SIP-14
REP 2017

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México siendo las 16:00 horas del día 30 del mes de junio del 2020 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Posgrado del: **Centro de Investigación en Computación** para examinar la tesis titulada:

"Cellular Automata Control with Deep Reinforcement Learning" del (la) alumno (a):

Apellido Paterno:	BECERRA	Apellido Materno:	SOTO	Nombre (s):	EMANUEL

Número de registro: A 1 8 0 9 6 8

Aspirante del Programa Académico de Posgrado: **Maestría en Ciencias de la Computación**

Una vez que se realizó un análisis de similitud de texto, utilizando el software antiplagio, se encontró que el trabajo de tesis tiene 6 % de similitud. **Se adjunta reporte de software utilizado**.

Después que esta Comisión revisó exhaustivamente el contenido, estructura, intención y ubicación de los textos de la tesis identificados como coincidentes con otros documentos, concluyó que en el presente trabajo **SI** [ ] **NO** [X] **SE CONSTITUYE UN POSIBLE PLAGIO.**

JUSTIFICACIÓN DE LA CONCLUSIÓN: *(Por ejemplo, el % de similitud se localiza en metodologías adecuadamente referidas a fuente original)*

El porcentaje de similitud es bajo y se localiza en frases cortas y comunes, asi como en textos adecuadamente referenciados.

****Es responsabilidad del alumno como autor de la tesis la verificación antiplagio, y del Director o Directores de tesis el análisis del % de similitud para establecer el riesgo o la existencia de un posible plagio.**

Finalmente, y posterior a la lectura, revisión individual, así como el análisis e intercambio de opiniones, los miembros de la Comisión manifestaron **APROBAR** [X] **NO APROBAR** [ ] la tesis, en virtud de los motivos siguientes:

_Cumple con todos los requisitos estipulados en el reglamento de estudios de posgrado del IPN.

COMISIÓN REVISORA DE TESIS

Dr. Rolando Menchaca Méndez
Director de tesis

Dr. Juan Carlos Chimal Eguía

Dr. Carlos Alberto Duchanoy Martínez

Dr. Ricardo Menchaca Méndez
2º Director de Tesis

Dr. Francisco Hiram Calvo Castro

Dr. Mario Eduardo Rivero Ángeles

INSTITUTO POLITECNICO NACIONAL
CENTRO DE INVESTIGACION EN COMPUTACION
DIRECCION

Dr. Marco Antonio Moreno Ibarra
PRESIDENTE DEL COLEGIO DE PROFESORES

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESIÓN DE DERECHOS

En la Ciudad de México el día 14 del mes septiembre del año 2020, el que suscribe Emanuel Becerra Soto alumno del Programa de Maestría en Ciencias de la Computación con número de registro A180968, adscrito al Centro de Investigación en Computación, manifiesta que es autor intelectual del presente trabajo de tesis bajo la dirección del Dr. Ricardo Menchaca Méndez y del Dr. Rolando Menchaca Méndez y cede los derechos del trabajo titulado "*Cellular Automata Control with Deep Reinforcement Learning*", al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección *elbecerrasoto@gmail.com*. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Emanuel Becerra Soto

Nombre y firma

Resumen

El novedoso campo del Aprendizaje Profundo por Refuerzo promete una nueva clase de algoritmos generales capaces de aprender directamente de interacciones con el mundo, ya sean reales o simuladas. Debido a esto la popularidad de este campo ha crecido en los últimos años. A pesar de su enorme potencial, aún faltan aplicaciones concretas en la vida real. En este documento se propone un nuevo ambiente de Aprendizaje por Refuerzo basado en Autómatas Celulares, los cuales son ejemplos característicos de sistemas complejos. Particularmente se usó el Autómata Celular de Incendios Forestales como la dinámica subyacente para un agente, el cual afecta la cuadrícula cambiando celdas tipo fuego por celdas tipo vacías. La semántica corresponde a la de un "Helicóptero" tratando de apagar un incendio forestal. Para tratar de resolver el ambiente propuesto se empleó el algoritmo de *Deep Q-networks* sobre una cuadrícula de 3×3. El agente fue evaluado durante 100,000 pasos de una simulación del ambiente. Nuestros resultados muestran que los Autómatas Celulares pueden ser usados como punto de

referencia para algoritmos de Aprendizaje por Refuerzo, además el uso de Autómatas Celulares, por sus capacidades de modelado, podría reducir la brecha entre simulación y aplicación a la vida real.

Abstract

The recent field of Deep Reinforcement Learning promises a new class of general algorithms capable of learning from real or simulated world interactions. Hence the field popularity has exploded in the last years. Despite its potential, real world concrete applications are lacking. This document proposes a novel Reinforcement Learning environment based on Cellular Automata, which are quintessential models of complex systems. Particularly a Forest Fire Cellular Automata was used as underlying dynamics for an agent that can affect the grid configuration by changing fire cells to empty cells. The semantics are those of a "Helicopter" trying to extinguishing a wild fire. We applied Deep Q-networks to solve the proposed environment on a 3×3 grid. We evaluate our agent on a 100,000 steps simulation of the environment. Our results show that Cellular Automata can be used as benchmark for Reinforcement Learning algorithms, likewise its modeling capabilities could help to narrow the gap to real world applications.

Acknowledgments

Writing is hard. But this is a known fact. What is not known is the particular environment and situations that any writer must endure finishing its work. I dedicate this thesis to all the persons that somehow, consciously or unconsciously, have reduced the burden of writing by supporting me with "everything else". Without them, this work would be impossible to finish.

To my family,
including pets, for always being there, no matter what.

To my friends,
who help me to put things in perspective and enjoy the ride.

To my professors,
not only for their top-notch classes but giving me invaluable advice.

To my advisors,
for transmitting me their passion for computer science and for showing me that work-life balance is possible in this day and age but more importantly that one should love what it does.

And to the "Centro de Investigación en Computación" and the "Instituto Politécnico Nacional", for providing a wonderful and life-changing experience.

This document was possible due to the economic support given by CONACYT.

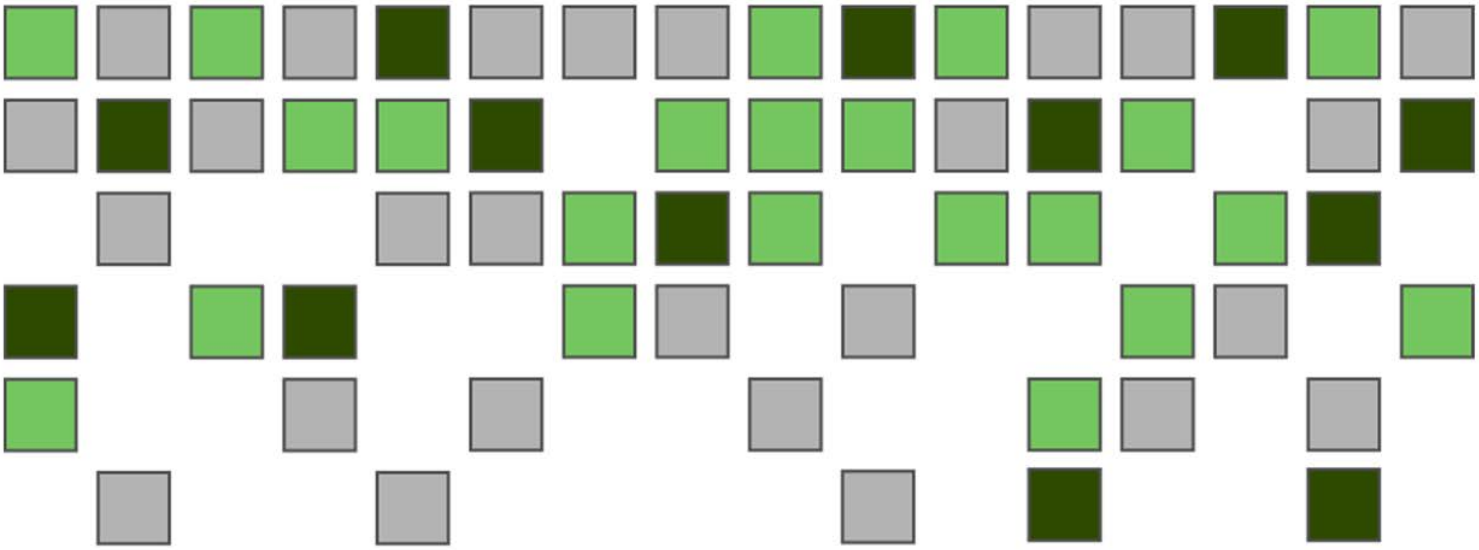
Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Explanation
AI	Artificial Intelligence
ALE	Arcade Learning Environment
ANN	Artificial Neural Networks
API	Application programming interface
CA	Cellular Automata or Cellular Automaton
CAM	Cellular Automata Machines
DQN	Deep Q-Networks
DRL	Deep Reinforcement Learning
DSM	Drossel ans Schwabl Model
FFEM	Forest Fire Environment Maker
MDP	Markov Decision Process
ML	Machine Learning
MLP	Multilayer Perceptron
OAGA	Open AI Gym API
POMDP	Partially Observable Markov Decision Process
RL	Reinforcement Learning



Cellular Automata Control with Deep Reinforcement Learning

By Emanuel Becerra Soto

1

Introduction

1.1 Motivations

Since antiquity the idea of building a thinking machine has always been in the minds of philosophers, artists, inventors, kings, and commoners alike, filling us with wonder, terror, and contemplation. A human creation capable of human feats would turn us, at least in an allegorical sense, into gods. This long time dream is an inspiration for the field of *Artificial Intelligence* (AI). AI was born in 1956 on the *Dartmouth Summer Research Project on Artificial Intelligence*. In the early days of AI, the kind of problems that were tackled were the ones that are easy for a computer to solve but difficult for humans. These kinds of tasks are logical and mathematical in their very nature. Ironically the kind of

tasks that are easy for humans have proved to be difficult to code on computers. Like visual identification of objects or navigation through complex environments. Because of their apparent easiness we seldom stop to think about how we actually manage to solve them and coming with a precise set of rules is labor-intensive and error-prone. An approach to solve those kinds of tasks is called *Machine Learning* (ML). The basic idea of machine learning is to use data on the task at hand and let the computer figure out how to best use the data to solve the problem, in a figurative way learning from observation. The promise of such methods is avoiding explicit programming to solve problems and perhaps getting us closer to the faraway dream of a thinking machine.

Reinforcement Learning (RL) is a branch of ML that is inspired by how we learn through interaction with the real world. An agent interacts with an environment a receives a signal, called *reward*. The main idea of RL is that the agent should accumulate as much *reward* as possible. In doing so it would accomplish whatever its purpose is. This is known as the *reward hypothesis*, in other words, getting plenty of favorable rewards is great for whatever the agent is trying to do. The agent, from its current knowledge, chooses an *action* and then receives a *reward* and an *observation* from the environment, this process is repeated over and over while the agents seeks for its goal by accumulating *reward*. In summary, the *RL problem* is how to correctly act to accomplish a goal by looking for behavior that maximizes an accumulated *reward* quantity, know as *return*.

A popular perspective on AI and ML is deep learning. *Deep Learning* is a myriad of techniques based on *artificial neural networks* (ANN). ANN are loosely based in neuroscience and the brain structure. They are composed of several arrays of computing units called *neurons*. Neurons are arranged in hierarchies called layers. The processing of an ANN is done in an orderly fashion by the layers. When an input is received it is processed by the first layer, then the second and so on, until the last layer, which outputs the final result. If a technique or method is designated as *deep learning* it only implies that such technique employs an ANN

with a bunch of layers.

Naturally, RL and deep learning can be combined, this union is known as *Deep Reinforcement Learning* (DRL). DRL has protagonized recent successes with superhuman performance on playing backgammon (Tesauro 1995), video games (Mnih et al. 2015), poker (Moravčík et al. 2017), multiplayer video games (Jaderberg et al. 2019), chess (Silver et al. 2018) and the decades-long challenging game of go (Silver et al. 2016)(Silver et al. 2018).

One important aspect of advancing an AI area is to have adequate benchmarks. *Benchmarks* are key problems that are widely used for the evaluation and design of algorithms. Having good performance on a benchmark is used to showcase the prowess of an algorithm. Nonetheless putting too much faith into a single benchmark could hurt research in the long run (Hooker 1995)(Stanley and Lehman 2015). Despite their pitfalls, they are essential tools of AI research. To name an important example, the *ImageNet* dataset (Deng et al. 2009) and its associated challenge the *Large Scale Visual Recognition Challenge* (Russakovsky et al. 2015) were fundamental for setting the current popularity of deep learning techniques and the bonanza in the field of computer vision.

In the scope of RL, a widely accepted benchmark is the *Arcade Learning Environment* (ALE) (Bellemare et al. 2013). ALE is a suite of more than fifty games from the Atari 2600 console. Moreover, frameworks such as *Open AI Gym* (Brockman et al. 2016) have lowered the barrier entry for using ALE, making Atari games easily accessible for RL algorithms and their training and evaluation. ALE's primary strengths are its variety of tasks and the independent creation of each game which reduces the bias of a single manufacturer. If a single algorithm is capable of performing well on a variety of very different and difficult tasks, it would be conceivable to claim its generality. Thus, ALE challenges any general RL algorithm to play well on as many as possible Atari games. A key event for the popularization of Atari games as a benchmark was their utilization on the seminal *Deep Q-Networks* (DQN) papers (Mnih et al. 2013)(Mnih et al. 2015).

The DQN algorithm achieved a remarkable performance on a dozen of Atari games, receiving as input just raw pixels, hence without the games rules and from a single set of hyperparameters. DQN revived the interest in DRL methods.

Other common benchmarks for RL are, GridWorlds (Sutton and Barto 2018), robotics simulations with the MuJoCo physics engine (Todorov, Erez, and Tassa 2012), classic control tasks (Brockman et al. 2016) and a heterogeneous variety of game theoretical tasks and board games (Lanctot et al. 2019).

Despite their broad success, Atari games suffer from their lack of environments with high stochasticity, which allows for the exploitation of such determinism by simply algorithms (Bellemare et al. 2015)(Machado et al. 2018). Other criticisms arises from the emulation component of the games, that functionally transforms them into black boxes. This issue is compounded when some RL methods are also black boxes, making debugging extremely difficult (Foley et al. 2018). The emulation also hinders the parametrization of the environments and the extraction of meaningful statistics to broaden the understanding of why some RL algorithms work and others do not.

The central idea of this document is to show that *Cellular Automata* (sg. *Cellular Automaton*) are suitable for the design and evaluation of RL environments. *Cellular Automata* (CA) are mathematical and computational systems with interesting characteristics and a rich history of theoretical background (Ilachinski 2001). CA are systems with discrete and local dynamics and a discrete evolution. They have been traditionally used to model complex systems, on the natural and social sciences like the weather, wildfires, the immune system, pattern formation, galaxy formation, earthquakes, ecology, chemotaxis, epidemics, cell to cell interaction, reaction-diffusion systems, city traffic, social group formation, racial segregation, economics, just to name a few broad examples (Ganguly et al. 2003)(Hoekstra, Kroc, and Sloot 2010). Furthermore, due to their complete discrete nature, CA can be seen as simple computational devices and provide an ideal model for parallel and unconventional computing. Of exceptional

importance is that some CA have *universal computation* capabilities, in other words, some CA have the same computing power of a *Turing Machine* and thus capable of executing any algorithm.

