University POLITEHNICA of Bucharest

Automatic Control and Computers Faculty,
Computer Science and Engineering Department

SUMMARY

PHD THESIS

Exploration in Reinforcement Learning for solving real world problems. Exploration is all we need.

Thesis advisors:
Prof. Adina Magda Florea

Author:
Andrei Cristian Nica

Bucharest, 2023

Contents

Summary

This documente provides a comprehensive summary of a Ph.D. thesis focused on exploration in Reinforcement Learning (RL) for solving real-world problems. The thesis delves into the fundamental aspect of exploration, which involves searching and discovering new knowledge and skills. The goal of the research work has been to address the challenges faced by RL agents in exploration, ultimately enhancing their learning capabilities. Throughout the thesis, various perspectives on exploration and RL have been considered, ranging from assessing learning progress in tasks with sparse rewards to leveraging hierarchical learning and exploring environments with multiple agents. Additionally, the thesis extends its exploration into real-world applications, such as indoor robot navigation and autonomous driving, as well as generative machine learning for de novo drug discovery. By examining these diverse domains, the thesis offers insights and recommendations to practitioners and researchers working in the field of RL.

Reinforcement Learning is a powerful approach to building intelligent agents and has the potential to be used in a wide variety of real-world applications. However, despite its potential and recent progress, there are still many challenges that must be addressed before RL can be widely adopted. Exploration is a fundamental aspect of learning. Exploration is the act of searching and discovering, and it is essential for the development of new knowledge and skills. Therefore it is an integral part of Reinforcement Learning as it allows the agent to explore different actions and states to improve its policy and adapt to changing environments. The goal of this thesis has been to research various methods and strategies aimed at addressing exploration challenges faced by Reinforcement Learning agents, thereby improving their learning capabilities. The research work includes studying the trade-offs between exploration and exploitation, developing new exploration techniques, analyzing the impact of exploration on performance and learning, evaluating and improving the effectiveness of different exploration methods in various tasks and environments. The objective has been to identify and understand key factors that contribute to effective exploration in Reinforcement Learning, and to provide new insights and recommendations for practitioners and researchers working in this field. Throughout the thesis, various exploration and reinforcement learning perspectives have been considered. When an agent is given a task with sparse rewards, we have looked into ways to assess the agent's learning progress. While task-based rewards would not influence policy changes in demanding exploration tasks, we evaluated intrinsic rewards for enhancing agent learning. We explore leveraging and improving hierarchical reinforcement learning for long-horizon tasks. We consider solving tasks with multiple reinforcement learning agents, and we point out difficulties and solutions for enhancing learning in these configurations. We also examine how a multi-agent perspective could enhance single-agent exploration challenges. Finally, we argue for the significance of policy generalization in reinforcement learning and its relation to the exploration problem. In addition to researching reinforcement learning in simulated domains, motivated by the careful design of the hypotheses investigated, we also consider two main real-world problems that elevate the exploration issue to a new plane. On the one hand, we consider learning policies for indoor robot navigation and autonomous vehicles. On the other hand, we research generative machine learning for de-novo drug discovery.

1 Introduction

Artificial Intelligence (AI) has started to take shape in the 1950s, when the naming became popular among cognitive scientists, experts, researchers, and mathematicians. In the 1956, John McCarthy coined the term 'artificial intelligence' and established the first AI conference. Even before, the idea of building "intelligence" has been at least an idea to contemplate upon for many decades, but since then, it has become to some degree even a goal to pursue scientifically. A promising and intriguing idea, that we could ourselves build intelligent machines, which could have the potential of solving problems beyond what humans could solve, both in scale and complexity.

Machine learning (ML) has made significant progress in the last few decades, and in recent years, the field has seen the development of many new techniques and technologies that have greatly expanded its capabilities. For example consider the following: Generative adversarial networks (GANs) [14], Transformers [56], Graph convolutional networks (GCNs) [25], Adversarial training [15], Federate learning [31], Self-supervised learning [54], Neural architecture search (NAS) [65], Automated machine learning [53] (AutoML), Population Based training [20], Diffusion Models [47]. One of the most consequential evolutions in the field, which has underpinned many of the aforementioned advancements, is the emergence of "Deep Learning" [27]. It is a type of machine learning that uses large neural networks to learn from data. This has led to the creation of many powerful machine learning models that are able to achieve state-of-the-art performance on a wide range of tasks, such as image, language, and speech understanding, tackling problems that we wouldn't have been able to solve before. Machine learning has advanced significantly in the last decade due to a number of factors, including the availability of large amounts of data, the development of more powerful computing hardware, and the growth of interdisciplinary research and collaboration between different fields. One of the key factors that has contributed to the advancement of machine learning in recent years is the availability of large amounts of data. Machine learning algorithms require data to learn from, and the availability of large and diverse datasets has allowed researchers to train more accurate and sophisticated models. This has been facilitated by the widespread adoption of digital technologies, which has led to the generation of vast amounts of data in a wide range of fields, including health care, finance, transportation, and social media. The other very important factor that has contributed to the advancement of machine learning is the development of more powerful computing hardware. Machine learning algorithms require a lot of computational power, and the development of faster and more efficient computer processors, as well as the widespread use of graphics processing units (GPUs) and other specialized hardware, has made it possible to train larger and more complex models in a shorter amount of time. Finally, the advancement of machine learning in recent years has also been facilitated by the growth of interdisciplinary research and collaboration between different fields. Machine learning has applications in a wide range of fields, including computer science, statistics, biology, physics, and engineering, and the growth of interdisciplinary research has allowed researchers from different fields to share ideas and collaborate on common problems. This has led to the development of new algorithms and techniques that have been applied to a wide range of real-world problems.

Reinforcement Learning is a type of Machine Learning technique that enables an agent to learn in an interactive environment by trial and error using feedback from the environment (reward), its own actions and experiences. While Machine Learning is a sub-field of artificial intelligence, which considers in general algorithms that allow machines to continuously learn and adapt from data "unassisted", Reinforcement Learning establishes a general paradigm where machines, or agents, learn to act in an environment so as to maximize a reward signal. The origins of RL can be traced back to two independently studied topics: one area was investigating solutions for optimal control problems by using value functions and dynamic programming, while the second major area can be identified in the psychological studies of animal learning where models of trial and error learning have been proposed. A better detailed history of reinforcement learning is described by Sutton and Barto (1998) ([48]).

In this thesis we will focus on the aspect of exploration in solving problems with Deep reinforcement learning (DRL). Deep reinforcement learning combines the techniques of reinforcement learning and deep learning, while using neural networks as computational models for learning the agents behaviour. Exploration, in the context of reinforcement learning, refers to the process by which an agent learns about its environment and gathers information that can be used to make better decisions. This can involve taking actions that may not immediately lead to the highest rewards, but that can provide the agent with valuable information about its environment. For example, an agent learning to navigate a maze may take actions that take it down unfamiliar paths, in order to gather more information about the layout of the maze and improve its performance. Exploration is an important aspect of reinforcement learning, as it allows agents to learn and adapt to their environments in a more effective way. It also helps to address the trade-off between exploitation (using the knowledge the agent has already gained to maximize its rewards) and exploration (gathering new information to improve its performance). We will analyse some of the challenges of exploration in reinforcement learning for solving different problems, we will research different perspectives in solving these challenges, propose new ways to improve learning and promising future directions of research.

2 Motivation

Reinforcement learning has already achieved impressive results in many applications, including game-playing (Go, StarCraft, Dota) [45, 57, 58, 5], robotics (grasping objects) , natural language process (dialogue refined using human feedback) [36] and even several real world problems [11, 63]. This success further bolsters the idea that reinforcement learning can be a strong and effective method for addressing a variety of real-world problems. While machine learning algorithms have been applied to a wide range of tasks and industries, including finance, health care, transportation, and social media, reinforcement learning, on the other hand, has the potential to be applied to even more complex and dynamic real-world problems, such as self-driving cars, and drug discovery. Both supervised learning and reinforcement learning can be effective for solving specific tasks or problems, but they do not necessarily involve intelligence in the same way that a human or animal might. Intelligence is a complex and multifaceted concept, and there is no consensus on a precise definition. However, many researchers would agree that true intelligence involves the ability to think and reason, to learn from experience, to adapt to new situations, and to understand and use language. Reinforcement learning as a paradigm encompass many of these ideas, and has the potential to enable the development of intelligent systems that can learn and adapt to their environments in a more human-like way. Learning with reward is a general-purpose framework for learning, which could have many important applications, and might be even enough ([46]) to learn intelligence. Additionally, the flexibility of reinforcement learning algorithms allows them to be applied in both simulation-based and real-world environments, further increasing their applicability.

Reinforcement learning is an active and rapidly-growing area of research, with many exciting developments and advances being made. This makes it an exciting and rewarding field for researchers to work in, not only for its future impact prospects in science and society but also for its learning opportunity. It is a diverse field of research, being able to draw inspiration from a wide range of other domains. Algorithms are often inspired by developments and insights from a variety of fields, including psychology, neuroscience, control theory, and computer science. For example, the concept of reinforcement learning draws heavily from the field of psychology, where it was first studied in the context of animal learning. Additionally, the development of modern reinforcement learning algorithms has been heavily influenced by the field of control theory, which provides a mathematical framework for analyzing the behavior of systems that receive external input. Research in this field can also help us better understand how intelligent systems learn and make decisions. This understanding can provide valuable insights into the nature of intelligence and how it can be developed in machines.

