « Automatic Speech Recognition using a Predictive Echo State Network Classifier ». *Neural Networks.* Special Issue: Echo State Networks and Liquid State Machines 20(3), pp. 414–423. DOI: 10.1016/j.neunet.2007.04.006 (cit. on p. 70).

Staahl, C. and Drewes, A. M. (2004). « Experimental Human Pain Models: A Review of Standardised Methods for Preclinical Testing of Analgesics ». *Basic & Clinical Pharmacology & Toxicology* 95(3), pp. 97–111. DOI: 10.1111/j.1742-7843.2004.950301.x (cit. on p. 11).

Stahlhut, C., Navarro-Guerrero, N., Weber, C., and Wermter, S. (2015a). « Interaction in Reinforcement Learning Reduces the Need for Finely Tuned Hyperparameters in Complex Tasks ». *Kognitive Systeme* 3(2). DOI: 10.17185/duepublico/40718 (cit. on pp. 93, 152).

Stahlhut, C., Navarro-Guerrero, N., Weber, C., and Wermter, S. (2015b). « Interaction Is More Beneficial in Complex Reinforcement Learning Problems Than in Simple Ones ». In: *4. Interdisziplinärer Workshop Kognitive Systeme: Mensch, Teams, Systeme und Automaten.* Bielefeld, Germany, pp. 142–150 (cit. on p. 152).

Stent, G. S. (1973). « A Physiological Mechanism for Hebb's Postulate of Learning ». *Proceedings of the National Academy of Sciences* 70(4), pp. 997–1001 (cit. on pp. 50, 115, 117).

Sternson, S. M. (2013). « Hypothalamic Survival Circuits: Blueprints for Purposive Behaviors ». *Neuron* 77(5), pp. 810–824. DOI: 10.1016/j.neuron.2013.02.018 (cit. on pp. 8, 10, 18, 20, 30, 33).

Stevenson, C. W. and Gratton, A. (2003). « Basolateral Amygdala Modulation of the Nucleus Accumbens Dopamine Response to Stress: Role of the Medial Prefrontal Cortex ». *European Journal of Neuroscience* 17(6), pp. 1287–1295. DOI: 10.1046/j.1460-9568.2003.02560.x (cit. on p. 114).

Suri, R. E. (2002). « TD Models of Reward Predictive Responses in Dopamine Neurons ». *Neural Networks* 15(4-6), pp. 523–533. DOI: 10.1016/S0893-6080(02)00046-1 (cit. on p. 93).

Sutton, R. S. (1984). « Temporal Credit Assignment in Reinforcement Learning ». PhD thesis. Department of Computer Science, University of Massachusetts, pp. 223+ (cit. on p. 59).

Sutton, R. S. (1988). « Learning to Predict by the Methods of Temporal Differences ». *Machine Learning* 3(1), pp. 9–44. DOI: 10.1007/bf00115009 (cit. on p. 62).

Sutton, R. S. (1991a). « Dyna, An Integrated Architecture for Learning, Planning, and Reacting ». *ACM SIGART Bulletin* 2(4), pp. 160–163. DOI: 10.1145/122344.122377 (cit. on p. 61).

Sutton, R. S. (1991b). « Planning by Incremental Dynamic Programming ». In: *International Workshop on Machine Learning (ML).* Learning Reaction Strategies. Evanston, IL, USA: Morgan Kaufmann, pp. 353–357 (cit. on p. 61).

Sutton, R. S. (1992). « Adapting Bias by Gradient Descent: An Incremental Version of Delta-Bar-Delta ». In: *National Conference on Artificial Intelligence.* Learning: Neural Network and Hybrid. Association for the Advancement of Artificial Intelligence. San Jose, CA, USA: The AAAI Press, pp. 171–176 (cit. on p. 56).

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction.* First ed. Adaptive Computation and Machine Learning. Cambridge, MA, USA: A Bradford Book/The MIT Press, pp. 342+ (cit. on pp. 50, 58–64, 76–78, 80, 81, 88, 91, 93, 95, 96, 107, 126).

Suzuki, K. A. O., Kemper Filho, P., and Morrison, J. R. (2012). « Automatic Battery Replacement System for UAVs: Analysis and Design ». *Journal of Intelligent & Robotic Systems* 65(1-4), pp. 563–586. DOI: 10.1007/s10846-011-9616-y (cit. on p. 37).

Swanson, L. W. and Petrovich, G. D. (1998). « What is the Amygdala? » *Trends in Neurosciences* 21(8), pp. 323–331. DOI: 10.1016/s0166-2236(98)01265-x (cit. on p. 23).