What makes a good benchmark? is a tough question to answer, since each field has its requirements and expectations, and tradeoffs must be made between many design dimensions such as *difficulty*, *relevance*, *accessibility*, *reproducibility*, *interpretability* and so on. We believe that CA would make a good benchmark for RL for the following reasons:

1. Easiness of implementation.
2. Capable of modeling complex systems.
3. Computing models (*universal*, *parallel*, *unconventional*).
4. Highly parameterizable.

1.2 Problem Statement

In this thesis, we consider the advantages of using Cellular Automata as underlying environments for RL tasks. To fully incorporate a Cellular Automaton into the RL framework a *Markov Decision Process* (MDP) should be defined. So we need to be able to find appropriate sets of *Actions* ($\mathcal{A}$), *States* ($\mathcal{S}$) and *Rewards* ($\mathcal{R}$), also specify an *Agent* that interacts with the CA and explicitly or implicitly define the transitions of the MDP $p(s', r|s, a) : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$.

The particular CA that we would be using through this document is the *Drossel and Schwabl model* (DSM) (Drossel and Schwabl 1992) also referred as the Forest Fire CA. DSM captures the dynamics of wildfires. Thus our first challenge is engineering a RL task that successfully integrates the DSM with some practical semantics. Thus, we propose an *Agent* or "Helicopter" flying over the forest trying to extinguish the wildfire. For an initial characterization, a grid of 3 *times*3 will be used. Therefore, the set of all states $\mathcal{S}$ of the proposed environment

will be the configurations of a 3×3 forest fire CA, with a tuple of helicopter positions (row, col) and an internal parameter m that takes two values $\{0, 1\}$ and controls the synchronization of helicopter movements and CA computations. The set of all actions $\mathcal{A}$ is the possible movement to all cells in a Moore's neighborhood from the agent position *{"Left-Up", "Up", "Right-Up", "Right", "Stay", "Left", "Left-Down", "Down", "Right-Down"}*. The rules that implicitly define the dynamics p of the environment are:

1. The selection of any action would affect the "Helicopter" position tuple (unless the *"Stay"* action was chosen), this is referred as "moving".
2. After moving to any cell the "Helicopter" extinguishes the fire by changing, if applicable, a current "fire" cell to an "empty" cell.
3. The m parameter is decreased by 1 after each "Helicopter" movement.
4. Before moving, if $m = 0$ then the grid is updated following the Forest Fire CA rules, then m is reset to its initial value ($m = 1$).

After precisely defining the environment (see Appendix A 5.1.1), the semantics are those of a helicopter trying to extinguish a wildfire in a simulated forest, thus the optimization problem that we want to solve is:

> Minimize the accumulated count of fire cells for a given time interval.

1.3 Objectives

1.3.1 Main Objectives

- Propose a novel environment for Reinforcement Learning tasks, based on Cellular Automata, that could be used as an alternative benchmark instead of Atari games.
- Characterize the environment by solving it by state of the art methods.

1.3.2 Specific Objectives

- Select a Cellular Automaton model for the environment, in this case the Forest Fire Cellular Automaton.
- Design a RL task incorporating the CA dynamics.
- Implementation of the RL environment, following the Open AI gym API.
- Apply DQN to solve the proposed task.

2

Cellular Automata

2.1 Introduction

Cellular Automata (sg. *Cellular Automaton*) are computational and mathematical systems with two essential characteristics: a discrete structure and a local interaction between its parts, but its killer feature is that they are simple, yet capable of producing complex behavior.

A quick *Google Scholar* search for the term "Cellular Automata", retrieves roughly 13,000 research articles (from 2019 to mid-2020). Despite not being as popular (by the same metric) as other topics, like "Cancer" (~128,000) or "Deep Learning" (~104,000), they are, arguably, still popular.

The topic of CA dates back to the 1950s spanning a history of 70 years. Since then the mathematical and practical properties of CA have been studied and its applications have been explored in different branches of science including physical and social.

Perhaps the reason behind this popularity is their simplicity. CA are composed of decentralized interactions of their individual parts (cells), these cells are usually arranged in a very regular grid and are updated following the same rules for all cells. The rules only take into account the vicinity of a given cell. From this design specifications, one could naively anticipate that a very homogeneous state is reached after some iterations of the CA. But this is not always the case, surprisingly for some simple configurations and rules, complex behavior emerges. This behavior is rich enough to be used to model natural systems. The structures that arise from local interactions were not designed *a priori* and their nature is capricious as they could be oscillating, chaotic, ordered, random, transient, stable. Their scale is also far from the initial size of the cell neighborhoods.

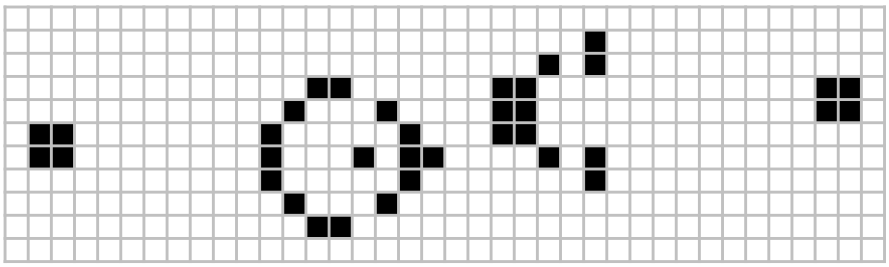


Figure 2.1: A glider gun, a famous pattern, from Conway's Cellular Automaton "Game of Life". Image taken from: Wikipedia.

2.2 Main Characteristics

Cellular Automata are mathematical objects that are mainly characterized by (Ilachinski 2001):

1. A discrete lattice of cells: A n-dimensional arrangement of cells, usually 1-D, 2-D or 3-D.

2. Homogeneity: Cells are equivalent in the sense that they share an update function and a set of possible states.

3. Discrete states: Each cell is in one state from a finite set of possible states.

4. Local Interactions: Cell interactions are local, this is given by the update function being dependent on neighboring cells.

5. Discrete Dynamics: The system evolves in discrete time steps. At each step the update function is applied simultaneously (synchronously) to all cells.

2.3 Mathematical Definition

The following is adapted from the book Probabilistic Cellular Automata (Louis and Nardi 2018).

Cellular Automata are dynamical systems of interconnected finite-state automata (cells). The cell evolution is through discrete time steps and it is dictated by a function dependent on a neighborhood of interacting cells.

The main mathematical aspects of a CA are:

- The network G: A graph G.

$$G = (V(G), E(G))$$

  The set of vertices $V(G)$ represents the location of the cells. The set of edges $E(G)$ describes the spatial relations between the cells.

- The alphabet S: Defines the states that each cell can take. In common CA settings S is a finite set. It is also called *local space* or *spin space*.

- The configuration space $S^{V(G)}$: This is the set of all possible states of the

CA. A specific configuration is denoted as:

$$\sigma = \{\sigma_k \in V(G)\}$$

σ_k is the configuration of the cell at position k.

- The neighborhoods V_k:

$$V_k \subset V(G)$$

The subset of nodes that can influence or interact with the cell at $k \in V(G)$ (ordinarily it includes itself). A typical configuration for V_k is: $G = \mathbb{Z}^2$, and $V_k = \{k, k \pm e_1, k \pm e_2\}$, where (e_1, e_2) is the canonical basis of $\mathbb{Z}^2$, (north/south, east/west). This is known as the *von Neumman neighborhood*.

- The global update F:

$$F : S^{V(G)} \to S^{V(G)}$$

$$(F(\sigma))_k = f_k(\sigma_{V_k})$$

The global update F is calculated by applying a local function f_k per cell at k. In the classical setting f_k is the same for all the cells.

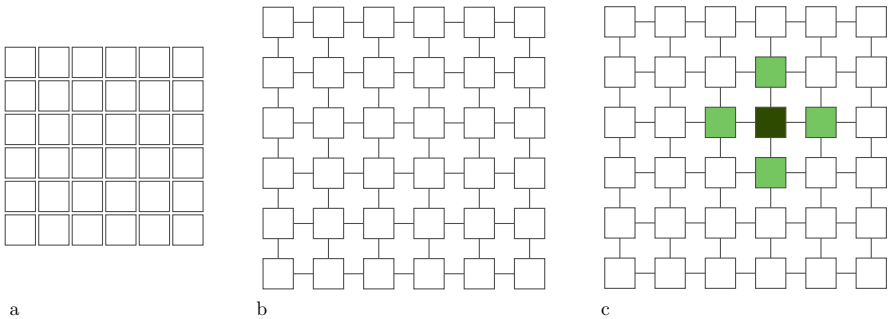


Figure 2.2: The underlying graph structure of a Cellular Automaton. A highlighted cell and its neighbors (c). Image adapted from (Hoekstra, Kroc, and Sloot 2010).

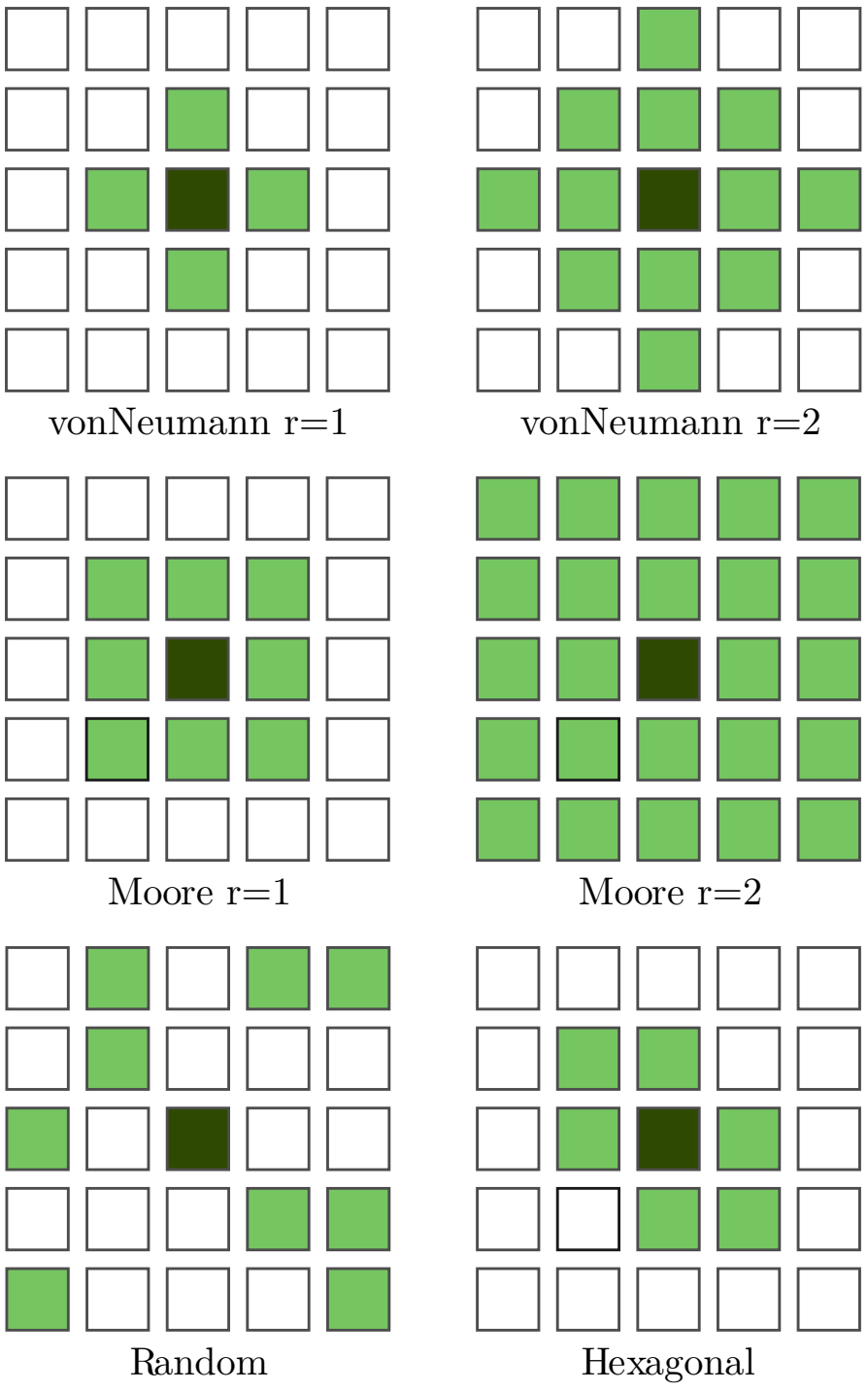


Figure 2.3: Different types of neighborhoods. Image adapted from (Hoekstra, Kroc, and Sloot 2010).

2.4 Classification

Wolfram from extensive simulations of 1-D CA grouped the general behavior of CA in four informally defined classes (Wolfram 2002). The classification is done by simulation from a variety of lattice initializations to broadly characterize the CA.

1. Class 1: A quick evolution towards a homogeneous global state is observed. All cells stop changing and all the randomness in the initial configuration disappears.

2. Class 2: The CA evolution leads to isolated patterns that could be periodic or stable.

3. Class 3: Pseudo-random or chaotic patterns emerge. Any stable structure is quickly destroyed by the surrounding noise.

4. Class 4: This is the most interesting type of behavior. Patterns that interact in a complex way emerge. These complex patterns are stable for long periods of time. Eventually, the complex patterns can settle into a global state like Class 2 behavior but this can take a vast amount of time. Wolfram has conjectured that many Class 4 CA are capable of universal computation (Wolfram 2002).

This classification was inspired by the behavior observed in continuum dynamical systems. For example, the homogeneous states of class 1 CA are analogous to fixed-point attracting states, or the repeating structures of class 2 CA are analogous to continuous limit cycles and class 3 chaotic patterns are analogous to strange attractors, while class 4 behavior does not have an obvious continuum analogy (Ilachinski 2001).

Other classification schemes can be found in the literature like based on the structure of their attractors (Kůrka 1997) or by the structure of their "Garden of Eden" states (non-reachable global states) (Kari 1994).

2.5 Modeling of Complex Systems

Since the consolidation of the scientific method in the XVII century, two methodologies emerged for generating and evaluating scientific knowledge. Them being the "experimental" and the "theoretical" paradigms. The experimental paradigm is concerned with observing, measuring and quantifying natural phenomena in order to test a hypothesis. Experimentation also can be made in a playful manner to collect and organize data. The theoretical paradigm seeks logical and mathematical explanations of natural phenomena. Both paradigms are complementary in the sense that predictions can be made by the theoretical paradigm and be tested using the experimental one, then if the experimental findings support the predictions the theory is kept otherwise is rejected. In other words, theory can be supported or falsified through experimentation.