The exploration versus exploitation dilemma is a fundamental challenge in reinforcement learning. This is because in many situations, an agent must choose between exploring its environment to gather new information, or exploiting its existing knowledge to maximize its rewards. The dilemma arises because exploration can be risky and may not lead to immediate rewards, while exploitation allows the agent to use its existing knowledge to maximize its rewards in the short term. However, if the agent only exploits its existing knowledge, it may miss out on opportunities to learn new strategies that could lead to even higher rewards in the long term. As a result, finding the right balance between exploration and exploitation is a crucial challenge in reinforcement learning. Exploration is essential for reinforcement learning agents to be able to learn and adapt to their environments. Exploring can lead to better solutions to the problem at hand, as well as more robust and reliable behavior. By studying exploration, researchers can develop strategies and algorithms that can help agents to strike the right balance between these two goals.

Although Reinforcement learning research has achieved incredible results in recent years, we believe it has yet to proven its potential, facing serious challenges for scaling to bigger problems [12]. Some of the key challenges include sample efficiency, generalization, exploration and safe exploration. These challenges are strongly inferring the need of research in exploration. It is also acknowledged that many of the results in the field have so far been achieved due to the scale of computation, scale of data and use of neural networks as feature extractors. Given the difficulty of learning solutions with RL, raises thus the question if scale is not enough or at least not desirable, which motivates further more fundamental developments and changes in the field.

We believe that this form of guidance of decision-making processes in real-time, which is the central aim of Reinforcement Learning, will form the foundation of leading algorithms for solving real-world problems in the future. These algorithms will likely be built upon a combination of unsupervised and supervised learning, and the use of rewards guiding the learning process will be a fundamental and scalable approach for achieving goals in the future.

3 Objectives

The primary objective of this thesis has been to investigate the problem of exploration in reinforcement learning algorithms and to explore new approaches for addressing this challenge. We have considered a holistic approach to analyzing this research question. There are several reasons why it might be important to having such an approach to researching exploration in reinforcement learning.

First, exploration is a complex and multifaceted problem that involves a range of different factors and considerations. By examining the problem of exploration from multiple angles and perspectives, it may be possible to gain a deeper and more comprehensive understanding of the challenges and opportunities involved in this task.

Second, a holistic approach can help to identify connections and relationships between different aspects of the problem, and to identify new approaches and strategies for addressing it. This can be particularly important in situations where traditional approaches may not be effective, and where novel solutions are needed.

Finally, a holistic approach can help to ensure that research on exploration in reinforcement learning is relevant and applicable to a wide range of real-world problems and applications. By considering the many different facets of exploration, it may be possible to identify approaches that are generalizable and that can be applied in a variety of different contexts.

Overall, a holistic approach to researching exploration in reinforcement learning can help to provide a deeper and more comprehensive understanding of this complex problem, and to identify new and innovative approaches for addressing it.

Specifically, this thesis aims to:

1. Examine the role of exploration in reinforcement learning and the challenges it presents in the face of unknowns and uncertainty.
2. Investigate the current state-of-the-art in exploration in reinforcement learning, and identify areas where further research is needed.
3. Examine the problem of exploration from a range of different perspectives, including intrinsic reward, hierarchical learning, multi-agent systems, and real-world applications.
4. Identify opportunities for improving exploration in reinforcement learning, and propose new approaches for achieving this goal.

Overall, this thesis aims to contribute to the broader understanding of exploration in reinforcement learning, and to provide new insights and approaches for addressing this important challenge.

4 Thesis structure

Chapter **2** begins with a short introduction in Reinforcement Learning, presenting context and background information needed to understand the problem and the potential approaches to improving these types of machine learning algorithms. This allows the reader to gain a better understanding of the problem and the rationale behind the research being conducted. Describing the foundation knowledge and theories helps us establish expertise in the field, identify gaps in the existing literature and provide direction for future research.

Further on in the thesis we research the problem of exploration in Reinforcement Learning from different perspectives, in different setups and for solving different real world problems. In each chapter we will provide a summary of prerequisite knowledge for that topic, analyse possible challenges, identify different perspectives for approaching solutions, propose new methods and analyse experimental results for the ideas presented.

Current Reinforcement Learning algorithms can easily be ineffective in sparse reward environments. In such setups the reward of the environment is a very weak signal for agents to learn useful behaviours for solving the task. This is one of the main problems which raises the challenge of exploration. In Chapter **3** we will start by analysing this setup, and one of the most common and intuitive approaches for enriching the learning signal, by using intrinsic reward. We will investigate how to evaluate learning progress in exploration, compare different methods and propose a learning signal based on the knowledge of ordering states by time.

In Chapter **4** we will investigate how learning temporal abstractions and leveraging them further to solve downstream tasks can help with exploration. We present advantages of an attention mechanism over temporally extended abstract actions and propose an online model-free algorithm for learning affordances and using them further to learn sub-goal options.

We further examine in Chapter **5** how Reinforcement Learning agents can learn to solve tasks in multi agent setups. We first investigate specific approaches and improvements useful for Multi Agent Reinforcement Learning, by improving performance in a game representing the Stag Hunt Dilemma. We then investigate and propose a different way of posing single agent problems as multi agent setups. This, in order to exploit the advantages of multiple agents, that can collaborate and communicate, and improve reinforcement learning with better exploration.

In Chapter **6** and **7** our research focuses on solving real-world problems and the new challenges that come with exploration.

In Chapter **6** we consider the problem of navigation in reality for both robots and autonomous vehicles. We start by analyzing the challenges of end-to-end learning policies for indoor robot navigation. Next, we investigate how we can leverage simulators and transfer policies to reality and show how memory and semantic information can help indoor localization. We continue by tackling the process necessary for building and testing a real-world autonomous vehicle that uses machine learning-based models for steering. We first describe the process of collecting and processing a local self-driving dataset. We then present results on this dataset for learning end-to-end models for steering and propose a self-supervised pipeline for learning driving policy only from monocular camera recordings.

In Chapter **7** we study the problem of exploration in the context of drug discovery, specifically how machine learning and reinforcement learning algorithms can identify potential new drug compounds by searching more efficiently the desired molecule state space. We research and propose improvements for both RL and GFlowNet algorithms for the task of small molecule generation and delve into the importance of generalization for better molecular design. Finally, we research the importance of properly evaluating the generalization of machine learning-based molecule design using simulator metrics in order to effectively transfer it to real-world binding affinity.

Following our research on the challenges of exploration in RL, both from different perspectives and in different problem setups, we propose in Chapter **8** a set of arguable opinions on what we believe will be essential for scaling Reinforcement Learning to complex real-world problems. First, we discuss the limitations of traditional reinforcement learning approaches and why they may not be well-suited for solving infinite state space problems, for which we propose a more general framework for RL. Second, we emphasize the strong connection between generalization and exploration in RL, which is essential for future progress.

Chapter **9** concludes the work, presents the main contributions of the thesis and future directions of research.

5 Chapter Summaries

5.2 Foundations

The Foundations chapter provides an introduction to Reinforcement Learning (RL), highlighting its significance as a research topic. It presents the origins of RL and discusses the fundamental reasons for its importance. The chapter also introduces the basic elements and mathematical frameworks essential for understanding RL. From a computer science perspective, it outlines the major challenges encountered when training an agent in an RL setup. The chapter explores various approaches, algorithms, and classification criteria that are crucial for solving RL problems.

Reinforcement Learning serves as the focus of Chapter 2, which begins with an examination of its motivation. The chapter delves into the concept of reward-driven agent behavior and presents the problem setup and learning framework. It explains Markov Decision Processes (MDPs) and provides the necessary background and notation for understanding RL. The theoretical challenges in RL are explored, including the curse of dimensionality, the exploration versus exploitation dilemma, and issues related to partial observability, stochasticity, discreteness, credit assignment, non-stationary environments, and generalization. Additionally, the chapter covers the integration of deep learning into RL, discussing neural networks, deep reinforcement learning, and the associated challenges and future directions. The significance of exploration in RL is also emphasized.

By addressing these foundational aspects, the Foundations chapter lays the groundwork for a deeper understanding of RL and sets the stage for the subsequent chapters' exploration of specific topics within the field.

5.3 Exploring with Intrinsic Reward

In the following chapter we will define and investigate theories for improving exploration in reinforcement learning algorithms using intrinsic reward. We will review state of the art algorithms for generating such reward, analyze current limitations for exploration in reinforcement learning on a set of computationally efficient environments, analyse options for evaluating learning progress in such setups and investigate solutions for improvement. Finally we will introduce a novel intrinsic reward mechanism that leverages, self-supervised, the knowledge of ordering states based on time in order to address exploration issues in RL. We show how learning the order of time from states, and using the "orderability" score of randomly shuffled states as intrinsic reward determines agents to converge towards less chaotic policies.

From a RL point of view we can define exploration as the policy which helps the agent interact with new states or new transitions in the environment. This would hep the agent explore the problem space enough in order to attempt to solve tasks and not get stuck on course. For this problem we will not be concerned with the dilemma of exploration versus exploitation but we will try to define a measure for the agent's novelty of experience and develop a RL algorithm for maximizing this. This does not however dismiss exploiting early actions in order to unlock novelty on a higher temporal resolution.

