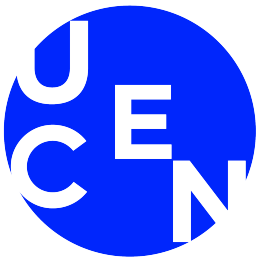


UNIVERSIDAD CENTRAL DE CHILE
FACULTAD DE INGENIERÍA Y ARQUITECTURA
ESCUELA DE INGENIERÍA

Agente de aprendizaje por refuerzo con enfoque basado en introspección implementado en un entorno competitivo

Memoria para optar al título profesional de
Ingeniero Civil en Computación e Informática.

Profesor Guía: **Francisco Cruz Naranjo**
Profesor Informante: **Alejandro Sanhueza Olave**
Profesor Informante: **Hernán Olmi Reyes**

Alfonso Enrique Opazo Muñoz

Santiago, Chile
2021

UNIVERSIDAD CENTRAL DE CHILE
FACULTAD DE INGENIERÍA Y ARQUITECTURA
ESCUELA DE INGENIERÍA

Agente de aprendizaje por refuerzo con enfoque basado en introspección implementado en un entorno competitivo

Memoria para optar al título profesional de
Ingeniero Civil en Computación e Informática.

Profesor Guía: **Francisco Cruz Naranjo**
Profesor Informante: **Alejandro Sanhueza Olave**
Profesor Informante: **Hernán Olmi Reyes**

QUIENES RECOMIENDAN QUE SEA ACEPTADA PARA COMPLETAR LAS EXIGENCIAS DEL TÍTULO DE INGENIERÍA CIVIL EN COMPUTACIÓN E INFORMÁTICA

Alfonso Enrique Opazo Muñoz

Santiago, Chile
2021

Esta tesis está dedicada a mi familia, a mi compañera de vida y a mi Bruce
Alfonso Opazo Muñoz

Resumen

Los algoritmos de inteligencia artificial o más específicamente el machine learning están inmersos en la sociedad más de lo que las personas creen. Desde la selección de un correo spam, hasta la sugerencia de la siguiente palabra al momento de escribir un mensaje de texto. El área del aprendizaje por refuerzo es considerado un paradigma de aprendizaje que basa su estructura en la psicología conductual y sirve para afrontar problemas donde no existen datos previos de dicho problema, es decir, el agente aprende mediante la implementación de ensayo y error. Introducir este tipo de algoritmos a procesos complejos dentro de la sociedad es considerado una tarea desafiante y emocionante a largo plazo. Es por eso, que surge una nueva área derivada del aprendizaje por refuerzo llamada aprendizaje por refuerzo explicable. Este nuevo enfoque viene a solucionar problemas relacionados con la confianza y la transparencia que personas sin conocimientos técnicos tienen sobre este tipo de sistemas.

De acuerdo a lo planteado anteriormente y mediante el uso del método científico, se implementó un agente de aprendizaje por refuerzo con la ayuda un algoritmo muy popular llamado Deep Q-Network. A dicho agente, se le incorporó adicionalmente un nuevo enfoque propuesto basado en introspección con la finalidad de poder obtener las probabilidades de éxito del agente al completar un juego en función principalmente de los Q-values obtenidos. Es decir, cuál es la probabilidad de éxito al realizar cierta acción durante el transcurso del juego.

Finalmente, el agente propuesto con enfoque basado en introspección, logra en primera instancia, ganar una cierta cantidad de rondas durante el proceso de entrenamiento, siendo en muchos pasajes del juego, más competitivo que sus contrincantes. Gracias a las probabilidades de éxito retornadas por el enfoque propuesto, dan cuenta que si bien el agente pudo completar una cantidad razonable de juegos y genero estrategias para poder ganar, posterior a una cierta cantidad de juegos completados, no puedo mantener un ritmo constante y el proceso de aprendizaje se

estancó, dando lugar a los otros dos tipos de agentes. Es decir, la toma de decisión del agente basado en introspección, no fue eficiente a largo plazo.

Palabras Claves: inteligencia artificial, inteligencia artificial explicable, aprendizaje por refuerzo, aprendizaje por refuerzo explicable, interacción humano-robot, entorno competitivo.