A third scientific paradigm recently appeared, namely the "computational" paradigm. In this approach, the study of nature is done through computer simulations. The computational paradigm works in partnership with the "experimental" and "theoretical" ones, so where observed phenomena are not easily tractable by analytical descriptions or direct experimentation is not allowed, computational simulation still permits further inquiry. Additionally when possible the outputs from the computations can be validated against experimental data and predictions from the theory, thus establishing helpful feedback between paradigms. The invention of the digital computer enabled the numerical solution of analytical models by means of discretization of quantities. Finally, the computational paradigm can sometimes be used as a shortcut to the theoretical or experimental paradigms, for example, if the problem at hand is well described, data obtained from a simulation could be employed as a proxy for real world data or by the contrary if the studied phenomena are poorly characterized a first computational approach could be performed to a gain further understanding that may enable an experimental or a theoretical approach.

Table 2.1: How different models handle *state*, *space* and *time*. "C" stands for continuous and "D" for discrete. The discrete nature of CA is highlighted. Table adapted from the book "Simulating Complex Systems by Cellular Automata" (Hoekstra, Kroc, and Sloot 2010).

Type of model	State	Space	Time
Partial differential equations (PDEs)	C	C	C
Integro-difference equations	C	C	D
Coupled ordinary differential equations (ODEs)	C	D	C
Interacting particle systems	D	D	C
Coupled map lattices (CMLs)	C	D	D
Systems of difference equations	C	D	D
Lattice Boltzmann equations (LBEs)	C	D	D
Cellular Automata (CA)	**D**	**D**	**D**
Lattice gas automata (LGAs)	D	D	D

The theoretical and computational paradigms force us to formally disclose our assumptions in the form of variables, processes, and relationships among them, thus forming what is known as an "abstract model" or simply a "model". Mathematical and computational models can be grouped according on how *state*, *space* and *time* variables are abstracted into the model (Hoekstra, Kroc, and Sloot 2010).

Complex Systems is a broadly defined term to encompass dynamical systems with more than a few interacting parts commonly in a non-linear way, in other words, systems that have emergent properties from individual interactions. These kinds of systems are common in the natural and social sciences. One of the classical examples is the formation of ant colonies and the organization that comes with it (Hölldobler, Wilson, and others 1994). Each ant is sensing external stimuli and acting upon them, the cumulative reactions of all the ants culminate in the building of the ant-hill, feeding it, defending it, and attacking other ant colonies. Everything from following 20 to 40 responses to a particular stimulus.

As seen by the ant-colony example, a correctly chosen set of interacting rules for a group of identical generic entities can create self-organization and/or emergent

behavior (Bak 2013). However, a general algorithm to find a correct set of local rules to produce a target global behavior is not know (Hoekstra, Kroc, and Sloot 2010). Regardless of going in the opposite way is as easy as running a simulation using a proposed set of rules and checking if they yield the target behavior. In summary, the question, What set of rules yields this behavior? Is extremely difficult, yet asking, Does this set of particular rules yield this behavior? Is rather easy.

In CA the same dichotomy arises. Going from local rules to global behavior (local to global mapping) is as easy as to just perform the computations defined by the CA but going in the other direction (global mapping to local) is extremely difficult. This issue is known as the "inverse problem". Regardless of its difficulty, numerous attempts have been tried, with limited success. A common approach is the usage of optimization techniques like evolutionary algorithms or simulated annealing to find rules driving the system to desired attractor basins (Ganguly et al. 2003).

The semantics of CA made them readily available to model complex systems. In fact, Ilachinski mentions that *"CA are, fundamentally the simplest mathematical representations of a much broader class of complex systems."* (Ilachinski 2001). The modeling of a complex system using CA often proceeds in a bottom-up manner. Initially essential properties of the interacting parts are abstracted and then codified into the updating rules of CA cells. Secondly, the simulation is run in order to learn the *mesoscopic laws* that emerge from the individual interactions. With domain knowledge this process could be iterated, first proposing essential rules and see if they produce reasonable emergent behavior and then improving the rules from this feedback. Finally, the system could be enlarged to a gigantic simulation (mainly in space) to get the *macroscopic* final behavior.

However, Toffoli points out that this is not the best allocation of computational budget and recommends using the learned *mesoscopic laws* as input for higher-level analytical or numerical models (Hoekstra, Kroc, and Sloot 2010). To

give an example he imagines a CA that predicts the formation of water droplets from simple interactions of particles. Then if the computational resources are available the model could be escalated millions and millions of times to model fog, clouds, all the way up to global weather. However scaling in this way is not really necessary as once the bulk properties of a water droplet have been found they could easily be feed as numerical parameters of a higher-level model (e.g. differential equation), so in the end, the CA had helped us to get something like droplets per cubic meters or temperature an so forth. In Toffoli words *"A successful CA model is the seed of its own demise!"*, nonetheless at the end of the day the CA had helped us taming the complex system and clearly expressing the essential microscopic dynamics of the system.

2.6 Computing

An analogy between CA and conventional computers can be made. The initial configuration of a CA could be though as input data to be computed over by the CA rules, producing results several time steps ahead and displayed on whatever configuration reached by the lattice.

This analogy is not a coincidence at all and it is further exposed by the history of CA. In the early 1950s, von Neumann was trying to build a machine that not only should be self-replicating but also capable of universal computability. Von Neumann's endeavors were successful and produced the first two-dimensional automaton formally shown to be Turing-complete (Von Neumann, Burks, and others 1966). Twenty years later John Conway's "Game of Life" was introduced and later was also found to be computationally universal (Elwin, Conway, and Guy 1982)(Poundstone 2013). More recently 1-D CA "Rule 110" has been proved to be universal and is one of the simplest known systems with such property (Cook 2004).

The usual strategy to prove that a given CA is universal is to show its equivalence

with other systems known to be universal. Other strategy is to directly build on the lattice all the primitive elements of computing, namely *storage* (memory), *transmission* (internal clock and wires), and *processing* (AND, OR and NOT gates) (Ilachinski 2001). Once a given system supports these computational primitives building a universal machine becomes a clerical work of assembling modules. The "Game of Life" is proven to be universal in this fashion.

On the other hand possessing the same power as a conventional digital computer plays an important role on our mathematical ability to make predictions on the behavior of CA because all universal computers require resources in the same order of magnitude to process a particular algorithm thus, in general, a computational shortcut to the simulation of any universal CA does not exist (Toffoli 1977). This implies that even if an analytical expression for exactly capturing the evolution of a universal CA is obtained, evaluating such expression would take asymptotically the same time as just running the CA and observing its own evolution. Thus remarkably the most efficient way to characterize a universal CA is through its own simulation (Ilachinski 2001).

A different explanation of why, in general, there is not an analytical expression to predict, for any time step t the grid configuration of a universal CA, is by means of the *Halting Problem*, which is known to be *undeciadable* (Turing 1936). Thus if a CA is taken as a running program there is no general way of knowing if it would *halt*, with the meaning of *halting*, for this case, being that the CA would reach any configuration that signals the results of the computations.

Furthermore, locality being one of the main ingredients of CA imposes an almost independent update on each cell that is only influenced by its neighbors. As the execution of the program by the CA is being carried out individually by its cells the computations are being executed in a fully parallel manner. Consequently, simulations on CA allow for efficient parallel implementations of any real world system that can be codified into the CA formalism. For example, CA based machines CAMs (CA Machines) have been proposed by Toffoli and others (Toffoli

1984). A hardware implementation of a CAM was developed at MIT (Margolus 1995) that for the modeling of complex systems it could achieve a performance of several orders of magnitude higher than a traditional computer at a comparable cost.

2.7 Generalized Cellular Automata

Generalizations to the classical attributes of CA can be conceived (Ilachinski 2001) enabling extensions like:

- Asynchronous CA: Allows asynchronous updating of the CA.
- Coupled-map Lattices: Allows real valued cell states. These systems are simpler than partial differential equations but more complex than standard CA.
- Probabilistic CA: The rules are allowed to be stochastic, assigning probabilities to cell state transitions.
- Non-homogeneous CA: Updating functions are allowed to vary from cell to cell. A simple example is a CA with two rules distributed randomly throughout the lattice. On the other extreme case, simulations have been performed with a random assignment of all Boolean functions with small number of inputs (Kauffman 1984).
- Boolean Networks: A type of non-homogeneous CA with an emphasis on variable inputs per node (Alonso-Sanz 2011).
- Mobile CA: In this model some cells can move through the lattice. The mobile parts of the CA can be thought as robots or agents and their movement is dictated from an internal state that reflects the features of the local environment.
- Structurally Dynamic CA: Considers the possibility of evolving cell

arrangement. In standard CA the lattice is only a substrate for the ongoing computation, but what happens when this passivity is removed.

2.8 History

Precursor ideas about Cellular Automata can be traced back to 1946 cybernetics models of excitable media of Wiener and Rosenbluth (Weiner and Rosenblunth 1946), however, their usual agreed upon inception was when in 1948 John von Neumann following a suggestion from mathematician Stanislaw Ulam introduced CA to study self-replicating systems, particularly biological ones (Von Neumann and others 1951)(Von Neumann, Burks, and others 1966).

Von Neumann's basic idea was to build a lattice in $\mathbb{Z}^2$ capable of copying itself, to another location in $\mathbb{Z}^2$. The solution, in spite of being elaborate and involving 29 different cell states, was modular and intuitive. Since then more constructions capable of the same feat have been found with a lesser number of states (Codd 1968).

In the 1960s theoretical studies of CA were made, especially as instances of dynamical systems and their relation to the field of symbolic dynamics. A notable result from the epoch is the Curtis-Hedlund-Lyndon theorem (Hedlund 1969), which characterizes translation-invariant CA transformations.

In 1969 Konrad Zuse published the book *Calculating Space* (Zuse 1970) with the thesis that the universe is fundamentally discrete as a result of the computations of CA-like machinery. Likewise during 1960s computer scientist Alvy Ray Smith demonstrated that 1-D CA are capable of universal computation and showed equivalences between Moore and von Neumann neighborhoods, reducing the first to the second (Smith III 1971).

A key moment came with the invention of 2-D CA Game of Life. Pure mathematician J.H. Conway created "Life" as a solitaire and simulation type

game. To play "Life" a checkerboard was needed, then counters or chips were put on top of some squares. This represented an initial alive population of organisms and the initial configuration would evolve following reproduction and dying rules. The rules were tweaked by Conway to produce unpredictable and mesmerizing patterns. The game was made popular when was published as recreational mathematics by Martin Gardner in 1970 (Gardner 1970). Despite its name and interesting properties "Life" has little biological meaning and should be only interpreted as a metaphor (Ermentrout and Edelstein-Keshet 1993).

During the 80s the notoriety of CA was boosted to the current status as CA became quintessential examples of complex systems. The focus of the research was shifted towards CA as modeling tools. Is in this decade that the first CA conference was held at MIT (Ilachinski 2001) and that a seminal review article of Stephen Wolfram was published (Wolfram 1983).

Since then applications have been coming in a variety of domains. In the biological sciences models of excitable media, developmental biology, ecology, shell pattern formation, and immunology, to name a few, have been proposed (Ermentrout and Edelstein-Keshet 1993). CA can be applied in image processing for noise removal and border detection (Popovici and Popovici 2002). For physical systems, fluid and gas dynamics are well suited for CA modeling (Margolus 1984). Also, they have been proposed as a discrete approach to expressing physical laws (Vichniac 1984).

2.9 Forest Fire Models

Forest fire models are a type of *Probabilistic Cellular Automata*. They try to capture the dynamics and general patterns of tree clusters that emerge from an evolving forest subject to perturbations.

They trace their origins to statistical physics and are closely related to percolation

Table 2.2: Key events in the history of Cellular Automata. Table adapted from the book Cellular Automata A Discrete Universe (Ilachinski 2001).

Year	Researcher	Discovery
1936	Turing	Formalized the concept of computability.
1948	von Neumann	Introduced self-reproducing automata.
1950	Ulam	Realistic models for complex extended systems.
1966	Burks	Extended von Neumann's work.
1967	von Bertalanffy, et al	Applied System Theory to human systems.
1969	Zuse	Introduced the concept of "computing spaces".
1970	Conway	Introduced the CA "Game of Life".
1977	Toffoli	Applied CA to modeling physical laws.
1983	Wolfram	Authored a seminal review article about CA.
1984	Cowan, et al	The Santa Fe Institute is founded.
1987	Toffoli, Wolfram	First CA conference held at MIT.
1992	Varela, et al	First European conference on artificial life.

phenomena and dissipative structures. However, they have been proved a valuable tool for ecological and natural hazard sciences as simple but powerful modeling tools (Zinck, Johst, and Grimm 2010).

Forest fire models help to tackle questions like Will the tree population eventually dies out?, What is the general shape of tree clusters?, What is the shape of the boundary between the forest and the fire?

At first glance forest fire models seem similar to epidemiological cellular automata models though they place emphasis on finite population and the persistence of a pathology over time in contrast to the infinite forest population and the emphasis on the spatial extension of the fire.

They broadly have the following characteristics (Louis and Nardi 2018):

1. Cells of at least three types:

- Non-burning tree
- Burning tree
- No tree (empty)

2. A rule for fire initiation:

- A starting configuration with fire cells, usually randomly chosen fire positions.
- Accident simulation, like with a small probability self-ignition of a tree cell.
- Space-time distributed ignition instances (e.g. Poisson distributed).

3. A rule for fire propagation. It involves a stochastic rule for fire spreading between neighborhoods that can be based on actual terrain conditions.

2.10 The Drossel and Schwabl forest fire model

The forest fire model that will be used through this document is the Drossel and Schwabl model (DSM) (Drossel and Schwabl 1992).

DSM was born from research on statistical physics about phase transitions and self-organized criticality (Bak, Chen, and Tang 1990)(Drossel and Schwabl 1992). For this reason, the CA was not intended as a modeling tool for real wildfires and was only a metaphor.

Nevertheless, data from real wildfires was compared against DSM predictions with the, no so surprising, observation that the model did not perfectly match real world datasets, as it was built with the only concern of generating fire sizes following a power law. DSM was overestimating the frequency of large fires (Millington, Perry, and Malamud 2006). Even though its origins and pitfalls DMS is still valuable, as it has a strong advantage against other wildfire models due to its simplicity and analytical tractability (Zinck, Johst, and Grimm 2010). Likewise, it provides a set of starting assumptions that can be augmented to the required complexity along with the usually seen trade-off between increasing a model's predictive capabilities and its generalizing power.

Consequently, success has been achieved using this simple model, for example, fire shape patterns have been obtained that closely resemble actual wildfires (Zinck

and Grimm 2008)(Zinck, Johst, and Grimm 2010).

The Drossel and Schwabl model consists of a lattice in $\mathbb{Z}^2$ populated with three types of cells: 0 (no tree), 1 (burning tree), and 2 (non-burning tree). All cells are synchronously updated according to the following rules: For each state $\sigma_k(n)$ at site k and time step n.