The problem that we will be focusing on in this chapter is that of better exploration. A classical reinforcement learning algorithm will try to maximize the extrinsic reward and thus will have no signal for learning until it receives a change in reward. There are many environments and situations where rewards are very sparse and the agent would be faced with random exploration in hope of reaching some reward. It is easy to see why we would want an intelligent agent to be able to learn much more from these experiences even without any extrinsic reward change. We could for example improve in these kind of situations by handcrafting heuristics for exploration or designing extra reward functions, also known as reward shaping, in order to make a problem easier to learn. But not only are we researching a universal reinforcement learning algorithm which is able to learn with as less as possible prior knowledge, but also designing the reward function and hand crafting heuristics can even hinder learning the best solution [1, 19].

We can however seek to transfer theories from information theory, biology and other domains and define other objectives for our agents. We could measure learning progress [43], reduce cognitive dissonance (incompatibility between two or more cognitive structures, experiences or behaviours [23], maximize empowerment [40] (the amount of control an agent has over it's environment), setup the agent for predictive coding [35]. Schmidhuber et al. [42] describe for example that agents should encompass multiple ingredients: they should continually compress or predict experiences, measure progress and optimize for bigger reward by improving their action policy.

We have so far argued why RL is an important research topic, how exploration is an important challenge in this domain and offered grounds for why intrinsic motivation seems to be exhibited by humans as well. Given all this, there seems to be good basis for researching improvements for exploration using intrinsic reward (IR). So far, there are various ways that have been studied in the literature to equip agents with the drive to explore simulated environments [7]. For example, agents can manifest curiosity depending on how boring an environment is, avoiding fully predictable states [37]. Other agents can be viewed as information-seeking. There are studies which analyze the concept of flow where an agent tries to maintain a state where learning is challenging, but not overwhelming [10]. The free-energy principle states that an agent seeks to minimize uncertainty by updating its internal model of the environment and selecting uncertainty-reducing actions. Empowerment is founded on information-theoretic principles and quantifies how much control an agent has over its environment.

In this chapter, we have delved deeper into a significant challenge that current RL algorithms face with learning, namely exploration. We have defined the objectives of exploration methods in an RL setup and examined how they are currently addressed in research. To measure the evolution of such algorithms, we have converged on a benchmark suite of environments (7 MiniGrid environments) with sparse rewards, which can be used to evaluate improvements with low computational resources. We have focused on the concept of intrinsic reward and explored how theories from information theory, hierarchical learning, predictive coding, and goal-conditioned policies can be employed to generate such signals.

In the second part of the chapter, we provide a baseline guide for assessing the performance of current IR algorithms in enhancing exploration in these environments. We also propose solutions for evaluating progress without measuring extrinsic reward. We outline some of the challenges still faced by IR methods in the literature and identify some of the flaws they exhibit in the selected environments. We have examined two pre-existing methods of IR for improving exploration in delayed or unknowable outcomes and presented our results.

In the final section, we propose a novel self-supervised signal for learning less chaotic policies. We first demonstrate how the "orderability" score of agents' experiences correlates with higher rewarding policies and improved exploration. We then present experimental results on how this signal can be used as an intrinsic reward for enhancing RL agents in sparse reward tasks. We provide compelling evidence that the orderability score is not only a useful metric for analyzing the learning progress of the agent but also that it can be employed as an IR, particularly in challenging exploration environments. It is important to note the discussion regarding the definition of learning progress we consider, as described in Sec. **3.4.2.**. While intrinsic reward remains an open and unsolved problem, there are several aspects of this method that show promising potential. Although we have not provided definitive experimental proofs of these advantages, we would like to outline them as follows. We believe this method has strong generalization properties compared to next-state prediction or other similar information gain IR methods. We argue this is due to the view invariance properties of the architecture and prediction target state space [21]. We propose this hypothesis of improved generalization because predicting next states is a considerably more challenging problem than ordering input observations. For instance, next-state prediction models can struggle with never-before-seen observations or parts of the observations, while this method may still be able to distinguish between the states provided in the input and order them correctly. We also believe this method will not be affected by the TV noise problem, as those sequences will have a poor orderability score.

5.4 Leveraging Hierarchical Learning

AI decision-making agents often grapple with two significant hurdles: the extent of the planning horizon and the branching factor arising from numerous alternatives. Hierarchical reinforcement learning techniques aim to address the former by introducing shortcuts that leapfrog multiple time steps. To manage the breadth, it's ideal to limit the agent's focus at each stage to a manageable number of possible options. The affordances theory (Gibson, 1977) proposes that only specific actions are viable in certain states. In this chapter introduce an online, model-free algorithm for learning affordances, which can then be used to learn subgoal options. We simulate "affordances" via an attention mechanism that curtails the range of available temporally extended options. We delve into the role of hard versus soft attention in gathering training data, the learning of abstract values in tasks with long horizons, and the management of an expanding array of options. We discern and empirically demonstrate situations in which the paradox of choice surfaces - that is, when having a smaller number of more meaningful options enhances the learning rate and efficacy of a reinforcement learning agent.

Decision making in complex environments can be challenging due to having many choices to consider at every time step. Learning an attention mechanism that limits these choices could potentially result in much better performance. Humans have a remarkable ability to selectively pay attention to certain parts of the visual input [22, 6], gathering relevant information, and sequentially combining their observations to build representations across different timescales [16, 64], which plays an important role in guiding further perception and action [34, 2]. In this work, we explore ideas for endowing reinforcement learning (RL) agents with these type of capabilities.

A RL agent interacts with an environment through a sequence of actions, learning to maximize its long-term expected return [49]. Temporal abstraction, e.g. options [50] allow it to consider decisions at variable time scales. Options allow the agent to reduce the depth of its lookahead when performing planning, and to propagate credit over a longer period of time. However, in large state-action spaces, the agent is still faced with many choices at every decision step, and options can in fact worsen this problem, as choices multiply when considering different timescales. Hence, temporal abstraction can lead to a larger branching factor.

A well-known solution is to *select* a small number of choices to focus on, for instance, applying an action only when certain preconditions are met in classical planning [13], or using initiation sets for options [50]. Recent work [24] proposed a generalization of initiation sets to *interest functions* [51, 59], which provide a way to learn options that specialize to certain regions of the state space. All these approaches can be viewed as providing an attention mechanism over action choices.

While attention has been widely studied in the field of computer vision [18, 6], the role of different types of attention over action choices has not been studied much, especially for temporally abstract actions. In this work, we explore attention over action choices in RL agents with a special interest in affordances associated with options, which can be viewed as a form of *hard* attention. On the other hand, having soft preferences over *all* actions without eliminating any can be viewed as *soft* attention.

In this work, we study the role of *hard* versus *soft* attention in decision making with temporal abstraction. We posit that in certain settings, restricting an agent's attention through affordances leads to fewer but more useful choices compared to using soft attention in the form of interest functions. We demonstrate empirically this *the paradox of choice:* fewer choices can contribute to faster learning, resulting in more rewards. As expected, this effect is more pronounced when the agent attends to choices that lead to better long-term utility. Attending to useful choices in a given state often has a compound effect, leading to further good choices in the near future.

Key Contributions: We measure the impact of *hard* versus *soft* attention over option choices: 1) when generating training data, 2) on abstract value learning in long-horizon tasks, and 3) when increasing the number of choices. We present an online, model-free algorithm to learn affordance-aware subgoal options and the policy over options with given subgoals and a high-level task. The affordances and option policies are learned simultaneously, informing each other. Leveraging the learned affordances to generate training data facilitates sample-efficient learning of the options. We empirically demonstrate that fewer choices can be better for both learning speed and final performance for an hierarchical reinforcement learning agent.

The focus of our work is to study affordances as hard attention and contrast it with soft-attention for temporally extended actions. We demonstrate this through non-tabular environments, consisting of both discrete and continuous action/state space in Minigrid [8] and Open-AI Robotics [44] respectively. The choice of these experimental setups were motivated by careful design with the aim to investigate and isolate the role of affordances as hard attention, while being able to vary the complexity in terms of number of options and tasks. We also emphasize that our approach scales to both discrete and continuous action spaces and it can be used in any domain where subgoals are known apriori.

5.5 Exploring Environments with Multiple Agents

In this chapter we examine how Reinforcement Learning agents can learn to solve tasks in multi agent setups. First we present a deep reinforcement learning approach for learning to play a time extended social dilemma game in a simulated environment. Agents face different types of adversaries with different levels of commitment to a collaborative strategy. Our method builds on recent advances in policy gradient training using deep neural networks. We investigate multiple stochastic gradient algorithms such as Reinforce or Actor Critic with auxiliary tasks for faster convergence. Second, we propose a novel way of posing single-agent problems as multi-agent setups and present methods for training RL agents in such a paradigm. We motivate this setup by being able to exploit the advantages of multiple agents, which can collaborate and communicate and benefit reinforcement learning with exploration challenges. We provide preliminary empirical evidence to support this hypothesis.

Machine learning research in years 2016 and 2017 yielded continuous progress in deep reinforcement learning algorithms for agents placed in simulated scenarios. A plethora of novel model-free algorithms explored the advantages of using deep neural networks for predicting state, or action values, and for approximating policies for continuous control in various visual environments (e.g. Atari [32], Vizdoom [26], or Minecraft [52], or board games (Go). Few recent studies focused on reinforcement learning in multi-agent setups where several learning entities must cooperate in competing, or collaborative games. The problem of non-stationarity, one of the core challenges in reinforcement learning, is aggravated by the continuously changing behaviors of the other actors. However, solving problems in a multi-agent reinforcement learning paradigm, could yield improved performance, due to the increased informational exchange and coordination of multiple actors.