Szita, I., Gyenes, V., and Lőrincz, A. (2006). « Reinforcement Learning with Echo State Networks ». In: *International Conference on Artificial Neural Networks (ICANN).* Vol. 4131. LNCS. Athens, Greece: Springer Berlin Heidelberg. Chap. 86, pp. 830–839. DOI: 10.1007/11840817_86 (cit. on p. 70).

T

Thelen, E., Corbetta, D., Kamm, K., Spencer, J. P., Schneider, K., and Zernicke, R. F. (1993). « The Transition to Reaching: Mapping Intention and Intrinsic Dynamics ». *Child Development* 64(4), pp. 1058–1098. DOI: 10.1111/j.1467-8624.1993.tb04188.x (cit. on p. 93).

V

van der Wal, E. (2012). « Object Grasping with the NAO ». MA thesis. Faculty of Mathematics and Natural Sciences Artificial Intelligence, University of Groningen, The Netherlands, pp. 1–76 (cit. on p. 92).

van Hasselt, H. and Wiering, M. A. (2007). « Reinforcement Learning in Continuous Action Spaces ». In: *IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning (ADPRL).* Honolulu, Hi, USA: IEEE, pp. 272–279. DOI: 10.1109/adprl.2007.368199 (cit. on pp. 94, 96, 127).

Vlachos, I., Herry, C., Lüthi, A., Aertsen, A., and Kumar, A. (2011). « Context-Dependent Encoding of Fear and Extinction Memories in a Large-Scale Network Model of the Basal Amygdala ». *PLoS Computational Biology* 7(3), e1001104+. DOI: 10.1371/journal.pcbi.1001104 (cit. on pp. 44, 46, 111).

Vollmer, A.-L., Ruciński, M., Alessandro, C., Wilkinson, N., Navarro-Guerrero, N., and Handl, A. (2013). « Special Session: Training in Robotics for Development of Cognition (RobotDoC) ». In: *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL).* Osaka, Japan (cit. on p. 153).

von der Malsburg, C. (1973). « Self-Organization of Orientation Sensitive Cells in the Striate Cortex ». *Kybernetik* 14(2), pp. 85–100. DOI: 10.1007/bf00288907 (cit. on p. 51).

W

Wächter, T., Lungu, O. V., Liu, T., Willingham, D. T., and Ashe, J. (2009). « Differential Effect of Reward and Punishment on Procedural Learning ». *The Journal of Neuroscience* 29(2), pp. 436–443. DOI: 10.1523/JNEUROSCI.4132-08.2009 (cit. on pp. 90–92, 107).

Watkins, C. J. C. H. (1989). « Learning from Delayed Rewards ». PhD thesis. Cambridge University, pp. 241+ (cit. on p. 61).

Watkins, C. J. C. H. and Dayan, P. (1992). « Q-Learning ». *Machine Learning* 8(3-4), pp. 279–292. DOI: 10.1007/bf00992698 (cit. on p. 61).

Weber, C., Elshaw, M., Wermter, S., Triesch, J., and Willmot, C. (2008). « Reinforcement Learning Embedded in Brains and Robots ». In: *Reinforcement Learning: Theory and Applications.* First ed. Artificial Intelligence. Vienna, Austria: I-TECH Education and Publishing. Chap. 7, pp. 119–142. DOI: 10.5772/5278 (cit. on p. 80).

Weber, C. and Triesch, J. (2009). « Goal-Directed Feature Learning ». In: *International Joint Conference on Neural Networks (IJCNN).* Georgia Tech Georgia Institute of Technology. Atlanta, GA USA: IEEE, pp. 3355–3362. DOI: 10.1109/IJCNN.2009.5179064 (cit. on p. 77).

Weber, C., Wermter, S., and Zochios, A. (2004). « Robot Docking with Neural Vision and Reinforcement ». In: *SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence.* Vol. 5. Cambridge, UK: Springer London, pp. 213–226. DOI: 10.1007/978-1-4471-0643-2_15 (cit. on p. 92).

Weinberger, N. M. (2011). « The Medial Geniculate, not the Amygdala, as the Root of Auditory Fear Conditioning ». *Hearing Research.* Central Auditory Pathways - A Tribute to Jeffery A. Winer 274(1-2), pp. 61–74. DOI: 10.1016/j.heares.2010.03.093 (cit. on pp. 20, 25, 113, 116).