- A burning tree is consumed at next time step:

$$\sigma_k(n) = 1 \mapsto \sigma_k(n+1) = 0 \quad \text{With probability 1}$$

- A new tree grows from an empty position k, dictated by parameter $p \in [0,1]$:

$$\sigma_k(n) = 0 \mapsto \sigma_k(n+1) = 2 \quad \text{With probability } p$$

- The fire is propagated through the vicinity or a lightning event occurs, which ignites a tree and is tuned by $f \in [0,1]$:

$$\sigma_k(n) = 2 \mapsto \sigma_k(n+1) = 1$$

$$\text{With probability} \begin{cases} 1, & \text{if at least one neighboring tree is burning} \\ f, & \text{if no neighboring tree is burning} \end{cases}$$

3

Reinforcement Learning

3.1 Machine Learning

Machine Learning is a subfield of Computer Science and Artificial Intelligence. It aims at creating entities capable of learning from examples. What distinguishes it from other AI approaches is the special emphasis on avoiding explicit programming for the task at hand, instead relying of only on examples (data) to solve it and from there generalizing to previously unseen cases. So at the heart of this approach lies data. In the classical machine learning setting the algorithm is provided with a set of questions and answers. Then after a period of computations over the pairs questions-answers, the algorithm is presented with new questions, some of them never seen before, that must be answered well

enough.

The formalization of this setting leads to an approach known as Supervised Learning, where the questions-answers pairs are codified into mathematical objects, usually numeric vectors for questions and real numbers or categories for answers and the proposed solution comes in the form of a function that maps questions to answers. The name supervised comes from the fact that the algorithm is provided with the answers to the questions so figuratively it is being supervised by a teacher.

Deviation from this classical setting leads to the other broad categories of machine learning. Unsupervised Learning algorithms are provided with just data (questions and not answers) and aim to uncover some structure in such data. Semisupervised Learning is halfway between Unsupervised and Supervised methods as only some answers are given, thus the algorithm must first find some structure on the questions (Unsupervised) to extend the answers to similar questions and then perform Supervised Learning. Finally, Reinforcement Learning algorithms get data and answers, but answers have only a grade on how favorable they are. The data and grades are obtained through interaction with an environment and the task here is to find answers that maximize the obtained grades.

3.2 Reinforcement Learning

The contents of this section have mostly been adapted from Sutton and Barto's Introduction to Reinforcement Learning book (Sutton and Barto 2018).

Reinforcement Learning aims to develop algorithms that can satisfactorily inform an agent which decisions to make to achieve a goal. The world in which the agent acts is called the environment. The decisions made by the agent could affect the environment and hopefully drive it closer to the goal. A signal of how well the

agent is performing is received at each decision step. The signal is called reward and it is an abstraction to represent great or poor sequences of actions. Plenty of this reward signal mean a successful agent closer to its objective.

The Reinforcement Learning Problem could be formalized with Markov Decision Processes (MDPs). The MDP framework models the interaction between an agent and an environment. The agent takes actions and the environment responds with observations and a reward signal, this process is repeated until termination or forever. The non-terminating case is known as a continuing task and the terminating one as an episodic task.

So at each time step $t = 0, 1, 2, 3, ..., T$ the agent using the information of state $S_t \in S$ of the environment must choose an action $A_t \in \mathcal{A}(S_t)$ from the available ones, then in the next time step $t + 1$ the environment transits to a new state $S_{t+1} \in \mathcal{S}$ and returns a reward $R_t \in \mathcal{R} \subset \mathbb{R}$. The dynamics between the agent and environment produces a trajectory of states, actions and rewards indexed by time.

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, ...$$

Of special importance is the case of a finite MDP, in which the cardinality of states, actions and rewards ($\mathcal{S}$, $\mathcal{A}$ and $\mathcal{R}$) is finite. In a finite MDP the random variables S_t, R_t have well defined discrete probability distributions that depend only on the previous action A_{t-1} and state S_{t-1}. This is known as the Markov property:

$$p(s', r|s, a) \doteq \Pr(S_t = s',\ R_t = r \mid S_{t-1} = s,\ A_{t-1} = a)$$

This $p : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ function is known as the *dynamics of the MDP*. The "|" (bar) symbol instead of a "," (comma) in the function arguments is just a reminder that p is calculating a conditional probability given s and a. From here on the assumed MDP would be a finite one.

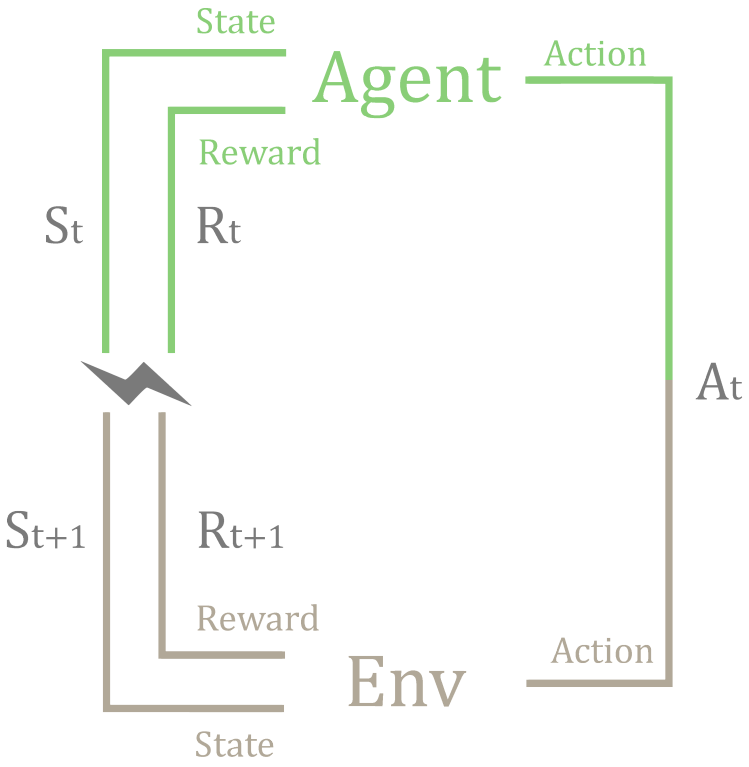


Figure 3.1: Diagram of the *Agent-Environment system* formalized as a MDP. At time step t the *"Agent"* with knowledge of observation S_t and reward R_t performs action A_t to which the *"Environment"* responds with a new observation S_{t+1} and reward R_{t+1}, the cycle is iterated. Figure adapted from (Sutton and Barto 2018).

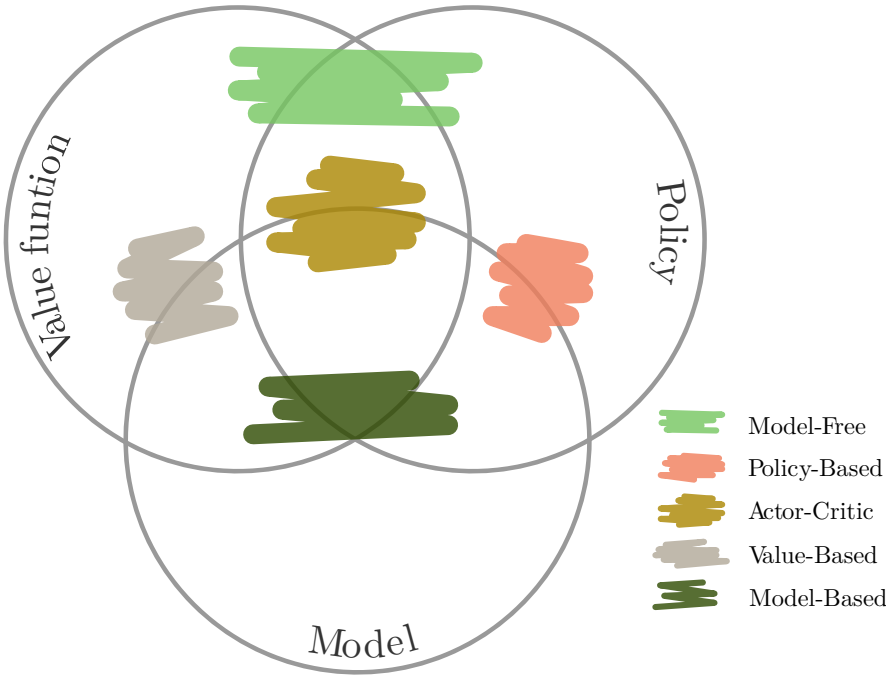


Figure 3.2: Broad classification of Reinforcement Learning algorithms. Adapted from David Silver's Lecture 1.

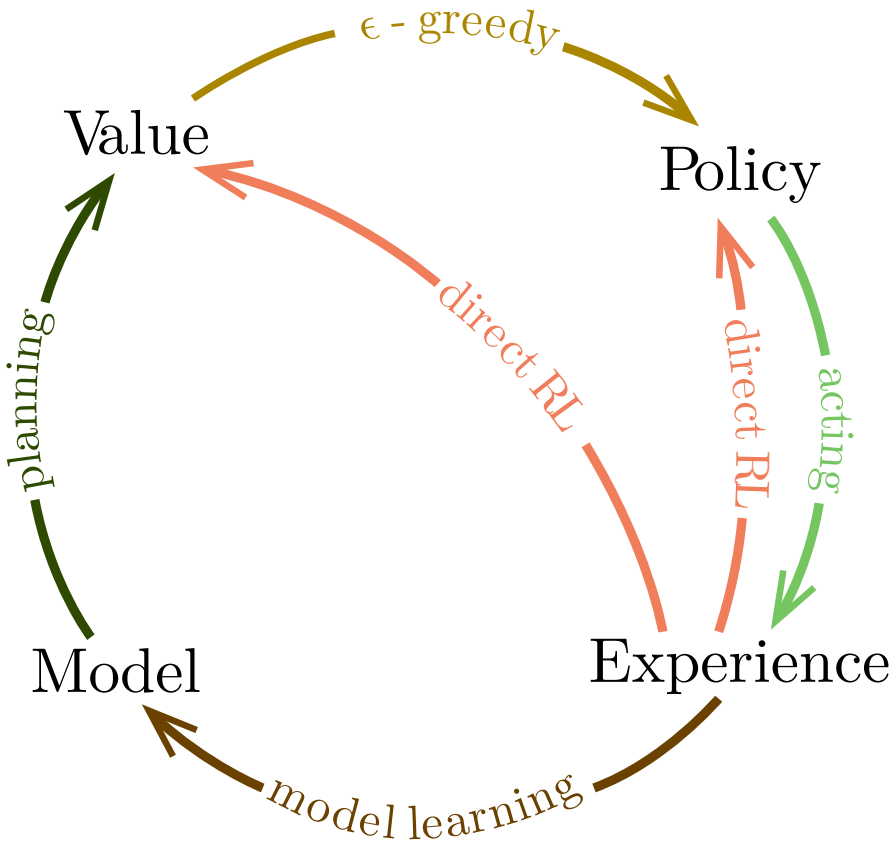


Figure 3.3: Broad relations between different types of strategies to solve the Reinforcement Learning Problem. Adapted from David Silver's Lecture 1.

3.2.1 Goals and Return

The reward signal must guide the agent into achieving a goal. The hint that the sequence of rewards provides is such that if the agent maximizes the total amount of received rewards, the goal would be reached. Thus the agent objective turns to, not maximizing immediate reward, but instead cumulative reward in the long run. This informal idea is known as the *reward hypothesis* and it is stated as follows:

> "All of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal." (Sutton and Barto 2018)

In practice, when designing a RL task rewards can be thought of being of two types:

1. **Sparse:** Rewards are only given when the goal is reached.
2. **Shaped:** Rewards are also given when reaching subgoals.

Sparse rewards are preferred since the reward signal should only communicate the *what* we want to accomplish and not the *how*. Similarly the RL framework provides better ways of inserting domain knowledge into the system, like in the initial policy or the initial value function. So it is not surprising that shaped rewards can bias learning when the subgoals are not aligned with the final goal or when accomplishing them is given more importance by the agent. Nonetheless, if well designed, shaped rewards can accelerate learning and are often much easier to learn from.

To formalize the previous discussion, the random variable G_t *return* is defined, for the time step t in an episodic task with final step T:

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + ... + R_T$$

To handle the continuing task case, where $T = \infty$, the notion of *return* can be

extended by the inclusion of the discounting parameter $\gamma \in [0 - 1]$.

$$G_t \doteq \sum_{k=0}^{\infty} \gamma^k R_{t+1+k}$$

This *return* formulation converges to a finite value if $\gamma < 1$ and the reward sequence R_k is bounded. The parameter γ represents how much weight far away rewards must be given, specifically a discount of γ^{k-1} for the reward R_{t+k}, k steps into the future. If $\gamma = 0$, the agent is "myopic" and is only concerned with maximizing immediate rewards. On the other hand as γ approaches 1, more and more distant rewards are being taken into account, thus the agent becomes more farsighted.

A generalization of *return* that combines both, episodic and continuing setting is as follows:

$$G_t \doteq \sum_{k=0}^{T} \gamma^k R_{t+1+k}$$

Where T is turned into a parameter of future steps to take into account by the convention that if the episode terminates, all the futures rewards would be 0. So in this formulation, T could be unbounded ($T = \infty$) or $\gamma = 1$, but not both at the same time.

The return G_t at time t follows and important recursive relation:

$$G_t = R_{t+1} + \gamma G_{t+1}$$

3.2.2 Policy

A *policy* π is a mapping from states to probabilities of selecting each particular action. Formally a function of 2 arguments, where the symbol "|" is just to remind us that the action depends on the state.

$$\pi(a|s) \doteq \Pr[A_t = a|S_t = s], \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$$

Policy codifies the behavior of the agent in the face of environment states. A well performing policy is one that generates behavior that gets the agent closer to its goal. An optimal policy is one that accomplishes the goal. Reinforcement Learning algorithms must find policies that maximize the *expected return* since by the *reward hypothesis* this is the same as accomplishing the goal. In other words, the problem that RL is trying to solve is to find a policy π that maximizes the *expected return* $\max_{\pi} \mathbb{E}_{\pi} G_t$.

3.2.3 Value Functions

Value functions capture the idea of how desirable each state is. They are always discussed in the context of a particular policy π. A whole family of methods uses them to solve the RL problem.

The *value function* $v_\pi(s)$ of a state s under a policy π is defined as:

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t | S_t = s], \quad \forall s \in \mathcal{S}$$

$$v_\pi(s) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+1+k} \middle| S_t = s\right], \quad \forall s \in \mathcal{S}$$

Similarly for action-state pairs ($a \in \mathcal{A}$, $s \in \mathcal{S}$) the *action-value* function for policy π, $q_\pi(s, a)$ is defined:

$$q_\pi(s, a) \doteq \mathbb{E}_\pi[G_t | S_t = s, A_t = a], \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$$

$$q_\pi(s, a) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+1+k} \middle| S_t = s, A_t = a\right], \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$$

The function q_π represents the value of forcing the agent into taking action a and then following the policy π.

The functions v_π and q_π can be estimated from interaction with the environment. The average, per state s, of rewards obtained following the policy π will converge

to $v_\pi(s)$. Similarly, if keeping track of state-action pairs, $q_\pi(s, a)$ can be estimated by the same method.