The return scheme of a multi-agent reward-based scenario is sometimes best described from a game theory perspective. A payoff matrix summarizes the gains for all players as a function of their chosen strategies. There is a large corpus of research in Artificial Intelligence on how agents can maximize their expected return through learning from iterated interactions. All those results focused on learning in stateless setups following various payoff schemes (e.g. prisoner's dilemma). Also, theoretical properties such as Nash equilibrium or Pareto optimality are assessed for learned strategies. A more difficult problem is identifying such a situation in a more complex scenario (e.g. during a chess game or a collaborative prey hunting) where the reward is a consequence of a (possibly large) sequence of decisions based on partial raw observations of the environment.

In the first part of this chapter we presents a deep reinforcement learning approach to such a scenario, where agents are situated in a cooperative episodic game based on stag hunt. Players have no persistent memory from one episode to the next. One of the goals of our research was to see if agents understand the underlaying macro scheme (which is an instance of the stag hunt game) through deep reinforcement learning techniques and not by being explicitly taught to choose between two strategies. Abstracting a high description of the interactions with the environment would be beneficial in many ways. For example, good performance in this game might offer a valuable prior experience before learning a second similar task (transfer learning). Knowing how to deal with a trust dilemma in general requires an agent facing a new situation just to learn how to interpret perception, and how to affect that specific environment reducing the burden on the learning process. The approach taken in our work is based on on-line policy gradient methods, more precisely variations of Reinforce and Actor-Critic algorithms with auxiliary tasks. All predictors are deep convolutional networks followed by recurrent layers trained using gradient-based updates per episode.

The scenario we tried to solve, called Malmo Platform PigChase[1] is one of the scenarios in Malmo Platform a reinforcement learning environment built on top of Minecraft. The higher complexity of this medium leads to a high resource, slow environment, which at first glance is incompatible with training deep neural predictors that require millions of samples before achieving high returns. For this reason we implemented a simplified replica of the original setup approximating the dynamics of the original setup which we will call from now on the PigChase Replica environment. We had two objectives in mind for this simplified environment: to be many times faster than the original and to work in batch mode. We then trained policy gradient based agents in this fast setup only to fine tune them at the end on the original one. Our intuition was that once the agent understands the general dynamics and the underlying trust dilemma, moving him on a similar environment requires small adaptation.

In single-agent RL, exploration can be challenging as the agent must balance between exploring new actions and exploiting actions that have been learned to be rewarding. Multiple agents, in multi-agent reinforcement learning (MARL), can help help however to improve exploration by leveraging the diversity of behavior, coordination, competition, and incentive alignment between agents. By having multiple agents explore the environment, there is a greater likelihood that some agents will explore new states and actions, increasing the overall exploration of the environment. Agents can coordinate their exploration efforts, such as splitting the exploration space among them or sharing information about the states they have visited. The presence of multiple agents in a competitive setting can drive exploration, as agents strive to find an advantage over each other. MARL can be used to align incentives between agents, encouraging them to collaborate and explore together, leading to improved exploration. These ideas motivate the research work undergone in Section **5.4** where we investigate how to adjust popular deep reinforcement learning algorithms, in order to pose single agent problems as a multi agent setups. In this section we present preliminary results that show our method is able to exploit the advantages of multiple agents, collaboration and communicate, in order to overcome RL exploration challenges.

5.6 Exploring and Navigating at Scale in Reality

In this chapter, we consider the problem of navigation in reality for both robots and autonomous vehicles. We start by analyzing the challenges of end-to-end learning policies for indoor robot navigation. Next, we investigate how we can leverage simulators and transfer policies to reality and show how memory

[1] https://www.microsoft.com/en-us/research/project/project-malmo/

and semantic information can help indoor localization. We continue by tackling the process necessary for building and testing a real-world autonomous vehicle that uses machine learning-based models for steering. We first describe the process of collecting and processing a local self-driving dataset. We then present results on this dataset for learning end-to-end models for steering and propose a self-supervised pipeline for learning driving policy only from monocular camera recordings.

The rapid evolution of mobile robotics has propelled the exploration of vast and diverse environments, with a primary focus on autonomous navigation. In this work, we delve into the dual domains of indoor robot navigation and autonomous driving, exploring the complexities and challenges that each field presents.

The field of mobile robotics is undergoing rapid development, with autonomous machines capable of performing a diverse range of tasks. However, navigating through these environments - especially indoor spaces and densely populated outdoor areas such as streets and highways. Thanks to advancements in machine learning algorithms, the utilization of raw sensory data such as cameras has become feasible in mobile robots, allowing machines to have access to as much information as possible in order to overcome the nosiness of real world scenarios. Although the data sources can be used in in conjunction with other sensors like odometry or laser, relying on raw sensory data, robots can learn and generalize from a diverse range of experiences, making them better equipped to handle novel and unseen situations. The scientific community is currently focused on addressing various critical challenges, particularly in the area of task navigation, which often involves the difficulty of exploration. To overcome this challenge, it is imperative for mobile robots to possess the ability to build robust models of their environment, solve localization problems, and navigate effectively. For improved autonomy, robots would be equipped with suitable exploration policies, which can continuously navigate and learn in real-world settings.

In the modern era, technological advancements are persistently reshaping our world, relieving humans from laborious tasks and enhancing productivity through automation. This evolution has seen an exponential rise in the application of mobile robots and autonomous vehicles across diverse environments. These autonomous entities are found in households, industrial spaces, educational institutions, and even on roads, adeptly performing a plethora of tasks. Significant strides in perception and computational abilities have contributed to broadening the scope of environments where these automated systems can operate. Navigation, a crucial facet of both mobile robotics and autonomous driving, is vital to the functionality of these systems. The intelligence of these machines manifests in their navigation abilities, as they are utilized in numerous applications like transportation, industry, and rescue missions. Path planning, a fundamental component of autonomous navigation, has been the subject of intensive research over the past two decades. Numerous algorithms have been developed and tested on various robotic systems such as micro air vehicles [61, 41], motion robots, wall-climbing robots, and underwater robots [62]. Similarly, autonomous vehicles on our roads also rely heavily on advanced path planning to ensure safe and efficient travel. The traditional understanding of navigation in mobile robots and autonomous vehicles lies in addressing three core questions (Levitt et. al. [28]): Where is the robot or vehicle, where are other places and things in relation to it, and how can it reach those other places from its current location. However, in intelligent organisms, these rules may not necessarily be linearly organised. Instead, they may be deeply interwoven, reflecting the intricate and dynamic realities these autonomous systems must navigate. This exploration of autonomous navigation in both indoor and outdoor contexts propels our understanding towards more comprehensive and adaptable solutions.

Artificial Intelligence aims to imbue systems with intelligence that can facilitate optimal decision-making and task execution. This foundational concept proves crucial in diverse scenarios ranging from indoor robot navigation to autonomous driving. The contemporary AI landscape is dominated by Machine Learning algorithms that enable machines to learn independently. One branch of ML, Reinforcement Learning, outlines a learning framework in which an agent, whether it's an indoor navigating robot or an autonomous vehicle, performs actions in an environment and strives to maximize its reward based on its gathered observations. This learning paradigm, extensively investigated in our previous research concerning exploration in the Markov Decision Process (MDP) space, can be tailored with necessary constraints to address navigation challenges in both domains. From an RL perspective, exploration can be seen as the policy guiding the agent to interact with new states or transitions in the environment. This exploration enables the indoor robot or the autonomous vehicle to sufficiently understand its surroundings and effectively perform its tasks without getting stuck. Thus, RL methods not only empower robots to navigate intricate indoor spaces but also equip autonomous vehicles with the capability to traverse complex road networks safely and efficiently.

Getting the software right is especially challenging for autonomous systems such as autonomous vehicles navigating complex traffic networks or indoor robots moving through dynamic spaces like airports, malls, and warehouses. These areas often contain narrow paths, shifting obstacles, and evolving patterns that demand intricate route planning. The key lies in devising software that can handle these issues while keeping user experience at the forefront. Rule-based approaches can often fall short in scalability in such unpredictable scenarios. It's essential that the autonomous system, whether a vehicle or a robot, requires minimal training for operators, avoids excessive environmental setup, learns quickly from demonstration, and provides productivity reporting. Variables influencing autonomous navigation are not merely physical obstacles occupying the system's working environment. Featureless terrains, varying times of the day, and countless other factors add complexity to autonomous navigation, whether on the road or indoors. Many of these challenges are edge cases that push the limits of conventional strategies, often only surfacing once the software is fully developed and the autonomous systems are tested in real-life situations. These complexities underline the need for a learning approach and the use of near real-world simulators. Nevertheless, truly functional autonomous navigation systems are not entirely bred in a lab; their success depends on the capability of the autonomous system, and the software it runs on, to adeptly navigate real-world environments. This holds true for both autonomous vehicles braving the uncertainties of the roads and indoor robots tasked with efficiently navigating bustling commercial spaces.