Westlund, K. N. and Sluka, K. A. (2013). « Nocifensive Behaviors, Muscle and Joint ». In: *Encyclopedia of Pain.* Springer Berlin Heidelberg, pp. 2284–2289. DOI: 10.1007/978-3-642-28753-4_2799 (cit. on p. 11).

Westlund, K. N. and Willis, W. D. (2012). « Pain System ». In: *The Human Nervous System.* Third ed. Academic Press/Elsevier. Chap. 32, pp. 1144–1186. DOI: 10.1016/b978-0-12-374236-0.10032-x (cit. on pp. 11–14).

Whalen, P. J. (1998). « Fear, Vigilance, and Ambiguity: Initial Neuroimaging Studies of the Human Amygdala ». *Current Directions in Psychological Science* 7(6), pp. 177–188. DOI: 10.2307/20182537 (cit. on pp. 23, 24, 27, 28).

Wilkinson, J. L. (1992a). « Brainstem ». In: *Neuroanatomy for Medical Students.* Second ed. Butterworth-Heinemann/Elsevier. Chap. 5, pp. 82–109. DOI: 10.1016/b978-0-7506-1447-4.50009-8 (cit. on pp. 15–21).

Wilkinson, J. L. (1992b). « Diencephalon and Internal Capsule ». In: *Neuroanatomy for Medical Students.* Second ed. Butterworth-Heinemann/Elsevier. Chap. 8, pp. 164–186. DOI: 10.1016/b978-0-7506-1447-4.50012-8 (cit. on p. 18).

Wörgötter, F. and Porr, B. (2005). « Temporal Sequence Learning, Prediction, and Control: A Review of Different Models and Their Relation to Biological Mechanisms ». *Neural Computation* 17(2), pp. 245–319. DOI: 10.1162/0899766053011555 (cit. on p. 58).

X

Xu, D. and Erdogmuns, D. (2010). « Renyi's Entropy, Divergence and Their Nonparametric Estimators ». In: *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives.* Information Science and Statistics. Springer New York. Chap. 2, pp. 47–102. DOI: 10.1007/978-1-4419-1570-2_2 (cit. on p. 68).

Xue, Y., Yang, L., and Haykin, S. (2007). « Decoupled Echo State Networks with Lateral Inhibition ». *Neural Networks.* Special Issue: Echo State Networks and Liquid State Machines 20(3), pp. 365–376. DOI: 10.1016/j.neunet.2007.04.014 (cit. on p. 69).

Y

Yan, W., Weber, C., and Wermter, S. (2012). « A Neural Approach for Robot Navigation Based on Cognitive Map Learning ». In: *International Joint Conference on Neural Networks (IJCNN).* Brisbane, QLD, Australia: IEEE, pp. 1146–1153. DOI: 10.1109/ijcnn.2012.6252522 (cit. on pp. 3, 4, 75).

Yan, W., Weber, C., and Wermter, S. (2013). « Learning Indoor Robot Navigation using Visual and Sensorimotor Map Information ». *Frontiers in Neurorobotics* 7(15). DOI: 10.3389/fnbot.2013.00015 (cit. on pp. 3, 4, 75).

Yildiz, I. B., Jaeger, H., and Kiebel, S. J. (2012). « Re-Visiting the Echo State Property ». *Neural Networks* 35, pp. 1–9. DOI: 10.1016/j.neunet.2012.07.005 (cit. on pp. 67, 68).

Z

Zang, P., Tian, R., Thomaz, A. L., and Isbell, C. L. (2010). « Batch Versus Interactive Learning by Demonstration ». In: *IEEE International Conference on Development and Learning (ICDL).* Ann Arbor, MI, USA: IEEE, pp. 219–224. DOI: 10.1109/devlrn.2010.5578841 (cit. on pp. 77, 78).

Zhang, J., Song, G., Li, Y., Qiao, G., and Li, Z. (2013). « Battery Swapping and Wireless Charging for a Home Robot System with Remote Human Assistance ». *IEEE Transactions on Consumer Electronics* 59(4), pp. 747–755. DOI: 10.1109/tce.2013.6689685 (cit. on p. 37).

Zhou, W. and Coggins, R. (2002). « Computational Models of the Amygdala and the Orbitofrontal Cortex: A Hierarchical Reinforcement Learning System for Robotic Control ». In: *Australian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence.* Vol. 2557. LNCS. Canberra, ACT, Australia: Springer Berlin/Heidelberg. Chap. 37, pp. 419–430. DOI: 10.1007/3-540-36187-1_37 (cit. on pp. 46, 111).