A fundamental property of value functions is that they satisfy a recursive relationship known as the *Bellman's Equation*:

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')], \quad \forall s \in \mathcal{S}$$

If the dynamics of the environment are completely known, the *Bellman's Equation* can be used to obtain the value function v_π for all states since a system of linear $|\mathcal{S}|$ equations with $|\mathcal{S}|$ unknowns arises.

Similarly the *Bellman's Equation* for value-action functions is:

$$q_\pi(s, a) = \sum_{s',r} p(s', r|s, a)[r + \gamma \sum_{a'} \pi(s'|a')\ q_\pi(s', a')], \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$$

3.2.4 Optimal Policies

Solving the Reinforcement Learning problem can be seen as finding a policy that achieves a high expected return. For finite MDPs the concept of an *optimal policy* can be defined. Value functions can be used to order policies in the following way: A policy π is said to be better than other policy π' if and only if $v_\pi(s) \geq v_{\pi'}(s)$ for all $s \in \mathcal{S}$. Thus value functions define a partial ordering over policies, that is to say:

$$\pi \geq \pi' \text{ if and only if } v_\pi(s) \geq v_{\pi'}(s), \quad \forall s \in \mathcal{S}$$

There is always at least one policy that is better than or equal to all others. The policies with this property are named *optimal policies* and are denoted by π_*. The value function under an *optimal policy* (*optimal state-value function*) is written v_* and can be expressed as:

$$v_*(s) \doteq \max_\pi v_\pi(s), \quad \forall s \in \mathcal{S}$$

In the same fashion, an *optimal action-value function* is defined as:

$$q_*(s,a) \doteq \max_\pi q_\pi(s,a), \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$$

Bellman's equation for value functions can be rewritten for the *optimal state-value function*. Intuitively it represents the agent perfectly responding with the best action to any situation of the environment. For all states $s \in \mathcal{S}$:

$$v_*(s) = \max_{a \in \mathcal{A}(s)} q_*(s,a)$$

$$v_*(s) = \max_a \sum_{s',r} p(s',r|s,a)[r + \gamma v_*(s')]$$

The above equalities are also present on they *action-value* $q_*(s,a)$ form, for all states $s \in \mathcal{S}$ and all actions $a \in \mathcal{A}$:

$$q_*(s,a) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \;\middle|\; S_t = s, A_t = a \right]$$

$$q_*(s,a) = \sum_{s',r} p(s',r|s,a)\left[r + \gamma \max_{a'} q_*(s',a') \right]$$

For finite MDPs *Bellman's optimality equation*, either for v_* or q_*, has a unique solution. If v_* or q_* are known for all states and actions, it is possible to solve the Reinforcement Learning problem and derive optimal behavior from them.

$$a = \operatorname*{argmax}_{a'} \sum_{s',r} p(s',r|s,a')[r + \gamma\ v_*(s')]$$

$$a = \operatorname*{argmax}_{a'} q_*(s,a')$$

3.3 Deep Q Networks

The *optimal action-value function* $q_*(s,a)$ also referred as $Q^*(s,a)$ is the maximum expected return achievable by selecting action a and then following an *optimal policy*. The *Bellman equation* for $Q^*(s,a)$ is:

$$Q^*(s,a) = \mathbb{E}\left[r + \gamma \max_{a'} Q^*(s',a') \;\middle|\; S_t = s, A_t = a\right]$$

If Q^* is known the RL problem can be solved, thus estimating Q^* is the basic idea behind many RL algorithms. The *Bellman equation* can be used as an iterative update: $Q(s,a)_{i+1} \leftarrow \mathbb{E}\left[r + \gamma \max_{a'} Q_i(s',a')|s,a\right]$, this *value iteration* technique converges to the *optimal action-value function* (Sutton and Barto 2018), $Q_i \rightarrow Q^*$ as $i \rightarrow \infty$. In practice *value iteration* must be truncated at some point, with the conventional practice of stopping after, under a threshold, no significative update change is obtained. For many tasks this approach is unsustainable due to the high dimensionality of states and the vast number of state-action pairs.

Model-free, value-function approximation RL tries to tackle the intractability of directly estimating $Q(s,a)$ for a high number of state-action pairs with a parameterized version $Q(s,a;\theta) \approx Q^*(s,a)$. In the case of Deep Reinforcement Learning (DRL) Q is approximated by a ANN with parameters θ.

A naive, nonetheless intuitive, approach is to directly translate $Q(s,a) : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ to a network architecture that predicts a Q-value for each state-action pair. However this approach incurs on $|\mathcal{A}|$ forward computations of the ANN since in order to get the next action following the *greedy policy*, we need to check for all values of the available actions $\forall a \in \mathcal{A}(S_t) : Q(s,a;\theta)$. One of the key insights of the Deep Q-Networks (DQN) algorithm (Mnih et al. 2013)(Mnih et al. 2015) was to change to an architecture of the form $Q_{network} : \mathcal{S} \rightarrow \mathbb{R}^{|\mathcal{A}|}$, predicting all Q-values in a single forward pass and also with the extra justification that perhaps the model would learn features about state s that can be reused to predict all Q-values for a given s. This architecture is known as the $Q_{network}$.

A $Q_{network}$ capable of predicting the values of the actions would solve the Reinforcement Learning problem (the agent would act greedily over the Q-values). Unluckily training a $Q_{network}$ could be tricky as theoretical convergence results are only available for linear approximators (Tsitsiklis and Van Roy 1997). Additionally, *Supervised Learning* techniques assume that the data was generated from the same distribution and its instances are *i.i.d*, both assumptions are violated because the training distribution is constantly changing as the implicit policy defined by the network is changing and the generated observations are highly correlated. To combat the instability that arises DQN uses an *experience replay* (*memory buffer*) and a *target network*.

For the Q-values update through the *Bellman's equation* a tuple of *state*, *action*, *reward* and *next reached state* (s, a, r, s') is required. Interaction with the environment supplies such tuples, however the environment must be explored in such a way that allows for favorable training. A straightforward manner to obtain the observation tuples is by an ϵ-greedy policy, that is, with a small ϵ a random action is selected, otherwise $a \leftarrow \underset{a'}{\text{argmax}}\, Q(s, a'; \theta)$, thus Q generates policy and it is fittingly called the *policy network*. The observation tuples are being stored on the *experience replay*. The *experience replay* is just a memory with fixed size to save past observations, when the memory is filled, the oldest observation is eliminated to make space for the new incoming observation. The training of Q is done by sampling a minibatch of past observations from the *experience replay*, this to alleviate the problem of highly correlated observations that are being produced by the policy network.

The target network $\hat{Q}_{network}$ is a mere copy of the $Q_{network}$, with the sole purpose of estimating the target update Q, for reward r_i and state s'_i:

$$y_i = r_i + \gamma \max_{a'} \hat{Q}(s'_i, a'; \theta)$$

The target network $\hat{Q}_{network}$ parameters are kept constant for a C number of

epochs, thus stabilizing the bootstrapping (*Bellman's equation* update). To illustrate how just using the policy network ($Q_{network}$) for learning and updating can lead to instability we can think of two relatively similar states s_1 and s_2. Suppose that s_1 is used to estimate the value of s_2, when updating the parameters to minimize the error on s_2 inadvertently we would be changing s_1 estimated Q-value, as both states share the same parameters and are similar, so if s_1 value is required again for bootstrapping the estimation would be deteriorated, this effect may compound and cause the whole network estimation to diverge. Decoupling estimation and acting is the main idea behind the use of two networks, each C number of epochs both networks are synchronized. Consequently the loss function of DQN is:

$$MSE(\theta) = \mathbb{E}\left[\left(y_i - Q(s_i, a_i; \theta)\right)^2\right]$$

The error is then minimized through any general optimization algorithm, in the case of DQN typically by any variation of gradient descent.

For an episodic task termination states must be handled. The needed adjustments are a change on the error calculation to include the termination case ($y_i = r_i$, if termination occurred at iteration i) and an outer loop for episodes iteration or without it, through the use of the Open AI Gym framework.

The complete DQN algorithm can be found at Appendix A 5.1.3.

3.4 The Atari benchmark

The *Arcade Learning Environment* (ALE) was introduced in 2013 (Bellemare et al. 2013). It was quickly adopted as a benchmark by the RL community. ALE consists on a wrapper around the Stella emulator[1] for the Atari 2600 console. This wrapper provides an interface to more than 50 Atari games. ALE was used by

[1] https://stella-emu.github.io/

Google's DeepMind team in the DQN paper (Mnih et al. 2015), arguably setting the current trend of deep reinforcement learning. Serving as a common benchmark ALE has been successfully used to develop new algorithms that later were applied to other domains, such as robotics (Risi and Preuss 2020). As 2020 a single agent is capable of outperforming standard human play for 57 games Atari games (Badia et al. 2020), however the end of the ALE benchmark is nowhere insight as a myriad of challenges still persists on Reinforcement Learning methodologies.

The desirable characteristics of Atari games for RL evaluation are their variety of tasks, their short episodic nature and scores that can be, almost directly, interpreted as rewards. Also, each game was created independently thus reducing the bias that would be introduced by a single party designing all the evaluation tasks. Agents interact with ALE through a set of 18 actions derived from 6 simple actions, namely four directions (up, down, left, right) and fire and NO-OP actions. The final 18 actions are combinations of the basic actions. The state information is raw pixel data with dimensions 210×160, with each pixel on a 7-bit color scale (0-127). Additionally, the whole console RAM of 128 bytes can be used as input to the learning algorithms. ALE can be used as generative model in the sense that previous states, if saved, can be restored.

Some criticism towards ALE is the scarce stochasticity in the environments (Bellemare et al. 2015)(Machado et al. 2018) and the black box emulation of the games. Recent efforts are trying to address those issues, like TOYBOX (Foley et al. 2018) which implements from scratch Atari games with added functionality to monitor meaningful representations of game internal states.

4

Experiments

4.1 Materials and Methods

4.1.1 Environment

A general tool for creating RL tasks was design (*Forest Fire Environment Maker (FFEM)*). The tool generates environments based on the forest fire CA, specifically the Drossel and Schwabl model (DSM) (Drossel and Schwabl 1992). We developed the tool from an initial idea of a "helicopter" flying over the symbolic wildfire, represented by the DMS, trying to extinguish the fire by means of allowing the "helicopter" to change fire cells to empty cells. During the implementation the environment was naturally generalized. The basic idea is to have an agent on top

of the CA lattice. The agent has its own state and parameters. Particularly it has a specified position at one of the nodes of the CA, where it can act by changing the cell on that position to another of its cell states. The agent navigates the CA changing cells states. When the agent reads its next instruction it moves to a new cell (or stays in place) and then affects, if applicable, the destination. After some agent movements the CA is updated and the full process is repeated.

The ability of allowing an agent to change the CA configurations express the attempt to drive the CA to a desired global state. In the RL framework, this is done by giving a reward to the agent at each emitted action. This was done by merely counting all cell states indicating unwanted behavior and then multiplying by some weight (usually negative). This was generalized into getting the frequency of any cell type and then multiplying by a weight to generate a score per cell type. Then all the scores were combined into a global score. In summary the used reward function was the dot product of the cell-states frequencies by some weights, with the semantics that weights for desirable states are greater than the unwanted ones.

The particular environment that was used for this thesis is one that has the semantics of a "helicopter" trying to extinguish a forest fire. This is represented by an agent influencing the dynamics of a DSM and is as follows: The first component is a DSM with a lattice of size 3x3 with parameters $p = 0.33$ (tree growth probability) and $f = 0.066$ (tree spontaneous burning probability). The used boundary conditions were set to invariant using an extra exogenous cell type that it is not involved in any CA dynamics, we just called them "rocks". The starting lattice configuration was made via random sampling by cell with probability of choosing "tree" of 0.75 and "fire" of 0.25. The second component, the agent ("helicopter") has a starting position in the middle of the lattice (*2nd row and column*), from where it can move in any direction of a Moore's neighborhood, thus the agent can choose from 9 actions (*"Left-Up", "Up", "Right-Up", "Right", "Stay", "Left", "Left-Down", "Down", "Right-Down"*), then after arriving at its new destination the helicopter extinguishes the fire, represented by changing, if

applicable, a "fire" cell to "empty". After some movements of the "helicopter" the forest fire CA is updated following the dynamics dictated by DSM. The specific sequence of agent movements and CA updates is: *move, update, move, move, update, move, move, update, move, move, …* (see 4.1). In accordance to the RL paradigm, at each performed action the environment must respond with a reward signal, in our setting this was $+1$ per "tree" cell and -1 per "fire" cell, in other words a balance between trees and fires. Finally, the RL task was continuous.

We selected the parameters of the DMS by seeking a separation between the scales of f and p ($f \ll p$), since a steady state is reached when $\lim \frac{f}{p} \to 0$ (Drossel and Schwabl 1992) and we wanted to see if the agent would significantly alter the dynamics of the CA. In our case $\frac{f}{p} = 0.2$. Alongside, empirical tuning was conducted from running the environment with a random policy to select the exact values of f and p that would make the grid relatively active at each step. It is important to note that the steady state only appears on large grids, as smaller ones are subject to finite size effects (Drossel and Schwabl 1992), thus we paid more attention to the empirical second criterion. Nevertheless balancing the two parameters would be important in future grid scaling experiments.

Moreover, the FFEM tool adds two new cell types. A "rock" type that does not interact with anything and is used as a barrier to the fire dynamics, it is always constant. A "lake" type that behaves exactly like a "rock" but it marks special positions in the grid that can alter the "helicopter" internal state. For example an idea (yet to be implemented) is to make the "power" of the "helicopter" limited so it would need to use it "wisely" and then would go to a "lake" to recharge it, akin to refilling its tank to keep fighting the wildfire. Also, the FFEM tool generalizes the "powers" of the "helicopter" allowing for the exchange of any cell type for another, this lets for the creation of environments with a diversity of semantics like: An environment where the agent has to put barriers or one where it has to strategically deforest to prevent the spread of fire or even one where it has to completely reorganize the CA lattice. Even more functionality was included

like a deterministic mode, generalized reward functions and, several termination conditions, to check all the FFEM features consult: *https://github.com/elbecerrasoto/gym-forest-fire*.

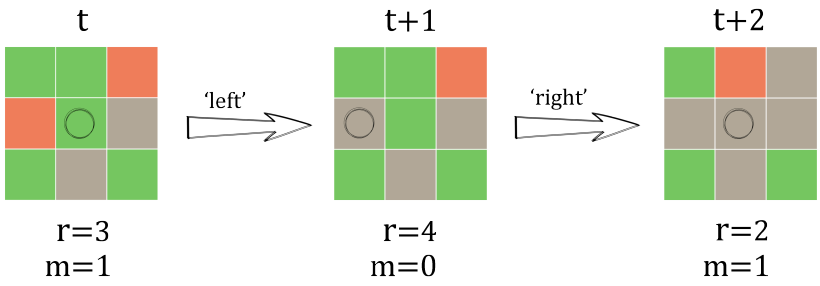


Figure 4.1: Illustration of two transitions in the proposed environment. On time t the helicopter is at 2nd row and column, then it moves to the left, on arriving at its new position it changes the "fire" cell to "empty". Then, at the next time step, it returns to its original position, however the CA was updated before its arrival, anyway the "fire" cell at the middle (not shown) is promptly eliminated and replaced by an "empty" cell. The parameter m is an internal state of the environment, that is decreased by 1 at each step, when it reaches 0 the CA is updated at the next step, before the action takes place, and then m is restored to it max value ($m = 1$). The current reward r is $+1$ per "tree" and -1 per "fire".