In this chapter, we have investigated the challenges and potential solutions for large-scale exploration in reinforcement learning with a focus on both indoor robot navigation and outdoor autonomous driving. We began by examining the process of learning end-to-end robot navigation policies for indoor spaces using RL. This involved transferring knowledge from simulation to real-world environments, implementing semantic attention, and addressing localization and long-range planning challenges. In the second part of the chapter, we shifted our attention to the domain of self-driving cars, delving into the intricacies of collecting and processing a robust dataset, as well as developing end-to-end steering policy models for autonomous vehicles. As we conclude the chapter, we reflect on the insights gained from these studies and discuss potential avenues for future work in both indoor robot navigation and outdoor autonomous driving.

In Sections **6.2, 6.3, 6.4,** we address the problem of robot navigation in indoor environments and explore the potential of using artificial neural networks as a fully learning-based approach to surpass the limitations of classical methods. We review these classical approaches and, based on the literature, identify areas where learning a navigation policy could potentially achieve state-of-the-art performance. We outline a set of navigation tasks that can be investigated individually, highlighting how incorporating semantic knowledge about the environment could enhance geometric approaches that have traditionally neglected this aspect. During this analysis, the challenge of space exploration emerges as a critical task in itself. The remainder of the report delves into this issue, focusing on training RL agents capable of navigating in real-world environments. Since training solely on a robot is currently infeasible and it would be imprudent to disregard the benefits and knowledge provided by simulators, we concentrate on three main challenges in indoor robot exploration: navigating using only RGBD information without precise localization; building a robust representation with implicit mapping and long-range localization by investigating attention mechanisms; and exploring methods for improving transfer to real-world environments. In this research we have independently evaluated each augmentation method, with depth noise clearly improving test scores. However, other methods demonstrate minor improvements, and we aim to explore the potential benefits of combining them. Although some methods may not significantly contribute to generalization in a supervised setup, we should not disregard their potential impact on policy learning, as suggested by previous work. Furthermore, we intend to investigate a more randomized approach to the "Noop" augmentation, bypassing the conditioning based on the distance to an obstacle. Thorough testing and refinement of the memory mechanism are essential, as it is a critical ability for an effective exploration agent. The current simple recurrent method appears to fail when calculating odometry for distances of five steps, indicating the importance of this component in the learning architecture. In Section **6.2** we will introduce the problem of robot navigation, what tasks can be defined within this frame and some of the current solutions. In Section **6.3** we present necessary resources for reproducing the experiments explained in this report. As for Section **6.4** we focus more on the problem of exploration, how to scale it to longer time horizons, and how to better transfer a robot policy to real world scenario.

Thus far, only preliminary tests have been conducted for training an exploration agent in the environment. Future steps include integrating auxiliary tasks and augmentation methods, evaluating the best policy on a real robot, and developing a metric for assessing the exploration ability of the policy in real-world scenarios. The explored surface scores may vary significantly depending on the clutter level in the environment, even with policies of similar quality, necessitating the development of an appropriate evaluation metric.

Our work then transitions to the domain of autonomous driving, where we present our research on creating and processing self-driving datasets, exploring end-to-end models for autonomous vehicles, and investigating self-supervised labeling techniques. We aim to demonstrate how machine learning and reinforcement learning techniques can be adapted to address these complex problems, and how these solutions can be scaled and transferred to real-world scenarios. In Section **6.5**, we elaborate on the process of collecting and processing a self-driving dataset at the UPB campus. We highlight the importance of dataset quality, particularly in safety-critical areas like autonomous driving, and provide a comprehensive guide covering every step from hardware setup to data validation. We believe that having a clean and efficient data collection process can enhance the benchmarking of autonomous driving solutions, accommodating local environmental nuances. In Section **6.6**, we explore the research on end-to-end models for self-driving cars on UPB campus roads. By transforming vehicle trajectory planning into a learnable process, we investigate the possibility of integrating steering predictions into existing autonomous driving frameworks. Our approach factors in specific geographical traits, employs data augmentation techniques, and assesses the model's quality based on a dataset collected from the UPB campus. Our findings suggest that the proposed models perform well for the gathered dataset. In Section **6.7**, we propose a training pipeline for Self-supervised Labeling for Autonomous Driving. This approach extracts valuable steering, acceleration, and braking information solely from monocular camera driving data. We connect this data to a reinforcement learning algorithm capable of predicting necessary commands for autonomous driving based on an abstract representation of the road. Key contributions of this section include a method for generating steering data annotations from unlabeled data and an innovative pipeline that produces path labels in an entirely unsupervised manner.

So far, we have presented and prototyped a pipeline for learning and producing steering angles using self-supervised, unlabeled driving scene recordings. This pipeline can distill valuable driving behavior from vast amounts of easily collectable data. We propose possible improvement ideas for this pipeline, which have only been tested to a limited degree. Firstly, the ground truth ground plane in our work has been roughly estimated using camera positions but could be generated using depth estimation and more advanced ground plane estimation to reduce errors produced from tilted or different camera positions in the recordings. The path labels can also be used to determine the car's speed by taking into account the length of the labeled path in terms of distance from the car to the furthest label in ground plane projection. To do this, we would need to determine the furthest "available" path the car can identify based on the current image. We consider an available path as the open road where the car can drive without any obstacles. We believe this can be determined self-supervised from videos using prediction errors both forward and backward in time. Using two neural networks to predict the future path both forward and backward through the video could be employed to determine obstacles that the car has avoided or slowed down for. This can be achieved by determining the patches which cannot be generated from both directions, helping us generate labeled paths only for the "drivable" available trajectory for a car at a given moment. Path predictions can be improved and evaluated using drivable area segmentation networks. Although the Ackermann steering model and camera parameters should be known for the deployment car, correctly calibrating them remains a significant challenge requiring constant maintenance. We proposed a method using reinforcement learning which can learn to drive using segmentation maps and self-adapt to different steering models of the car or to different camera positions or parameters. This model can be trained end-to-end using predicted paths on collected data, and we suggest that the observation space makes it much simpler for augmentation and simulation to cover the real data distribution during deployment.

5.7 Exploring Chemical Space for De Novo Molecular Design

In the following chapter we will study the problem of exploration in the context of drug discovery, specifically how machine learning and reinforcement learning algorithms can identify potential new drug compounds by searching more efficiently the desired molecule state space. We research and propose improvements for both RL and GFlowNet algorithms for the task of small molecule generation and delve into the importance of generalization for better molecular design. Finally, we research the importance of properly evaluating the generalization of machine learning-based molecule design using simulator metrics in order to effectively transfer it to real-world binding affinity.

The application of machine learning techniques to drug discovery and molecular design holds significant potential for revolutionizing the way we identify and develop new therapeutic agents [29]. Traditional drug discovery methods are often time-consuming, labor-intensive, and costly, with many potential compounds failing to reach the market due to issues in efficacy, safety, or pharmacokinetics. The average clinical drug development time reaches more than nine years and median development cost nearing 1 billion USD [60]. Preclinical early-stage drug discovery is typically an iterative optimization process that consists of the generate, assay, learn cycle. Moreover, the vastness and complexity of the chemical space make it difficult for conventional approaches to explore and identify novel, biologically active molecules efficiently. Machine learning offers the opportunity to overcome these challenges by leveraging the advantages of data-driven algorithms to accelerate the drug discovery process, optimize molecular properties, and uncover novel chemical structures with desired biological activities.

The drug discovery process [39] is a complex and multi-faceted endeavor that typically consists of several key stages, including target identification, target validation, hit identification, lead optimization, and preclinical development. In target identification, researchers focus on finding a biological target, such as a protein or a gene, that is implicated in a particular disease. Target validation involves confirming the target's role in the disease and its suitability for therapeutic intervention. Hit identification entails finding a compound with confirmed activity against a biological target (those that interact with the target). In this step, new drug candidates or molecules are designed or identified using various methods, such as computer-aided drug design, combinatorial chemistry, or high-throughput screening of existing compound libraries. The newly generated drug candidates are then tested in vitro (in test tubes or cell cultures) or in vivo (in animal models) to evaluate their biological activity, potency, selectivity, and other properties relevant to their potential therapeutic use. This step helps researchers understand how well the candidates interact with their target proteins or pathways and their potential side effects. Lead optimization refines these initial "hits" into more potent and selective "lead" compounds with improved drug-like properties. Lastly, preclinical development involves rigorous testing of the lead compounds in cellular and animal models to assess their safety, efficacy, and pharmacokinetic properties before progressing to clinical trials. Machine learning and reinforcement learning can be integrated into these steps to enhance their efficiency and effectiveness [9, 55].

By automating and enhancing various aspects of the drug discovery pipeline, these advanced computational techniques can help reduce the time and cost associated with bringing new drugs to market while also improving the likelihood of discovering effective and safe therapeutic agents. In the context of de novo molecular design, when designing novel molecules with desired properties from scratch, reinforcement learning (RL) and other machine learning (ML) based techniques have emerged as promising approaches for the efficient exploration and optimization of chemical space. These methods aim to address the limitations of traditional techniques by automating and expediting the discovery of novel molecular structures with desired properties. Generative models have brought about a revolution in the field of de novo drug discovery, introducing groundbreaking approaches to the process. Techniques such as long short-term memory recurrent neural networks (LSTM-RNNs), variational autoencoders (VAEs), generative adversarial networks (GANs), adversarial autoencoders (AAEs), evolutionary algorithms, gated recurrent unit (GRU-RNNs), and diffusion models have made significant contributions to this advancement [30, 33, 17]. For many years, reinforcement learning algorithms, such as Q-learning and policy gradient methods, have been proposed as a complementary approach to molecular design [38]. Emerging techniques in the field, such as graph neural networks and attention mechanisms continuously drive progress. These techniques enable iterative optimization of specific objectives through trial-and-error interactions within the chemical space, providing a valuable perspective for tackling complex molecular design challenges. The integration of RL with deep generative models can yield more targeted and efficient exploration, as the RL agent guides the generative model to produce molecules with optimal properties. Additionally, multi-objective reinforcement learning can address the challenge of optimizing multiple physicochemical properties simultaneously, leading to a more comprehensive design process.

