# Literature Review

Word count: 1812

*Keep going in the same way if things are getting better, and otherwise move around.* (Selfridge, 1978).

[Introduction to follow]

The following literature review outlines the major contributions, to the development of the field *Reinforcement learning,* in a sparse format, in the light of the limited extent of this dissertation. The interested reader should consult (Sutton and Barto, 2018) and (Juliani, 2018) among others, on which the below is largely inspired, along with footnotes for more detail.

The main topic of this dissertation is reinforcement learning, more specifically used for autonomous navigation using spatial cues. The essence of reinforcement learning (RL) is very well captured in the quote from Selfridge above; behaviour associated with positive feedback should dominate behaviour associated with negative feedback, and ultimately result in an *optimal* behaviour.   
Reinforcement learning has its roots in multiple disciplines, with the most prominent being neuroscience and mathematics (Sutton and Barto, 2018). The largest contributor within the field of neuroscience is the subfield of *trial-and-error learning*, which is an area that extends into psychology, and the subfield is concerned with learning behaviour of animals. One of the major theoretical contributions was *Law of Effect[[1]](#footnote-1)*, proposed by Edward Thorndike in 1911, which describes the effect of reinforcing events on the tendency to select actions (Sutton and Barto, 2018).  
The greatest contribution from the field of mathematics comes from the subfield *Optimal Control Theory,* and as a derivative hereof *Dynamic programming*. The field of optimal control is concerned with dynamic systems, and the biggest contributor from this field is without doubt Richard Bellman and his work in the mid-1950s. Bellman is the farther of dynamic programming, a solution method for optimal control problems, and Markov decision processes (MDP), which is the discrete time stochastic version of the classic optimal control problem (Sutton and Barto, 2018).   
Trial-and-error learning and optimal control theory along with the concepts of MDPs is undoubtedly the corner stones of reinforcement learning, yet the computational embedding is what takes reinforcement learning from theory to practice. Furthermore, this is what has driven much of the recent progress with the field, which is touched upon in a subsequent paragraph.   
One of the first persons to embed trial-and-error learning in a computer, was no else than Alan Turing with his design for a *pleasure-pain system* in 1948*.*

Not much interest was shown in, what is today known as, reinforcement learning up doing the 1960s and 1970s, resulting in lack of research in the area. However, two major contributions made their way, namely Minsky’s paper *Steps towards Artificial Intelligence,* from 1961, and Harry Klopf research in general conducted around 1975, where the latter was the main source of inspiration for (Sutton and Barto, 1998, 2018) through their earlier work, which is the main reference on reinforcement learning today.  
Minsky addressed the problem today known as the *credit assignment problem*, the main objective of most modern reinforcement learning algorithms, which is the problem of correctly rewarding individual actions leading to the desired behaviour.   
Klopf’s influence was widespread and sparked many great papers in the two decades to follow his research, with some of the most prominent papers being (Sutton, 1984), (Anderson, 1986), (Sutton, 1988) and finally (Tesauro, 1995).   
The work of (Sutton, 1984) and further extended in (Anderson, 1986) was the first work on the actor-critic architecture, as it is known today.  
(Sutton, 1988) introduced what is today known as *TD-lambda*, inspired by the early work of Ian Witten in (Witten, 1976) which is some of the first work on temporal difference learning, which TD-lambda is a part of.

In the mid-1990s came perhaps one of the most influential papers on applied reinforcement learning up until that date, namely (Tesauro, 1995). Tesauro trained a reinforcement agent to play backgammon and achieved near grand master level over two extensions to his initial work (Tesauro, 1995). Tesauro’s work influenced the field of reinforcement learning along with the backgammon community, with the influence on the latter resulting in a change in playing style among the world’s best players (Tesauro, 1995)[[2]](#footnote-2). Tesauro’s influence on the reinforcement community came in form of the level of play that his agent achieved, which underlined the potential of temporal difference learning.  
Tesauro was not the first to use reinforcement learning to solve games nor the last, to say the least.