4.1.2 Code Availability

The source code for this thesis is openly available at:

1. *https://github.com/elbecerrasoto/gym-forest-fire* for the RL environments.
2. *https://github.com/elbecerrasoto/CARL* for the DQN implementations.

4.1.3 Deep Q Networks

We tried to solve the proposed environment using model-free, value-function approximation RL. This approach is justified by the lack of general analytical models for CA dynamics (Ilachinski 2001) and their combinatorial explosion of

states. For illustration in our small 3x3 environment, the combinatorics are 3^9 grid states multiplied by 3^2 "helicopter" positions and 2 internal states for the agent-environment synchronization, giving a total of $2 \times 3^2 \times 3^9 = 3.54294 \times 10^5$, a still manageable quantity by tabular RL standards, however merely scaling to a 16x16 grid leads to an untenable $\approx 10^{124}$ states.

DQN (Mnih et al. 2013)(Mnih et al. 2015) meets the previously stated conditions and is well suited for tasks with a discrete set of actions and when sampling from the environment is low-cost. We implemented DQN with n-step unrolling of the Bellman optimality equation for Q values (Sutton 1988). The implementation was made in Python 3.x, the *lingua franca* of AI and RL. The non-linear Q function approximator was an ANN. We implemented the *experience replay* memory as a *double-ended queue* (deque) data structure, which behaves like a list but allows for efficient "appends" and "pops" from either side ($\mathcal{O}(1)$). (see *https://docs.python.org/3.8/library/collections.html?highlight=collections#collections.deque*). Thus for a fixed size of an *experience replay*, incoming observations are appended to one end and old observations are "popped" from the other.

A high level explanation of our DQN implementation (see Apendix 1) is as follows: The logic of the e-greedy policy was abstracted into an "Agent class" which also holds a memory of past observations (*experience replay*), during the training loop the agent performs an action and advances the environment one step, saving the transition tuple (s, a, r, s') into its memory. Then the agent samples a minibatch of transitions from the memory buffer and feeds it to the ANN model. A forward prediction of Q values for all 9 actions is computed followed by a backward optimization step. These sequence of agent acts and ANN train steps is iterated during T steps. Each C training steps the weights of the policy and target networks are synchronized.

For the full code details refer to: *https://github.com/elbecerrasoto/CARL*

4.1.4 Preprocessing

The FFEM tool generates environments that follow the guidelines of the *Open AI Gym API* (OAGA) (Brockman et al. 2016). The API specifies that the observation should be a numerical vector, that is commonly returned as a *numpy* n-dimensional array (Walt, Colbert, and Varoquaux 2011). In our implementation we instead returned a tuple of three *numpy arrays* representing: the grid data (cell states), the position of the helicopter and the remaining moves (m) to CA updating. We made the knowledge of m available to the agent to guarantee the Markov property. A more difficult RL task can be made by not knowing m (just dropping its value from the tuple), thus shifting the system to a *Partially Observable Markov Decision Process* (POMDP), where the hidden state m should be inferred by the agent. Strictly speaking our implementation is departing from the *Open AI Gym API* observation specification, however it does it in favor of facilitating the data processing and if purity is insisted on, merely concatenating the data and reshaping it into a single *numpy array* would fix it.

The input to the ANN models was a one-hot encoding of the CA grid concatenated with the "helicopter" position and the m remaining moves parameter. This was supported by a FFEM feature to indicate the output format of the CA lattice, that can be returned in plain numeric representation, one-hot encoding or by channels as the CA lattice can be interpreted as an image.

4.1.5 ANN Architectures

Three different $Q_{network}$ architectures were used. The models were build using the *Pytorch* library (Paszke et al. 2019). They were Multilayer Perceptrons (MLPs) with four or three hidden layers and ReLU activations. The weights and biases were initialized under the *Pytorch* defaults (*Kaiming uniform* (He et al. 2015)). The tensors transformations of the layers are described below:

- Architecture 1 (A1):

  - Number of learnable parameters: 7.4313×10^4

$$(30 \times 1) \rightarrow (128 \times 1) \rightarrow (256 \times 1) \rightarrow (128 \times 1) \rightarrow (32 \times 1) \rightarrow (9 \times 1)$$

- Architecture 2 (A2):
  - Number of learnable parameters: 4.9673×10^4

$$(30 \times 1) \rightarrow (256 \times 1) \rightarrow (128 \times 1) \rightarrow (64 \times 1) \rightarrow (9 \times 1)$$

- Architecture 3 (A3):
  - Number of learnable parameters: 2.87881×10^5

$$(30 \times 1) \rightarrow (256 \times 1) \rightarrow (512 \times 1) \rightarrow (256 \times 1) \rightarrow (64 \times 1) \rightarrow (9 \times 1)$$

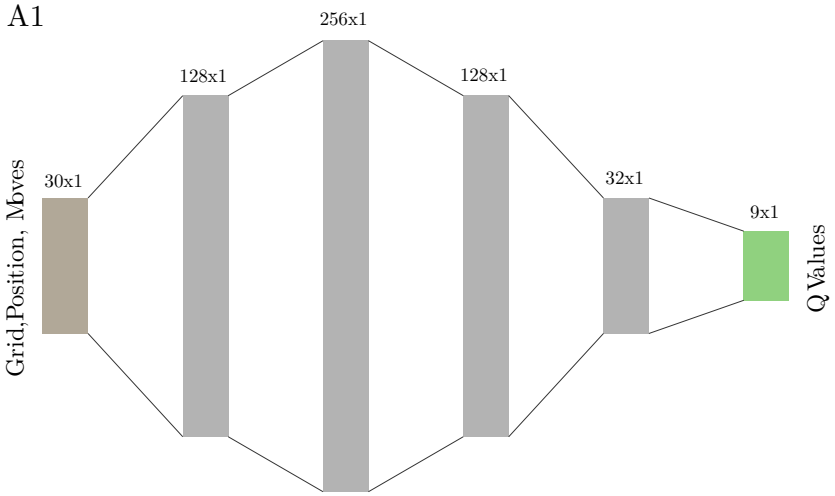


Figure 4.2: Diagram of Architecture 1 (A1).

4.1.6 Training Details

Training DRL algorithms is difficult. One of the reasons is that they are quite sensitive to hyperparameters (Irpan 2018). The improvement of training stability is a substantial part of DRL research. An advice to train RL algorithms (see *ht tp://rll.berkeley.edu/deeprlcourse/docs/nuts-and-bolts.pdf*) is to start with small

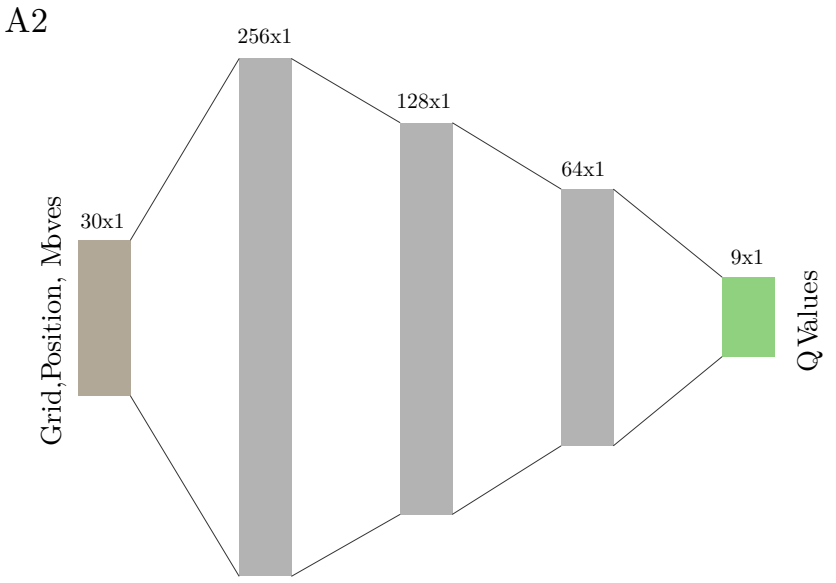


Figure 4.3: Diagram of Architecture 2 (A2).

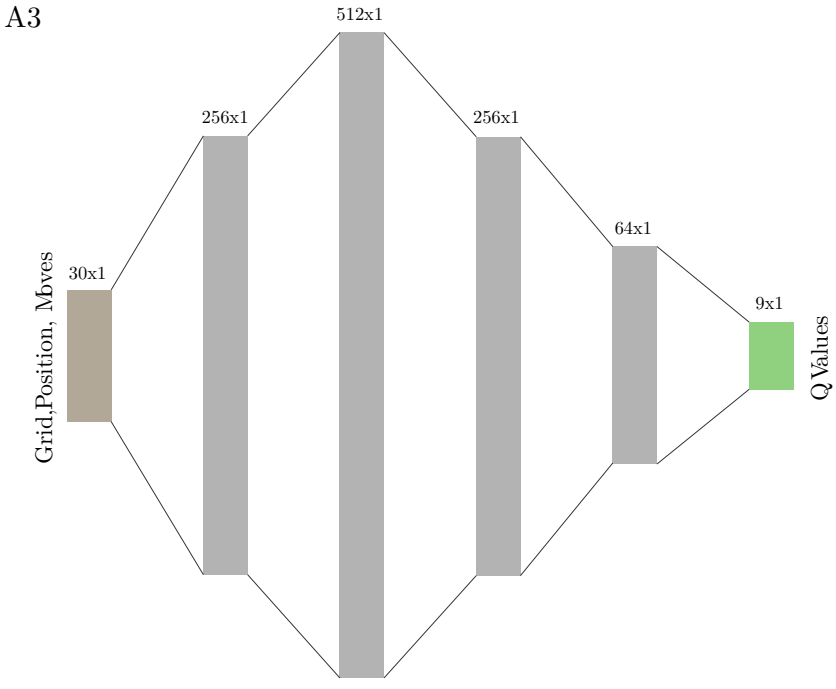


Figure 4.4: Diagram of Architecture 3 (A3).

instances of the problem and then proceed with a hyperparameter search to try to detect *"life signals"* (a successful DRL run), if despite this the DRL algorithm is not learning anything, the difficulty of the task should be decreased until *"life signals"* are detected. This training strategy was followed. Consequently, profiling runs were made with grid configurations of 1×9, 3×3, 5×5 and 8×8 (data not shown). We decided to settle on the 3×3 grid, to further extend the hyperparameter search, as bigger grid sizes were not showing *"life signals"*.

Nine experiments were run using the described environment, between experiments some hyperparameters were allowed to vary, namely the type of initial exploration, the unrolling of the Bellman's update, the network architecture, learning rate, batch size, and γ discount parameter. The overall values for the hyperparameters were taken from (Mnih et al. 2015) and then informally scaled to observe its effects on performance. We did not perform a systematic grid search due to the high computational cost. The hyperparameters that were varied are shown in Table 4.2, while the constant hyperparameters can be seen in Table 4.1.

Each experiment was run on *Google Colaboratory* cloud service using their provided GPUs and computing resources. Their run for 1,200,000 epochs each, divided in 400,000 exploration and 800,000 of exploiting epochs. Two exploration schemes were used, for some models a linear annealing of ϵ from 1.0 to 0.10 and for others a simple heuristic that moves the helicopter to any fire cell on the neighborhood, choosing randomly when more than one fire cell was present.

The employed optimization algorithm was Adam (Kingma and Ba 2014) with default *Pytorch* values, with the exception of *learning rate* that was set manually per experiment (see Table 4.2). Each transition tuple in the *experience replay* only estimates a single $Q(s, a)$ value, however the output of the $Q_{network}$ is a vector of Q values for all actions, so during optimization steps, for each example in the minibatch, only the weights involved in the calculation of a single $Q(s, a)$ must be taken into account for the updating of the weights.

Table 4.1: Constant DQN hyperparameters values across the 9 experiments.

Hyperparameter	Value
Total Training Steps	1,200,000
Exploration Steps	400,000
Exploitation Steps	800,000
Policy and Target Networks Synchronization	10,000
Experience Replay Size	200,000

4.1.7 Evaluation Procedure

The trained agents were evaluated by playing 100,000 steps in the environment following ϵ-greedy policies (ϵ of 0.00, 0.02, 0.05, 0.10) over the learned Q values. The *return*, *sample mean* and *sample standard deviation* of all evaluation *rewards* per run were computed. To provide a comparison baseline the same statistics were calculated for two other agents, a heuristic and a random one. The heuristic was the previously described of following fire cells in a 1-step Moore neighborhood. Due to the continuing nature of our RL task, the return is essentially the summation of all the *rewards* obtained during the evaluation steps, no discount is used and no resets to starting states are happening between evaluation steps per experiment.

4.2 Results

4.2.1 Hyperparameter Comparison

The obtained *return*, per experiment, from interacting with the environment 100,000 steps is shown in Table 4.2. Five experiments were able to perform better than the base heuristic (c, a, d, i, g runs). Three were worse than random (h, f, e runs) and one (b run) was in between. The maximum obtained *return* was 640,643 a 1.27 fold increase in performance over the base heuristic (return of 503,521).

Table 4.2: Comparison of obtained returns from the nine runs. The return was computed from playing 100,000 steps per run following the learned policy. The runs are ordered from best to worst and are named from a to i, the "heuristic" and "random" baselines are marked as "H" and "R" respectively.

Run	Return	Exploration	Unrolling	Architecture	LR	Batch Size	Gamma
c	640,658	heuristic	2	A3	0.0001	16	0.99
a	638,094	linear	2	A1	0.0003	32	0.99
d	615,091	heuristic	3	A3	0.0003	32	0.99
i	591,021	linear	2	A3	0.0003	32	0.99
g	507,313	linear	10	A3	0.0003	32	0.99
H	**503,521**						
b	327,597	linear	1	A2	0.0001	256	0.90
R	**319,833**						
h	294,965	linear	1	A3	0.0001	16	0.99
f	293,722	heuristic	1	A3	0.0003	32	0.99
e	293,600	heuristic	1	A2	0.0001	16	0.90

The comparison of the obtained *mean reward per step* and *reward standard deviation* is shown in Figure 4.5. It can be observed that the best runs (c, a, d, i, g) also have a smaller *standard deviation*.

Furthermore, a comparison of agents following ϵ-*greedy* policies is shown in Figure 4.6. It could be seen that the performance is close to their respective *greedy* policies.

4.2.2 Training Dynamics

During the training of each DQN run the change in *rewards* was monitored, logging the *reward* every 10 iterations. From this loggings the *mean reward per step* was calculated by a sliding window of size 1,000. The *mean reward per step* during learning are shown in Figure 4.7.