In section **7.3** we have conducted extensive experiments on Reinforcement Learning algorithms to enhance small molecule generation, specifically focusing on improving both docking scores and diversity . An essential section of our work also involved the analysis of a proxy neural network model serving as a reward mechanism, a common approach in Drug Discovery with ML due to the costly processes of gathering training data through wet-lab experiments or expensive simulations **7.2**. As a result of researching improvements of GflowNet molecule generation [4] in section **7.4**, we introduced for example a new metric, the TopKDiverse, to better align with our core objectives of discovering high scoring and diverse molecules. In consideration of the unique challenges presented by this domain, we proposed a new benchmarking suite tailored to small molecule discovery, thereby providing a robust standard for future research and development in this area. Lastly we engaged in a deep analysis of generalization within the GFlowNet algorithms, further enhancing their application in small molecule generation **7.5**.

The systematic investigation into de novo molecular design utilizing reinforcement learning methodologies and GFlowNet generative algorithms has yielded noteworthy and substantial results. As the key findings suggest, when faced with an ill-defined reward function, RL algorithms could still maximize specific regions of our state space. However, the quality of the proxy reward function was directly proportional to the amount of training data used.

In the context of RL training for diversity, it became apparent that several factors were of critical importance, including reward discounting with count-based exploration, efficient horizon determination, and careful calibration of the clip parameter and entropy coefficient. It was also observed that the neural network's size could significantly impact performance, given sufficient training data.

GFlowNet, on the other hand, presented its unique set of challenges and lessons. Evaluation of the model's performance was a critical aspect, and aligning the goal with the true objective was fundamental. Dealing with multi-objective rewards posed a significant challenge, but employing the best discovered trajectory as experience replay during training was found to be beneficial.

In terms of generalization, our research has yielded robust techniques for evaluating GFlowNet performance on a test set. We found a strong correlation between a specific metric and the ability to generate diverse, high-scoring batches, proving to be a promising area for further study. In this work we presented a thorough evaluation of the generalization performance of GFlowNets for molecule design, using our propsed metrics *GFNEval* and *pHighestKbins*. A key limitation of the metrics presented in this work is the cost to compute the exact probability of sampling a molecule under the GFlowNet. This can make evaluation difficult with large test sets and larger training runs. One direction to explore would be ways to approximate this quantity more efficiently, while still retaining the properties of the proposed metrics. We also demonstrate the discriminative power of the metrics within a training run. This property can be helpful in active learning settings with a GFlowNet generator [3] for early stopping the generator training. Future work should also focus on using these metrics for further analysis of GFlowNet learning dynamics, to make practical recommendations for training.

However, as with any scientific endeavor, our research was not without its challenges. The transition to real scores and the optimization of graph neural networks posed significant hurdles, and multi-objective maximization proved difficult to tackle. For the future, we have uncovered potential avenues for further research. Protein generation, using the Rosetta environment, offers an exciting potential extension of our work. Rosetta's sophisticated suite of algorithms for computational modeling and analysis of protein structures could provide a rich ground for applying our learned lessons from de novo molecular design. Additionally, decision transformers presented a promising area for constructing policies for exploring protein space. Leveraging their ability to model entire trajectories could potentially enable more effective exploration and better alignment of RL policies with our objectives.

5.8 Generalization and Challenging the Status Quo

In this chapter, I undertake a comprehensive exploration of the challenges inherent in Reinforcement Learning (RL), a field of study that has seen significant advancements but still faces considerable hurdles. This exploration is conducted from a variety of perspectives and within different problem setups, providing a broad and nuanced understanding of the issues at hand. The chapter presents a series of thoughtfully argued opinions that I believe are critical for scaling RL to address complex real-world problems.

One of the primary discussions in this chapter centers around the limitations of traditional reinforcement learning approaches. These methods, while foundational to the field, have shown themselves to be potentially ill-suited for solving problems that involve infinite state spaces. These are complex problems that require a level of adaptability and scalability that traditional RL approaches may not be able to provide. Moreover, I argue that the current focus and evaluation setup in RL research could be hindering the progress of developing and tracking algorithms for solving such complex problems. Without a shift in how we approach and evaluate RL, we risk stagnating in our progress and failing to realize the full potential of RL in tackling real-world problems.

To address this, I propose a more general framework for RL. This framework is designed to be more flexible and adaptable than traditional methods, capable of learning from the complexities and nuances of infinite state space problems, and from the learning progress itself while interacting in those environments. The development and argumentation of this framework form a significant portion of the chapter, providing both a critique of existing methods and a potential path forward.

Another key point of discussion in this chapter is the strong connection between generalization and exploration in RL. This is a critical relationship that, I argue, has not been given the attention it deserves in the field. Generalization, the ability of an RL agent to apply learned knowledge to new situations, and exploration, the process of seeking out new information, are deeply intertwined. Understanding and leveraging this connection is, I believe, essential for future progress in RL. In this chapter I also provide experimental evidence supporting the hypothesis that the setup of evaluation can determine what kind of algorithmic advantages we maximize for, which are not aligned with solving large state space problems.

I emphasize this point throughout the chapter, arguing that a more focused approach to understanding the relationship between generalization and exploration could unlock significant advancements in RL. This could lead to more effective RL agents, capable of learning and adapting in ways that bring us closer to solving complex real-world problems.

In summary, the eighth chapter of my thesis offers a detailed and critical exploration of the challenges and potential solutions in RL. It provides a series of thoughtfully argued opinions and proposals that I believe will contribute significantly to the ongoing development of the field. Through a critique of traditional methods and the proposal of a new, more general framework, as well as an emphasis on the connection between generalization and exploration, this chapter represents a significant contribution to the discourse on RL.

6 Conclusion

In this thesis, we explored the role of exploration in Reinforcement Learning from a variety of angles and in different real-world problems. The process of exploration is integral and highly entangled in the entire reinforcement learning process of agents reaching optimal behaviours for achieving their goals. Through our research, we have gained a deeper understanding of the challenges and opportunities associated with exploration in RL, and have identified and developed a number of novel approaches for addressing these challenges.

Overall, our work has contributed to the field of RL by providing new insights into the mechanisms of exploration, and by demonstrating the significance of these mechanisms in a range of real-world applications. In particular, our findings have highlighted the importance of incorporating exploration strategies that can be tailored to the specific characteristics of the problem at hand, in order to achieve optimal performance.

While the contributions presented in this thesis advance our understanding of reinforcement learning, particularly in the domain of exploration, it is important to recognize that numerous questions and challenges persist within the field. Further research is necessary to comprehensively elucidate the role of exploration in reinforcement learning. In the following paragraphs, we will summarize our main findings and contributions (Section 6.1), and discuss the limitations and future directions of our research (Section 6.2).

6.1 Contributions

Key contributions of the research presented in this thesis, including any new insights or understanding that have been gained as a result of the study are as follows:

- An in-depth analysis and evaluation of several intrinsic reward mechanisms for improving RL in a suite of sparse reward environments which has led to defining preferable metrics for evaluating exploration progress, and also the proposal of an original intrinsic reward method.
- An original Intrinsic Reward mechanism that leverages, self-supervised, the knowledge of ordering states based on time in order to address exploration issues in RL. We show how learning the order of time from states, and using the "orderability" score of randomly shuffled states as intrinsic reward determines agents to converge towards less chaotic policies.

- An online, model-free algorithm to learn affordance-aware subgoal options and policy over options. We investigate the role of hard versus soft attention in training data collection, abstract value learning in hard exploration tasks, and handling a growing number of choices. We identify and empirically illustrate the settings in which the paradox of choice arises, i.e. when having fewer but more meaningful choices improves the learning speed and performance of a reinforcement learning agent.
- Demonstrate the advantage of using Deep Reinforcement Learning to learn tabula rasa, a robust policy in a collaborative mini-game built based on an extension of the theoretical game "stag hunt." Our research work, which identifies and integrates several improvements for a classical deep reinforcement learning algorithm, was able to learn in a multi-agent setup a policy that outperformed in a public online competition other RL policies and policies that used expert knowledge.
- We propose a novel way of posing single-agent problems as multi-agent setups and present methods for training RL agents in such a paradigm. We motivate this setup by being able to exploit the advantages of multiple agents, which can collaborate and communicate and benefit reinforcement learning with exploration challenges. We provide preliminary empirical evidence to support this hypothesis.
- Sim-to-real research transfer for RL in both indoor robot navigation and autonomous vehicle control. We first collected a large dataset of both simulated and real-world data, which we used to train RL algorithms for both tasks. We then improved the performance of these algorithms through a combination of hyperparameter optimization and exploration strategies, and adapted them to the real world by adjusting the policies to account for differences between the simulated and real-world environments. For the indoor robot navigation we propose a method to improve localization based on semantic knowledge transfer. For the autonomous vehicles policy we propose a self-supervised method for learning path prediction from which we extract steering angles. Finally, we evaluated the performance of the RL algorithms in closed-loop systems through a series of experiments, including collecting additional real-world data and comparing the algorithms' performance to a baseline or other control strategies. Overall, our results demonstrated the effectiveness of using simulation to train RL algorithms for both indoor robot navigation and autonomous vehicle control, and showed that it is possible to effectively transfer these algorithms to the real world through careful fine-tuning and adaptation.
- By researching de novo molecular design using machine learning algorithms we were able to present new insights for small molecule generation and propose improvements for increasing positive docking scores and diversity. We develop techniques for evaluating GFlowNet performance on a test set, and identify the most promising metric for predicting generalization for small molecule generation, which is a crucial aspect to consider for data efficiency and real world application transfer. Using Reinforcement Learning with specific designs for these types of problems, we are able to achieve both the desired diversity of generated small molecules and extremely high binding energies of docked compounds (using AutoDock Vina).
- We propose a novel Reinforcement Learning framework that can theoretically capture, with ease, more information necessary for learning more complex agent behaviors across different environments. This Reinforcement Learning training setup would empower the necessary information flow in order to meta-learn higher-order behaviors through a meta-agent that would learn to train other agents.
- We propose rethinking generalization in Reinforcement Learning through evaluation and inductive biases. We raise the importance of evaluating generalization for scaling Reinforcement Learning to real world complex problems and present experiments motivating this.