Reinforcement learning is well suited to be employed on a wide range of games, because the structure of many games resembles with the dynamics of reinforcement learning. Some of the first cases were (Shannon, 1950) and (Samuel, 1959), studying the games of chess and checkers respectively.  
More recent, and deeply influential, studies are (Bellemare et al., 2013) and in continuation hereof (Mnih et al., 2013).  
(Bellemare et al., 2013) introduced a platform for researchers, for the purpose of exploring existing algorithms and encourage progress in *domain-independent AI technology* (Bellemare et al., 2013). The platform, Arcade Learning Environment (ALE), provides researchers with easy access to hundreds of Atari 2600 games, which are suitable for benchmarking algorithms. One of the first to explore the opportunities embedded in ALE was (Mnih et al., 2013), with the introduction of a *deep* structure for the neural network to learn the Q-function, which at the same time became the start of the next era. Their motivation came from recent advances in the field of deep learning, and the most influential implications was the ability to train the neural network on raw image input, combined with the use of *experience replay*. The power of using raw data is the term *raw*, i.e. no handcrafted features, which allows the neural network to truly uncover potentially non-linear structures of the data. The use of experience replay allows the agent to exploit the data to a greater extent, implying faster training, without the risk of overfitting to the data, by continuously updating its experience (Lin, 1993).

The time following these two papers, for the field of reinforcement learning, has been high-paced and truly revolutionary – which is often the characteristics of the start of a new era. Many great and important papers emerged in relatively short time, with some of the more interesting contributions being (Mnih et al., 2015) and (Silver et al., 2016).  
(Mnih et al., 2015) was especially interesting because of the generalisable ability of their implementation, which obtained above-professional-human level on 49 Atari games, and smash the performance of all previous work evaluated on the ALE. Using their own words;

*This work bridges the divide between high-dimensional sensory inputs and actions, resulting in the first artificial agent that is capable of learning to excel at a diverse array of challenging tasks.* (Mnih et al., 2015)*.*

(Silver et al., 2016) was truly eye-opening because of their ability to obtain near-master[[3]](#footnote-3) level on the boardgame GO, which was regarded as one of the grand challenges for artificial intelligence, because of its enormous state space consisting of possible moves. The architecture employed consisted of two different neural networks, one to learn the policy function and one to learn the value function, combined with a tree-search implementation to locate the most visited move doing simulation.  
The before mentioned papers is perhaps some of the most influential papers in recent time, yet many great contributions has been seen over the last six years, developing novel and sophisticated methods. For now, these novel and sophisticated methods are left uncomment, as at least one will be explained later.

The recent advances in both deep learning and reinforcement learning has given rise to some major advances within the subfield of autonomous navigation, through the papers (Zuo et al., 2014), (Zhang et al., 2017) and not the least (Faust, A. and Francis A., 2019).  
(Zuo et al., 2014) embed Q-learning in a small robot, to safely and smoothly navigate it out of a maze, using input from internal sensors giving information about the robots relative position.   
(Zhang et al., 2017) goes a step further, with their objective of outlining a framework which is generalisable across many tasks. Their implementation builds on the assumption that the reward structure of a state, and thereby the future states following, can be expressed as a linear combination of *successor* features. More specifically, the idea is that different, yet somewhat similar, tasks can be expressed by the same features with different weighting. The implication is that transferring between tasks ends up being a question about optimizing for new weights, compared to initialising weights, for the base model of the new task, which substantially reduces training time.

Perhaps the most interesting reference at this point, within the subfield of autonomous navigation, is a collection of three novel papers. (Faust, A. and Francis A., 2019) presents the work done in three recent papers combining automated reinforcement learning and sampling-based planning.  
Automated reinforcement learning implies searching the reward space and the parameter space of a neural network, to obtain an optimal combination leading to the desired behaviour. In relation to sampling-based planning, (Faust, A. and Francis A., 2019) utilises *probabilistic roadmaps[[4]](#footnote-4)* before advancing to *simultaneous localisation and mapping[[5]](#footnote-5)*, to support the planning procedure by locating feasible roadmaps for the robot to follow.

The final contribution needed to be highlighted, considering the previously mentioned recent advancements, is based on the paper by (Juliani et al., 2018). (Juliani et al., 2018) introduce a toolkit for Unity – *Unity ML-agents Toolkit*, with the aim of taking the recent advances within artificial intelligence and reinforcement learning even further. They note themselves;

*As the state of the field becomes more developed, existing environments and the benchmarks based on them become less informative, the need for novel environments presents itself.* (Juliani et al., 2018).