Similarly, the values of the error were logged each 10 steps. The error behavior

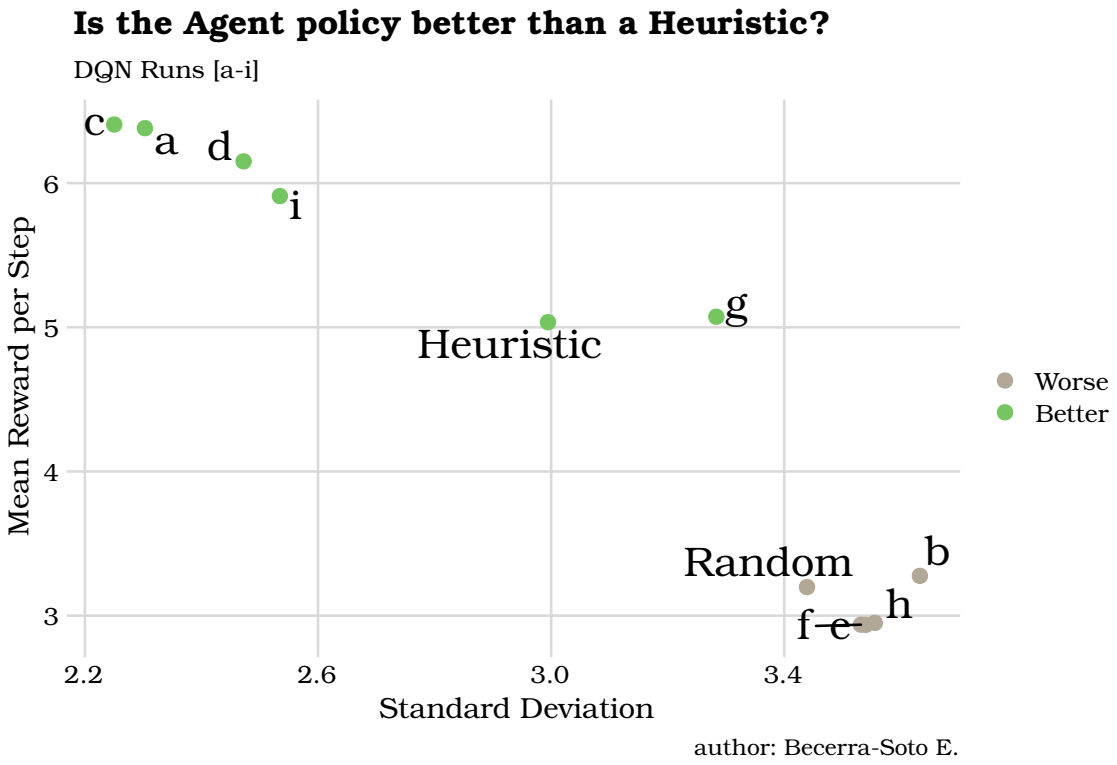


Figure 4.5: The mean and standard deviation for each model was computed from playing the environment 100,000 steps, following greedy policies ($\epsilon = 0$).

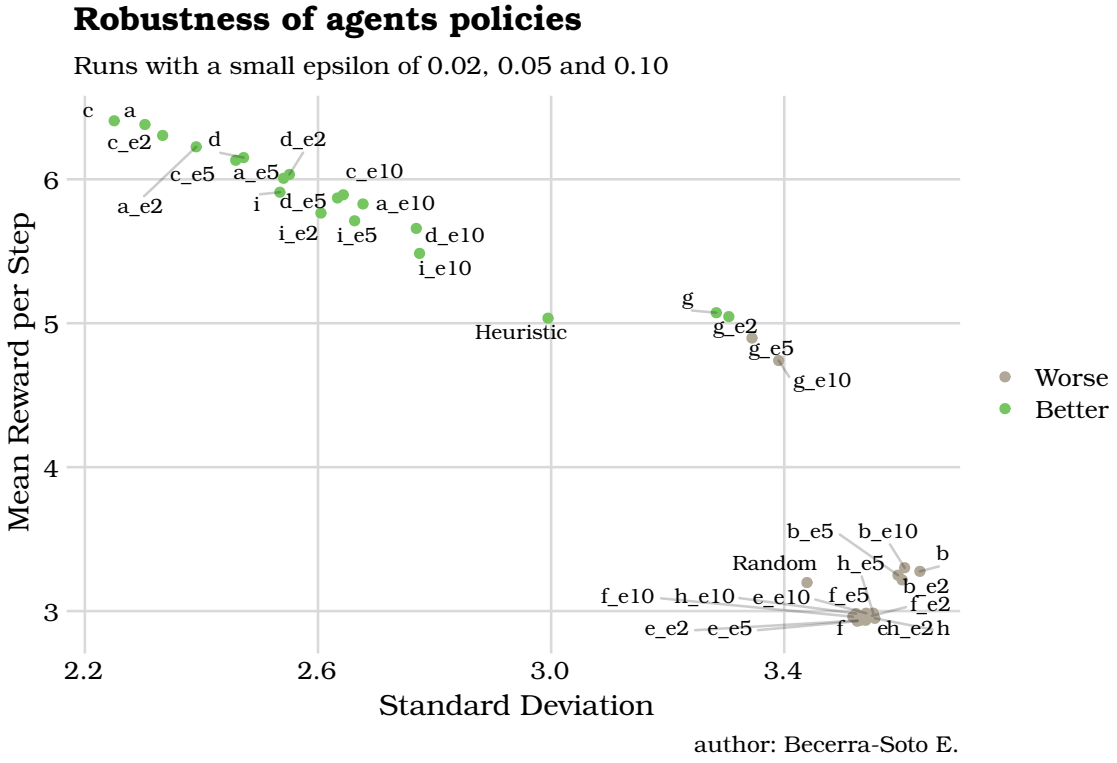


Figure 4.6: Performance of ϵ-*greedy* policies with ϵ values of 0.02, 0.05 and 0.10.

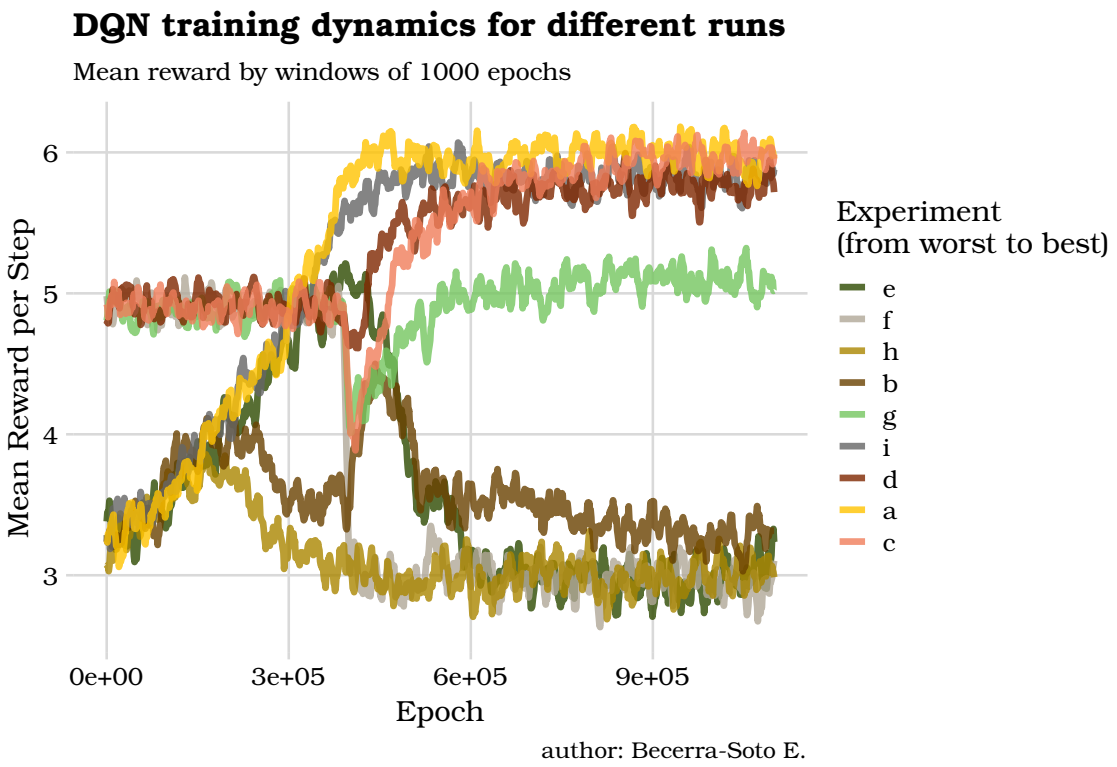


Figure 4.7: Mean Rewards per step during training (1,000 steps sliding window).

during training can be seen in Figure 4.8.

Visual demonstrations are available for the best two runs:

- Run C: *https://www.youtube.com/watch?v=9qFw78___sSM*
- Run A: *https://www.youtube.com/watch?v=DHdRd96KEZA*

4.3 Discussion

The main objective of this thesis is to design and characterize a novel environment for *Reinforcement Learning*. It must be based on Cellular Automata, with the rationale that a handful of well constructed CA based environments would provide an "interesting" set of benchmark tasks for Reinforcement Learning algorithms.

The appeal of such tasks arises from the properties of CA. Some theoretical and practical properties that could make CA a good RL benchmark are:

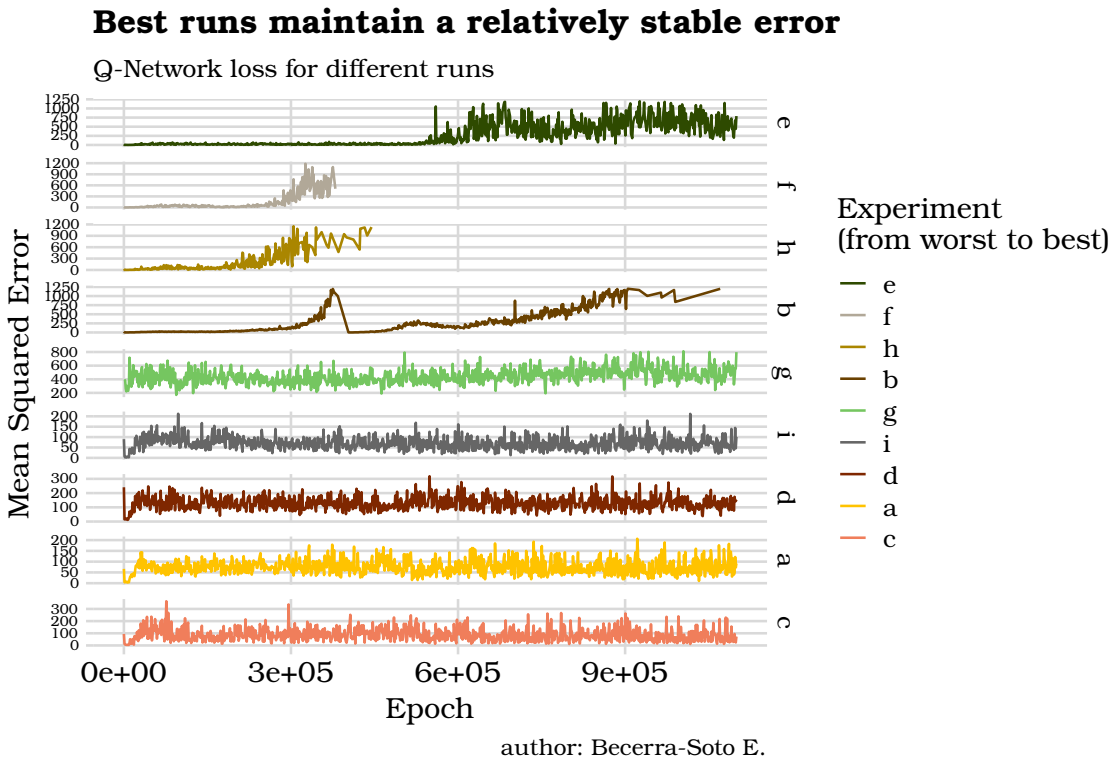


Figure 4.8: Mean Squared Error (MSE) during training. The scale of the y-axis varies between experiments as different values of γ and *unrolling* were used. Also the error was capped at 1,200 due to the increasing error of the divergent models.

1. They are easy to implement.
2. They can model complex real world phenomena.
3. Some CA have *universal computation* capabilities.

Under the previous justifications, we have shown that is possible to design a CA based Reinforcement Learning task. Additionally, the task has real world semantics, namely a Helicopter extinguishing a wildfire. Our particular proposed environment is an agent affecting cells on top of a Drossel and Schwabl forest fire model (Drossel and Schwabl 1992). To further characterize the proposed environment we have applied DQN to try to solve it.

4.3.1 Effect of Hyperameters

The obtained results from the non-exhaustive hyperparameter search can be seen in Table 4.2. Answering the question of to what degree the hyperparameters affect the results of each run is not trivial and mostly experimental arguments can be provided. This is because attribution becomes difficult when multiple hyperparameters are changed at the same time since their interactions can be non-linear, yet carefully tuning a single hyperparameter can be inefficient as we would only explore a single dimension of hyperparameter space.

The experiments c, a, d, i and g have a better performance than the baseline heuristic. The hyperparameter that seems to have the greater effect is *Unrolling*. When it is equal to 1 the experiments failed to learn (beating the heuristic). *Unrolling* helps to "accelerate" learning as future steps rewards information is used to estimate the Q values. However, in our case, this "acceleration" meant the difference between learning something or nothing at all. A comparison between experiments c, h, and f is interesting as they have similar hyperparameters values but h and f failed to learn and arguably the difference can be attributed to the *Unrolling* hyperparameter.

The tried *learning rate* values were 0.0001, 0.0003, both of them worked. A small

minibatch size (16 and 32) and a γ of 0.99 also worked.

One of the approaches to make the task easier was to allow some agents to train following a heuristic policy during exploration. Even though the best performing run used the heuristic exploration, other agents were capable of learning by a simple ϵ linear decrease from 1.0 to 0.10 during the first 400,000 iterations. Thus learning is achievable without the heuristic and in future experiments within the same environment it could be discarded as it can bias learning. The impact of exploring with the heuristic can be seen in the training dynamics plot (Figure 4.7) where the agents using the heuristic start the training with much better performance, but after exploration is finished some manage to learn from the heuristic (c, d) and others not and its performance plummets (f, e). Similar behavior can be observed in the epochs vs loss plot (Figure 4.8), where, for the heuristic runs, the error started low and it was kept in that way during the exploration phase, since the linear exploration policy is steadily changing as opposed to the heuristic. Then when exploration finished and the 0.10-greedy policy took effect, some runs had learned and its error was maintained and others had not and its errors blew up.

The chosen ANN architectures are in increasing order of complexity: A2, A1 and A3 (see Figures 4.3, 4.2 and 4.4). The best run employed the more complex network (A3). The A3 architecture has a number of parameters in the same order of magnitude as the environment states (2.87881×10^5 parameters vs. 3.54294×10^5 states), so it is expected to be considerably overfitting. The number of parameters of the next performing network (4.9673×10^4 parameters) is of an order below, but still, it is expected to be overfitting.

Training for more iterations presumably would not increase significantly the performance of any experiment as it could be seen from the training dynamics plot (see Figure 4.7).