6.2 Future exploration in RL

This thesis has explored several key areas in reinforcement learning, from intrinsic rewards, hierarchical learning, multi-agent environments, robotics, to drug discovery. Throughout this journey, we have identified both strengths and limitations in the current state of RL, which together set the stage for future work in this exciting field.

Our first area of focus was exploring with intrinsic rewards. Although we were able to evaluate different intrinsic rewards and improve exploration with knowledge from time-based patterns, the benchmarks used did not cover a sufficiently diverse range of environments. The temporal intrinsic reward also did not disregard patterns outside the agents' control, which could be addressed by considering an inverse dynamics model. Future work should aim to introduce more diverse and complex environments to better evaluate and enhance the capacity of RL agents in handling various situations. Additionally, refining the temporal intrinsic reward mechanism to better distinguish and respond to various complex patterns is a promising area for further investigation. Intrinsic rewards play a pivotal role in reinforcement learning, presenting opportunities not just for the enhancement of RL but for a better understanding of intrinsic motivation in future real world agents or biological systems. Future research should aim at deciphering the patterns of motivation in biological entities, which could provide insights into engineering more efficient artificial IR mechanisms. Most of the current intrinsic reward methods work well for short-term dependencies but not for tasks that require long-term strategic thinking. Developing ways to handle long-term dependencies and even blending different IR methods could potentially foster more sophisticated, adaptive learning behaviors.

In studying hierarchical reinforcement learning, the scalability of affordances to larger environments was not fully realized. This points to a crucial direction for future work: to develop mechanisms that allow for the efficient scaling of affordances in increasingly complex environments, thereby enhancing the hierarchical understanding and decision-making capabilities of RL agents. Our findings from this research work underline the substantial role of focusing on rewarding experiences in improving RL. Looking ahead, we believe a promising research direction lies in investigating mechanisms that enhance the RL agent's attention towards high-signal experiences, as opposed to noise. The intent would be to devise more sophisticated strategies for distinguishing between relevant and irrelevant information in experiences, thereby optimizing the learning process. This line of inquiry could yield significant advancements, enabling systems to learn more effectively from their environment.

In the exploration of multi-agent environments, we were able to solve single tasks distributed by multiple agents that communicate. However, scaling the number of agents, and incorporating mechanisms such as distillation and restart of agent weights could potentially enhance the efficiency and adaptability of multi-agent systems. Future RL research holds promise in the development of multi-agent systems using learned communication. Key areas of focus could include adaptive exploration techniques and learning from heterogeneous experiences. This involves devising strategies that allow agents to dynamically modify their exploration based on peer feedback, potentially improving collective decision-making and learning. Concurrently, leveraging the diverse experiences of multiple agents offers an opportunity to provide RL systems with a more comprehensive understanding of their environment. By further exploring how to effectively assimilate these varied experiences, we can enhance the overall efficacy and adaptability of multi-agent RL systems. We can further enrich this idea by considering multi-agent systems, where multiple RL agents interact with the complex environments. Each agent could work on individual tasks, contributing to a larger collective goal. Through communication, these agents could share their hypotheses, findings, and learned behaviors, effectively building a shared 'cultural knowledge'. This concept could revolutionize the way we understand multi-agent systems, promoting more efficient problem-solving and knowledge-building through a form of 'collective intelligence'.

We looked into memory mechanisms, imitation learning, and transfer learning in the context of both indoor robotics and autonomous vehicles as part of our investigation into real-world robotic navigation. Despite these developments, it is still difficult to make agents self-adapt, so they can easily integrate with real-world data and adapt to new environments after simulator training. Future research should focus on creating models that enable RL agents to independently recalibrate while making inferences in response to changing environmental conditions. Such advancement would greatly improve their ability to navigate, strengthen their capacity to learn from practical experiences, and make the transition from simulated to real-world environments more seamless.

Our research made important advancements in enhancing diversity for de novo molecular discovery and facilitating generalization across a variety of proposed compounds within the broad field of drug discovery. The inclusion of uncertainty in the training process, however, has received little attention. Thus, further study should focus on incorporating uncertainty into RL (reinforcement learning) models for drug discovery. This development might significantly improve the proposed compounds' robustness and generalizability. It is also important to look into how the exploration-exploitation trade-off in the RL framework can be guided by a model's learned uncertainty, potentially accelerating the discovery process. Deep ensemble methods or the application of Bayesian approaches may also provide a way to effectively capture and make use of uncertainty in this situation.

As we continue to advance the domain of reinforcement learning, a critical area of future exploration lies beyond specific research areas and extends into broader, overarching directions. The development of RL agents that can interact with open-ended worlds and self-set goals promises a more dynamic mode of exploration and learning. This notion challenges traditional RL precepts, proposing a shift wherein agents determine their state-space rather than being limited by pre-defined environments. This transition empowers agents to shape their world, creating a more adaptable and responsive learning process. Simultaneously, there's potential to rethink our reliance on the traditional Markov Decision Process framework. By embracing a continuous flow of state-space interactions, we could enhance our agents' capabilities to handle novel situations, thereby pushing the boundaries of generalization and exploration. This approach would allow our agents to handle the vast complexities of a constantly changing, open-ended world, where they may never encounter the same state twice. Another promising avenue for future research could be the application of meta-learning to the exploration policy itself. This approach entails training RL agents to learn how to explore, instead of designing the intrinsic motivation that lead to better exploration. The goal would be to equip agents with the ability to adaptively determine the most effective exploration strategies based on their experiences, environment, and tasks at hand. By learning to generate an exploration policy, RL agents could better navigate unfamiliar environments or scenarios, leading to more efficient and effective learning. However, it's important to recognize that the pursuit of scale should not be seen as the sole solution, or even a desirable objective, in all RL research. Scaling up does bring about increased complexity and capabilities, but it can also introduce redundancy and inefficiency. An emphasis should be placed on creating more efficient, intelligent systems that leverage what they've learned to improve their performance. The focus should be on 'learning better', not just 'learning more'. With this balanced approach, we can strive towards the ultimate goal of RL: the creation of intelligent systems that can navigate and adapt to their environment in the same way humans do.

In conclusion, even though RL research has advanced significantly, vast areas of potential research to improve these algorithms are yet to be explored. However, by acknowledging our limitations and building upon our current knowledge, we can continue to push the boundaries of RL, moving ever closer to our goal of creating intelligent, adaptable, and efficient learning systems.

Bibliography

[1] Dario Amodei and Jack Clark. Faulty Reward Functions in the Wild. *https://blog.openai.com/faulty-reward-functions/*, 2016.

[2] Ryan Paul Badman, Thomas Trenholm Hills, and Rei Akaishi. Multiscale Computation and Dynamic Attention in Biological and Artificial Intelligence. *Brain Sciences*, 10(6):396, 2020.

[3] Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation. *arXiv preprint arXiv:2106.04399*, 6 2021. URL http://arxiv.org/abs/2106.04399.

[4] Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation. *Advances in Neural Information Processing Systems*, 33:27381–27394, 6 2021. ISSN 10495258. URL http://arxiv.org/abs/2106.04399.

[5] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with Large Scale Deep Reinforcement Learning. 2021. URL https://www.facebook.com/OGDota2/.

[6] Ali Borji, D N Sihite, and L Itti. Salient Object Detection: A Benchmark. Computer Vision—ECCV 2012: the 12th European Conference on Computer Vision; 2012 Oct 7-13; Florence, Italy, 2012.

[7] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.

[8] Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic Gridworld Environment for OpenAI Gym. \url{https://github.com/maximecb/gym-minigrid}, 2018.

[9] Suresh Dara, Swetha Dhamercherla, Surender Singh Jadav, CH Madhu Babu, and Mohamed Jawed Ahsan. Machine Learning in Drug Discovery: A Review. *Artificial Intelligence Review*, 55(3): 1947–1999, 3 2022. ISSN 0269-2821. doi: 10.1007/s10462-021-10058-4. URL https://link.springer.com/10.1007/s10462-021-10058-4.

[10] Ildefons Magrans de Abril and Ryota Kanai. Curiosity-driven reinforcement learning with homeostatic regulation. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2018.