Ziemke, T. and Lowe, R. (2009). « On the Role of Emotion in Embodied Cognitive Architectures: From Organisms to Robots ». *Cognitive Computation* 1(1), pp. 104–117. DOI: 10.1007/s12559-009-9012-0 (cit. on pp. 2, 30–33, 36, 126).

B Appendix
Publications Originating from this Thesis

Three papers and an abstract have been published for presenting the concepts of this research.

Navarro-Guerrero, N., Weber, C., Schroeter, P., and Wermter, S. (2012). « Real-World Reinforcement Learning for Autonomous Humanoid Robot Docking ». *Robotics and Autonomous Systems* 60(11), pp. 1400–1407. DOI: 10.1016/j.robot.2012.05.019

Navarro, N., Weber, C., and Wermter, S. (2011). « Real-World Reinforcement Learning for Autonomous Humanoid Robot Charging in a Home Environment ». In: *Annual Conference Towards Autonomous Robotic Systems (TAROS).* vol. 6856. LNCS. Sheffield, UK: Springer Berlin Heidelberg, pp. 231–240. DOI: 10.1007/978-3-642-23232-9_21

Navarro-Guerrero, N., Lowe, R., and Wermter, S. (2012). « A Neurocomputational Amygdala Model of Auditory Fear Conditioning: A Hybrid System Approach ». In: *International Joint Conference on Neural Networks (IJCNN).* Brisbane, QLD, Australia: IEEE, pp. 214–221. DOI: 10.1109/IJCNN.2012.6252392

Navarro, N., Lowe, R., Weber, C., and Wermter, S. (2011). « Many-Routes Hypothesis of Fear Conditioning: A Dynamical Reservoir Based Approach ». In: *Marie-Curie Researchers Symposium Poster.* Warsaw, Poland

Other publications not included in this thesis:

Stahlhut, C., Navarro-Guerrero, N., Weber, C., and Wermter, S. (2015a). « Interaction in Reinforcement Learning Reduces the Need for Finely Tuned Hyperparameters in Complex Tasks ». *Kognitive Systeme* 3(2). DOI: 10.17185/duepublico/40718

Stahlhut, C., Navarro-Guerrero, N., Weber, C., and Wermter, S. (2015b). « Interaction Is More Beneficial in Complex Reinforcement Learning Problems Than

in Simple Ones ». In: *4. Interdisziplinärer Workshop Kognitive Systeme: Mensch, Teams, Systeme und Automaten.* Bielefeld, Germany, pp. 142–150

Vollmer, A.-L., Ruciński, M., Alessandro, C., Wilkinson, N., Navarro-Guerrero, N., and Handl, A. (2013). « Special Session: Training in Robotics for Development of Cognition (RobotDoC) ». in: *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL).* Osaka, Japan

Arredondo, T., Freund, W., Navarro-Guerrero, N., and Castillo, P. (2013). « Fuzzy Motivations in a Multiple Agent Behaviour-Based Architecture ». *International Journal of Advanced Robotic Systems* 10(313), pp. 1–13. DOI: 10.5772/56578

Appendix

C Acknowledgements

I want to thank my supervisor Prof. Dr. Stefan Wermter, for his valuable advice and support during the ups and downs of my PhD studies. I also want to thank Dr. Cornelius Weber and Dr. Robert Lowe for the helpful discussions and feedback.

I want to thank all the members of the Knowledge Technology Group. Specially, to Dieter Jessen and Heidi Oskarsson for their help after my arrival to Hamburg. I also want to thank Katja Kösters, Erik Strahl, Reinhard Zierke and Tim Scharfenberg for their patience, prompt and skilled support.

I want to thank the RobotDoc Collegium for their amazing workshops and fascinating conversations. Particularly, I want to thank Prof. Dr. Angelo Cangelosi for his caring attitude towards all the fellows. I also want to thank the Cognition & Interaction Lab. from the University of Skövde, Sweden for their hospitality during my research secondment.

There are countless others, close and from far away, who have been there for me. Particularly, I want to thank my parents, sister and Regina Rentsch, to my friends Elena Villanueva, Gabriel Cisternas and Annika Kühn for her constant support and encouragement.

Throughout my studies, I was financially supported mostly by the EU project RobotDoC, under 235065 ROBOT-DOC from the 7th Framework Programme, Marie Curie Action ITN and also by the University of Hamburg through my supervisor Prof. Dr. Stefan Wermter.

D Appendix

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Declaration on Oath

I hereby declare, on oath, that I have written the present dissertation by my own and have not used other than the acknowledged resources and aids.

Hamburg, den

Nicolás Ignacio Navarro Guerrero