From the evaluation data, it can be observed that the best performing runs have a lower *standard deviation* to those that failed to learn (Figure 4.5), thus the better

runs not only played better but they consistently did it. Likewise, from the same data, three overall regions, in terms of *mean reward* and *standard deviation*, can be seen. The first one groups the best runs together. The second groups the worst runs. The third one contains the Heuristic and the g run. From the third region is interesting to notice that even so agent g plays better than the heuristic it does it with a higher *standard deviation*, presumably because it performs well in some environment states but badly in others, besides g used an *Unrolling* of 10 which could explain that some estimations of Q values were biased with such large *Unrolling*.

To test for the robustness and maximization bias of the learned policies, the agents were also evaluated with ϵ-greedy policies of 0.02, 0.05 and 0.10. We did not find anything unexpected, since the performance of better than the heuristic agents deteriorated (Figure 4.6).

5

Conclusion

In this document, we have explored the advantages of using Cellular Automata as Reinforcement Learning environments. In particular, a novel task has been proposed. It consists of an agent navigating through a forest fire CA with the goal of reducing the extension and amount of fire cells. The task has the semantics of a "Helicopter" trying to extinguish a wildfire. The optimization problem was put in terms of the Reinforcement Learning paradigm and then tried to be solved with the Deep Q-Networks algorithm.

Our results suggest that CA can successfully be used as RL environments. Furthermore, we believe that CA could be a promising benchmark for RL due to their simple formulation yet they are capable of modeling complex systems and have interesting computing characteristics. Consequently, CA based

environments might be closer to real applications than current video game based benchmarks.

Future work would include extending the grid size, as for now our results are very preliminary on a 3×3 grid. Further characterization of our proposed environment remains, as only one algorithm was applied, so a natural step is to try other families of RL methods, like policy gradients and comparisons to other RL complementary methodologies such as *searching* and *planning* would be necessary.

Encouraging a CA based RL benchmark must involve featuring a broad number of interesting tasks, as this was one of the key factors for the success of ALE. Modifications of the parameters of our proposed environment would change the semantics of our task. For example, changing the effects of the agent to transform tree cells to empty cells or to rock cells would modify the semantics to one of a "builder" or "worker" trying to extinguish the fire by means of barriers or calculated tree cut downs. This task could be translated to an algorithmic context of finding cuts on a graph, represented by the forest structure, that minimize the spread of a controlled initial fire. Besides completely new and relevant tasks could be proposed from the CA literature.

Appendices

5.1 Appendix A Algorithms

5.1.1 Agent-Environment interactions

Algorithm 1: Agent-Environment interactions.

Input : Steps to play T, Synchronization parameter m.

Output: Return G (accumulated reward).

Initialize: Automaton Grid $C_{i \times j}$
Initialize: Helicopter H: $H = (row, col) : row \in \{1, ..., i\}, col \in \{1, ..., j\}$
$G \longleftarrow 0$
$k \longleftarrow m$
for *1, ..., T* **do**
    if *k has value* 0 **then**
        $C_{i \times j} \longleftarrow$ `computeForestFire`$(C_{i \times j})$
        $k \longleftarrow m$
    else
        $k \longleftarrow k - 1$
    end
    $a \longleftarrow$ `policy`$(C_{i \times j}, H, m)$
    $H \longleftarrow$ `updatePosition`(H, a)
    $C_{i \times j} \longleftarrow$ `applyEffect`$(C_{i \times j}, H)$
    $r \longleftarrow$ `getReward`$(C_{i \times j})$
    $G \longleftarrow G + r$;
end

5.1.2 Agent-Environment interactions following Open AI Gym API

Algorithm 2: Agent-Environment interactions following Open AI Gym API.

Input : Steps to play T, *Environment* and *Agent* objects.

Output: Return G (accumulated reward).

```
/* Commonly used names: s and s' for current and new states, a for
   action and r for reward.                                        */
/* Environment object is reset to starting state, it returns the
   first observation                                               */
1  s ⟵ Environment.reset()
2  G ⟵ 0
3  FALSE ⟵ terminationSignal
4  for 1, ..., T do
5      if terminationSignal then
           /* The episode is over, reset Environment again.       */
6          s' ⟵ Environment.reset()
7          r ⟵ 0
8      else
9          a ⟵ Agent.policy(s)
10         stepData ⟵ Environment.step(a)
11         s' ⟵ stepData[0]
12         r ⟵ stepData[1]
13         terminationSignal ⟵ stepData[2]
14     end
15     G ⟵ G + r
16     s ⟵ s'
17 end
```

5.1.3 Deep Q-networks (DQN) using Open AI Gym API

Algorithm 3: Deep Q-networks (DQN) using Open AI Gym API.

Input : Total iterations T to train Q network.

Output: Trained Q network with weights θ.

Initialize replay memory D to capacity N

Initialize policy network Q with random weights θ

Initialize target network $\hat{Q}$ with weights θ^- such that $\theta^- = \theta$

Get initial observation: $s \longleftarrow$ `Environment`.$reset()$

for *t=1, ..., T* **do**

  With probability ϵ: select random action a

  Otherwise select: $a \longleftarrow \arg\max_{a'} Q(s, a'; \theta)$

  Execute `Environment`.$step(a)$ to obtain reward r and next observation s'

  Store transition (s, a, r, s') in D

  Sample a minibatch of transitions (s_i, a_i, r_i, s'_i) from D

  $$y_i \longleftarrow \begin{cases} r_i & \text{If the episode was over at iteration } i \\ r_i + \gamma \max_{a'} \hat{Q}\ (s'_i, a'; \theta^-) & \text{Otherwise} \end{cases}$$

  Perform optimization step on $\Big(y_i - Q(s_i, a_i; \theta)\Big)^2$ with respect to θ

  Every C steps synchronize networks weights: $\theta^- = \theta$

  if *Episode was over at iteration t* **then**

    $s \longleftarrow$ `Environment`.$reset()$

  else

    $s \longleftarrow s'$

  end

end

References

Alonso-Sanz, Ramon. 2011. *Discrete Systems with Memory*. Vol. 75. World Scientific.

Badia, Adrià Puigdomènech, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, and Charles Blundell. 2020. "Agent57: Outperforming the Atari Human Benchmark." *arXiv Preprint arXiv:2003.13350*.

Bak, Per. 2013. *How Nature Works: The Science of Self-Organized Criticality*. Springer Science & Business Media.

Bak, Per, Kan Chen, and Chao Tang. 1990. "A Forest-Fire Model and Some Thoughts on Turbulence." *Physics Letters A* 147 (5-6): 297–300.

Bellemare, Marc G, Yavar Naddaf, Joel Veness, and Michael Bowling. 2013. "The Arcade Learning Environment: An Evaluation Platform for General Agents." *Journal of Artificial Intelligence Research* 47: 253–79.

———. 2015. "The Arcade Learning Environment: An Evaluation Platform for General Agents (Extended Abstract)." *In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 4148–52.

Brockman, Greg, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. "OpenAI Gym."

Codd, EF. 1968. "Cellular Automata, Academic Press." *New York*.

Cook, Matthew. 2004. "Universality in Elementary Cellular Automata." *Complex Systems* 15 (1): 1–40.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. "Imagenet: A Large-Scale Hierarchical Image Database." In *2009 Ieee Conference on Computer Vision and Pattern Recognition*, 248–55. Ieee.

Drossel, Barbara, and Franz Schwabl. 1992. "Self-Organized Critical Forest-Fire Model." *Physical Review Letters* 69 (11): 1629.

Elwin, R Berkelamp, John H Conway, and Richard K Guy. 1982. "Winning Ways for Your Mathematical Plays." Academic Press.

Ermentrout, G Bard, and Leah Edelstein-Keshet. 1993. "Cellular Automata Approaches to Biological Modeling." *Journal of Theoretical Biology* 160 (1): 97–133.

Foley, John, Emma Tosch, Kaleigh Clary, and David Jensen. 2018. "Toybox: Better Atari Environments for Testing Reinforcement Learning Agents." *arXiv Preprint arXiv:1812.02850*.

Ganguly, Niloy, Biplab K Sikdar, Andreas Deutsch, Geoffrey Canright, and P Pal Chaudhuri. 2003. "A Survey on Cellular Automata."

Gardner, Martin. 1970. "Mathematical Games." *Scientific American* 222 (6): 132–40.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification." In *Proceedings of the Ieee International Conference on Computer Vision*, 1026–34.

Hedlund, Gustav A. 1969. "Endomorphisms and Automorphisms of the Shift Dynamical System." *Mathematical Systems Theory* 3 (4): 320–75.

Hoekstra, Alfons G, Jiri Kroc, and Peter MA Sloot. 2010. *Simulating Complex Systems by Cellular Automata*. Springer.

Hooker, John N. 1995. "Testing Heuristics: We Have It All Wrong." *Journal of Heuristics* 1 (1): 33–42.

Hölldobler, Bert, Edward O Wilson, and others. 1994. *Journey to the Ants: A Story of Scientific Exploration.* Harvard University Press.

Ilachinski, Andrew. 2001. *Cellular Automata: A Discrete Universe.* World Scientific Publishing Company.

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

Jaderberg, Max, Wojciech M Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, et al. 2019. "Human-Level Performance in 3D Multiplayer Games with Population-Based Reinforcement Learning." *Science* 364 (6443): 859–65.

Kari, Jarkko. 1994. "Reversibility and Surjectivity Problems of Cellular Automata." *Journal of Computer and System Sciences* 48 (1): 149–82.

Kauffman, Stuart A. 1984. "Emergent Properties in Random Complex Automata." *Physica D: Nonlinear Phenomena* 10 (1-2): 145–56.

Kingma, Diederik P, and Jimmy Ba. 2014. "Adam: A Method for Stochastic Optimization." *arXiv Preprint arXiv:1412.6980*.

Kůrka, Petr. 1997. "Languages, Equicontinuity and Attractors in Cellular Automata." *Ergodic Theory and Dynamical Systems* 17 (2): 417–33.

Lanctot, Marc, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, et al. 2019. "OpenSpiel: A Framework for Reinforcement Learning in Games." *arXiv Preprint arXiv:1908.09453*.

Louis, Pierre-Yves, and Francesca R Nardi. 2018. *Probabilistic Cellular Automata*. Springer.

Machado, Marlos C, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. 2018. "Revisiting the Arcade Learning Environment: Evaluation Protocols and Open Problems for General Agents." *Journal of Artificial Intelligence Research* 61: 523–62.

Margolus, Norman. 1984. "Physics-Like Models of Computation." *Physica D: Nonlinear Phenomena* 10 (1-2): 81–95.

———. 1995. "CAM-8: A Computer Architecture Based on Cellular Automata." *arXiv Preprint Comp-Gas/9509001*.

Millington, James DA, George LW Perry, and Bruce D Malamud. 2006. "Models, Data and Mechanisms: Quantifying Wildfire Regimes." *Geological Society, London, Special Publications* 261 (1): 155–67.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. "Playing Atari with Deep Reinforcement Learning." *arXiv Preprint arXiv:1312.5602*.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, et al. 2015. "Human-Level Control Through Deep Reinforcement Learning." *Nature* 518 (7540): 529–33.

Moravčík, Matej, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. 2017. "Deepstack: Expert-Level Artificial Intelligence in Heads-up No-Limit Poker." *Science* 356 (6337): 508–13.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, 8024–35. Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Popovici, Adriana, and Dan Popovici. 2002. "Cellular Automata in Image Processing." In *Fifteenth International Symposium on Mathematical Theory of Networks and Systems*, 1:1–6. Citeseer.

Poundstone, William. 2013. *The Recursive Universe: Cosmic Complexity and the Limits of Scientific Knowledge*. Courier Corporation.

Risi, Sebastian, and Mike Preuss. 2020. "From Chess and Atari to Starcraft and Beyond: How Game Ai Is Driving the World of Ai." *KI-Künstliche Intelligenz* 34 (1): 7–17.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, et al. 2015. "Imagenet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115 (3): 211–52.

Silver, David, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, et al. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* 529 (7587): 484–89.

Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, et al. 2018. "A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go Through Self-Play." *Science* 362 (6419): 1140–4.

Smith III, Alvy Ray. 1971. "Simple Computation-Universal Cellular Spaces." *Journal of the ACM (JACM)* 18 (3): 339–53.

Stanley, Kenneth O, and Joel Lehman. 2015. *Why Greatness Cannot Be Planned: The Myth of the Objective*. Springer.

Sutton, Richard S. 1988. "Learning to Predict by the Methods of Temporal Differences." *Machine Learning* 3 (1): 9–44.

Sutton, Richard S, and Andrew G Barto. 2018. *Reinforcement Learning: An Introduction*. MIT press.

Tesauro, Gerald. 1995. "Temporal Difference Learning and Td-Gammon."

Communications of the ACM 38 (3): 58–68.

Todorov, Emanuel, Tom Erez, and Yuval Tassa. 2012. "Mujoco: A Physics Engine for Model-Based Control." In *2012 Ieee/Rsj International Conference on Intelligent Robots and Systems*, 5026–33. IEEE.

Toffoli, Tommaso. 1977. "Cellular Automata Machines."

———. 1984. "CAM: A High-Performance Cellular-Automaton Machine." *Physica D: Nonlinear Phenomena* 10 (1-2): 195–204.

Tsitsiklis, John N, and Benjamin Van Roy. 1997. "Analysis of Temporal-Diffference Learning with Function Approximation." In *Advances in Neural Information Processing Systems*, 1075–81.

Turing, Alan Mathison. 1936. "On Computable Numbers, with an Application to the Entscheidungsproblem." *J. Of Math* 58 (345-363): 5.

Vichniac, Gérard Y. 1984. "Simulating Physics with Cellular Automata." *Physica D: Nonlinear Phenomena* 10 (1-2): 96–116.

Von Neumann, John, Arthur W Burks, and others. 1966. "Theory of Self-Reproducing Automata." *IEEE Transactions on Neural Networks* 5 (1): 3–14.

Von Neumann, John, and others. 1951. "The General and Logical Theory of Automata." *1951*, 1–41.

Walt, Stéfan van der, S Chris Colbert, and Gael Varoquaux. 2011. "The Numpy Array: A Structure for Efficient Numerical Computation." *Computing in Science & Engineering* 13 (2): 22–30.

Weiner, N, and A Rosenblunth. 1946. "The Mathematical Formulation of the Problem of Conduction of Impulses in a Network of Connected Excitable Elements Specifically in Cardiac Muscle."

Wolfram, Stephen. 1983. "Statistical Mechanics of Cellular Automata." *Reviews*

of Modern Physics 55 (3): 601.

———. 2002. *A New Kind of Science*. Vol. 5. Wolfram media Champaign, IL.

Zinck, RD, and V Grimm. 2008. "More Realistic Than Anticipated: A Classical Forest-Fire Model from Statistical Physics Captures Real Fire Shapes." *The Open Ecology Journal* 1 (1).

Zinck, Richard D, Karin Johst, and Volker Grimm. 2010. "Wildfire, Landscape Diversity and the Drossel–Schwabl Model." *Ecological Modelling* 221 (1): 98–105.

Zuse, Konrad. 1970. *Calculating Space*. Massachusetts Institute of Technology, Project MAC Cambridge, MA.