[11] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de las Casas, Craig Donner, Leslie Fritz, Cristian Galperti, Andrea Huber, James Keeling, Maria Tsimpoukelli, Jackie Kay, Antoine Merle, Jean Marc Moret, Seb Noury, Federico Pesamosca, David Pfau, Olivier Sauter, Cristian Sommariva, Stefano Coda, Basil Duval, Ambrogio Fasoli, Pushmeet Kohli, Koray Kavukcuoglu, Demis Hassabis, and Martin Riedmiller. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature 2022 602:7897*, 602(7897):414–419, 2 2022. ISSN 1476-4687. doi: 10.1038/s41586-021-04301-9. URL https://www.nature.com/articles/s41586-021-04301-9https://www.nature.com/articles/s41586-021-04301-9%E2%80%A6.

[12] Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. Challenges of Real-World Reinforcement Learning. *arXiv preprint arXiv:1904.12901*, 2019.

[13] Richard E Fikes, Peter E Hart, and Nils J Nilsson. Learning and executing generalized robot plans. *Artificial Intelligence*, 3:251–288, 1 1972.

[14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. 6 2014. URL http://arxiv.org/abs/1406.2661.

[15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. 12 2014. URL http://arxiv.org/abs/1412.6572.

[16] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9 (4):188–194, 2005.

[17] Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant Diffusion for Molecule Generation in 3D. 3 2022. URL http://arxiv.org/abs/2203.17003.

[18] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[19] Alex Irpan. Deep Reinforcement Learning Doesn't Work Yet. \url{https://www.alexirpan.com/2018/02/14/rl-hard.html}, 2018.

[20] Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M. Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, Chrisantha Fernando, and Koray Kavukcuoglu. Population Based Training of Neural Networks. 11 2017. URL http://arxiv.org/abs/1711.09846.

[21] Andrew Jaegle, Vahid Mehrpour, and Nicole Rust. Visual novelty, curiosity, and intrinsic reward in machine learning and the brain. *Current Opinion in Neurobiology*, 58:167–174, 10 2019. ISSN 09594388. doi: 10.1016/j.conb.2019.08.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0959438819300054.

[22] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 2106–2113. IEEE, 2009.

[23] Jerome Kagan. Motives and development. *Journal of personality and social psychology*, 22(1):51, 1972.

[24] Khimya Khetarpal, Martin Klissarov, Maxime Chevalier-Boisvert, Pierre-Luc Bacon, and Doina Precup. Options of Interest: Temporal Abstraction with Interest Functions. *AAAI*, 2020.

[25] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. 9 2016. URL http://arxiv.org/abs/1609.02907.

[26] Guillaume Lample and Devendra Singh Chaplot. Playing FPS Games with Deep Reinforcement Learning. In *AAAI*, pages 2140–2146, 2017.

[27] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 5 2015. ISSN 0028-0836. doi: 10.1038/nature14539. URL http://www.nature.com/articles/nature14539.

[28] Tod S Levitt. Qualitative navigation for mobile robots. *Int. J. Artificial Intelligence*, 44:305–360, 1990.

[29] Kit-Kay Mak and Mallikarjuna Rao Pichika. Artificial intelligence in drug development: present status and future prospects. *Drug discovery today*, 24(3):773–780, 3 2019. ISSN 1878-5832. doi: 10.1016/j.drudis.2018.11.014. URL http://www.ncbi.nlm.nih.gov/pubmed/30472429.

[30] Dominic D. Martinelli. Generative machine learning for de novo drug discovery: A systematic review. *Computers in Biology and Medicine*, 145:105403, 6 2022. ISSN 00104825. doi: 10.1016/j.compbiomed.2022.105403. URL https://linkinghub.elsevier.com/retrieve/pii/S0010482522001950.

[31] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data, 4 2017. ISSN 2640-3498. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

[32] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, and others. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[33] Varnavas D. Mouchlis, Antreas Afantitis, Angela Serra, Michele Fratello, Anastasios G. Papadiamantis, Vassilis Aidinis, Iseult Lynch, Dario Greco, and Georgia Melagraki. Advances in De Novo Drug

Design: From Conventional to Machine Learning Methods. *International Journal of Molecular Sciences*, 22(4):1676, 2 2021. ISSN 1422-0067. doi: 10.3390/ijms22041676. URL https://www.mdpi.com/1422-0067/22/4/1676.

[34] Anna C Nobre and Mark G Stokes. Premembering experience: a hierarchy of time-scales for proactive attention. *Neuron*, 104(1):132–146, 2019.

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[36] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. 3 2022. URL http://arxiv.org/abs/2203.02155.

[37] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017.

[38] Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Deep Reinforcement Learning for De-Novo Drug Design. *Science Advances*, 4(7), 11 2017. doi: 10.1126/sciadv.aap7885. URL http://arxiv.org/abs/1711.10907http://dx.doi.org/10.1126/sciadv.aap7885.

[39] V Srinivasa Rao and K Srinivas. Modern drug discovery process: An in silico approach. *Journal of Bioinformatics and Sequence Analysis*, 2(5):89–94, 2011. ISSN 2141-2464. URL http://www.academicjournals.org/JBSA.

[40] Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment and state-dependent noise-an intrinsic motivation for avoiding unpredictable agents. In *Artificial Life Conference Proceedings 13*, pages 118–125. MIT Press, 2013.

[41] F Schler. 3d path planning for autonomous aerial vehicles in constrained spaces [Ph. D. thesis]. *Department of Electronic Systems, Faculty of Engineering and Science, Aalborg University, Aalborg, Denmark*, 2012.

[42] Jürgen Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In *Workshop on Anticipatory Behavior in Adaptive Learning Systems*, pages 48–76. Springer, 2008.

[43] Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

[44] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[45] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 10 2017. ISSN 0028-0836. doi: 10.1038/nature24270. URL http://www.nature.com/articles/nature24270.

[46] David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021. doi: 10.1016/j.artint.2021.103535. URL www.elsevier.com/locate/artint.

[47] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. 3 2015. URL http://arxiv.org/abs/1503.03585.

[48] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[49] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[50] Richard S Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

[51] Richard S Sutton, Ashique Rupam Mahmood, and Martha White. An Emphatic Approach to the Problem of Off-policy Temporal-Difference Learning. *Journal of Machine Learning Research*, 17: 73:1–73:29, 2016.

[52] Chen Tessler, Shahar Givony, Tom Zahavy, Daniel J Mankowitz, and Shie Mannor. A Deep Hierarchical Approach to Lifelong Learning in Minecraft. In *AAAI*, pages 1553–1561, 2017.

[53] Chris Thornton, Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Auto-WEKA. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, volume Part F128815, pages 847–855, New York, NY, USA, 8 2013. ACM. ISBN 9781450321747. doi: 10.1145/2487575.2487629. URL https://dl.acm.org/doi/10.1145/2487575.2487629.

[54] Isaac Triguero, Salvador García, and Francisco Herrera. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information Systems*, 42(2): 245–284, 2 2015. ISSN 0219-1377. doi: 10.1007/s10115-013-0706-y. URL http://link.springer.com/10.1007/s10115-013-0706-y.

[55] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, and Shanrong Zhao. Applications of machine learning in drug discovery and development. *Nature reviews. Drug discovery*, 18(6): 463–477, 6 2019. ISSN 1474-1784. doi: 10.1038/s41573-019-0024-5. URL http://www.ncbi.nlm.nih.gov/pubmed/30976107http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6552674.

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. 6 2017. URL http://arxiv.org/abs/1706.03762.

[57] O Vinyals, I Babuschkin, J Chung, M Mathieu, M Jaderberg, W Czarnecki, A Dudzik, A Huang, P Georgiev, R Powell, and others. AlphaStar: Mastering the Real-Time Strategy Game StarCraft II, 2019.

[58] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, and others. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[59] Martha White. Unifying Task Specification in Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning, {ICML} 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 3742–3750, 2017.

[60] Olivier J Wouters, Martin McKee, and Jeroen Luyten. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA*, 323(9):844–853, 2020.

[61] Fei Yan, Yi-Sha Liu, and Ji-Zhong Xiao. Path planning in complex 3D environments using a probabilistic roadmap method. *International Journal of Automation and computing*, 10(6):525–533, 2013.

[62] Namik Kemal Yilmaz, Constantinos Evangelinos, Pierre F J Lermusiaux, and Nicholas M Patrikalakis. Path planning of autonomous underwater vehicles for adaptive sampling using mixed integer linear programming. *IEEE Journal of Oceanic Engineering*, 33(4):522–537, 2008.

[63] Xianyuan Zhan, Haoran Xu, Yue Zhang, Xiangyu Zhu, Honglei Yin, and Yu Zheng. DeepThermal: Combustion Optimization for Thermal Power Generating Units Using Offline Reinforcement Learning. 2 2021. URL http://arxiv.org/abs/2102.11492.

[64] Ruohan Zhang, Calen Walshe, Zhuode Liu, Lin Guan, Karl S Muller, Jake A Whritner, Luxin Zhang, Mary M Hayhoe, and Dana H Ballard. Atari-head: Atari human eye-tracking and demonstration dataset. *arXiv preprint arXiv:1903.06754*, 2019.

[65] Barret Zoph and Quoc V. Le. Neural Architecture Search with Reinforcement Learning. 11 2016. URL http://arxiv.org/abs/1611.01578.