Ground-breaking discoveries are based on the formulation of new grand challenges, and the introduction of Unity ML-agents Toolkit brings endless possibilities, in terms of formulating tasks and environments with ever increasing complexity. By the authors;

*When examining the areas of human intelligence typically studied, we find four major ones which have been the focus in recent years, and which we expect to continue to be the focus into the future: sensory, physical, cognitive, and social. … To challenge algorithms within these four domains, environments should be able to test all four axes of intelligence simultaneously.* (Juliani et al., 2018).

It is worth to think about where the work of this dissertation is placed, looking back at the topics and advances covered. The aim of this dissertation is to explore the opportunities within the Unity ML-agents Toolkit, by using embedded state-of-the-art deep learning methods to solve autonomous navigational tasks with increasing complexity. Furthermore, aims the dissertation at shredding light on the possibilities to use external sensor information in the navigational procedure, which distance the work herein from the highlighted references on autonomous navigation.

## References

1. **Anderson, C. W. (1986). Learning and Problem Solving with Multilayer Connectionist Systems. Ph.D. thesis, University of Massachusetts, Amherst.**
2. **Sutton, R. S. and Barto A. G. (2018). Reinforcement Learning – An introduction. Second Edition, MIT Press, Cambridge, MA.**
3. **Sutton, R. S. and Barto A. G. (1998). Reinforcement Learning – An introduction. First Edition, MIT Press, Cambridge, MA.**
4. **Tesauro, G. (1995). Temporal difference learning and TD-Gammon. Communications of the ACM, 38 (3): 58–68.**
5. **Sutton, R. S. (1984). Temporal Credit Assignment in Reinforcement Learning. Ph.D. thesis, University of Massachusetts, Amherst.**
6. **Sutton, R. S. (1988). Learning to predict by the method of temporal differences. Machine Learning, 3(1):9–44 (important erratum p. 377).**
7. **Witten, I. H. (1976a). Learning to Control. University of Essex PhD thesis.**
8. **Samuel, A. (1959). Some studies in machine learning using the game of checkers. IBM J. of Research and Development 3, 210-229.**
9. **Shannon, C. E. (1950). Programming a computer for playing chess. Philosophical Mag. 41, 265-275.**
10. **Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M. (2013). Playing atari with deep reinforcement learning. ArXiv:1312.5602.**
11. **Bellemare, M. G., Naddaf, Y., Veness, J., Bowling, M. (2013). The arcade learning environment: An evaluation platform for general agents. Journal of Artificial Intelligence Research, 47:253–279.**
12. **Long-Ji Lin. (1993). Reinforcement learning for robots using neural networks. Technical report, DTIC Document.**
13. **Mnih V, Kavukcuoglu K, Silver D, et al. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540): 529-533.**
14. **Silver, David, et al. (2016). Mastering the game of Go with deep neural networks and tree search. Nature 529.7587: 484-489.**
15. **Zuo, B. et al. (2014). A reinforcement learning based robotic navigation system.** [**2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)**](https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6960119)**.**
16. **Zhang, J. et al. (2017). Deep Reinforcement Learning with Successor Features for Navigation across Similar Environments.**
17. **Faust, A. and Francis A. (2019). Long-Range Robotic Navigation via Automated Reinforcement Learning. Google AI Blog. Online:** <https://ai.googleblog.com/2019/02/long-range-robotic-navigation-via.html> **Accessed: 01/06/2019.**
18. **Faust, A. et al. (2018). PRM-RL: Long-range Robotic Navigation Tasks by Combining Reinforcement Learning and Sampling-based Planning. *IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, Australia (2018), pp. 5113-5120.**
19. **Francis, A. et al. (2019). Long-Range Indoor Navigation with PRM-RL.**
20. **Juliani, A. et al. (2018). Unity: A General Platform for Intelligent Agents.**

1. Page 15, (Sutton and Barto, 2018) [↑](#footnote-ref-1)
2. Page 9. [↑](#footnote-ref-2)
3. Page 5, (Silver et al., 2016). [↑](#footnote-ref-3)
4. See (Faust et al. 2018) for details. [↑](#footnote-ref-4)
5. See (Francis et al. 2019) for details. [↑](#footnote-ref-5)