Abtract

Artificial intelligence algorithms or more specifically machine learning are embedded in society more than people realize. From the selection of a spam email, to the suggestion of the next word when writing a text message. The area of reinforcement learning is considered a learning paradigm that bases its structure on behavioral psychology and serves to face problems where there is no previous data of said problem, that is, the agent learns through the implementation of trial and error. Introducing these types of algorithms to complex processes within society is considered a challenging and exciting task in the long term. That is why a new area derived from reinforcement learning arises called explainable reinforcement learning. This new approach comes to solve problems related to trust and transparency that people without technical knowledge have about this type of system.

According to the above and through the use of the scientific method, a reinforcement learning agent was implemented with the help of a very popular algorithm called Deep Q-Network. To this agent, a new proposed approach based on introspection was additionally incorporated in order to be able to obtain the probability of success of the agent when completing a game based mainly on the Q-values obtained. That is, what is the probability of success when performing a certain action during the course of the game.

Finally, the proposed agent with an introspection-based approach, manages in the first instance, to win a certain number of rounds during the training process, being in many parts of the game, more competitive than his opponents. Thanks to the probabilities of success returned by the proposed approach, they realize that although the agent was able to complete a reasonable number of games and generated strategies to be able to win, after a certain number of games completed, I cannot maintain a constant rhythm and the learning process stalled, leading to the other two types of agents. In other words, the agent's decision-making based on introspection was not efficient in the long term.

Keywords: artificial intelligence, explainable artificial intelligence, reinforcement learning, explainable reinforcement learning, human-robot interaction, competitive environment.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

No es nada nuevo leer en los periódicos o escuchar en la televisión algún reportaje sobre los avances de la tecnología y lo rápido que esta se mueve en sus diferentes campos. Uno de los campos que más ha tenido aportes es el de la Inteligencia Artificial (AI por sus siglas en inglés). La Inteligencia Artificial según Margaret A Boden (2016) tiene como objetivo que los computadores puedan realizar la misma clase de cosas que puede hacer la mente humana. Según lo planteado por esta autora, el desarrollo de la Inteligencia Artificial tiene como objetivos principales aportar a los campos de la tecnología (usar los computadores para realizar cosas útiles) y las ciencias (ayudar mediante algoritmos a resolver problemas de los seres humanos y de los demás seres vivos) (Boden, 2016).
Uno de los conceptos que se deben tomar en cuenta para considerar que una máquina sea realmente inteligente es su forma de aprender. Dentro de los paradigmas del aprendizaje, podemos distinguir entre el aprendizaje supervisado, no supervisado y por refuerzo (Russell y Norvig, 2010). Este último, el aprendizaje por refuerzo, será el dominio en el que nos enfocaremos con respecto a esta área de la IA.

El enfoque principal del aprendizaje por refuerzo, demuestra en términos generales, como un agente puede interactuar con el entorno y recibir una recompensa de acuerdo al estado y a una acción tomada en determinado momento. El objetivo primordial de un agente de este tipo es obtener la mayor recompensa de su entorno.

Un elemento muy importante en el desarrollo de agentes de aprendizaje por refuerzo es el entorno con el cual el agente interactúa. El entorno virtual elegido para el aprendizaje del agente tiene la característica principal de estar basado en competencia y permite la interacción entre personas y robots (Barros, Bloem y Barakova, 2020).

El objetivo principal de este proyecto de tesis es desarrollar un agente de aprendizaje por refuerzo en un entorno competitivo, utilizando un enfoque basado en fenomenología.

Este proyecto busca disminuir la brecha que existe en la confiabilidad que los usuarios no expertos en inteligencia artificial, o derechamente en aprendizaje por refuerzo, perciben de acuerdo a todos estos tipos de sistemas. La carencia de transparencia en la toma de decisión en sistemas basados en inteligencia artificial es uno de los temas más difíciles y complejos de abordar hoy en día.

1.1. Motivación

Gracias a los avances en materia de robótica e inteligencia artificial, se ha generado un progreso significativo en estas áreas. Estos avances han contribuido de forma positiva en las sociedades. Pero aún queda mucho camino por recorrer.

Se espera que, en un futuro no muy lejano, los robots controlados por sistemas autónomos puedan aportar en diferentes áreas de la sociedad como lo son la industria, manufactura, la banca, el rescate de personas y animales y el cuidado de personas con poca movilidad. Pero nada de esto se puede lograr sin antes trabajar en asuntos principalmente humanas como es la confianza, la transparencia y la ética. Este conjunto de características, principalmente del ser humano son la base fundamental que motiva a trabajar en algoritmos que permitan a todos, expertos y no expertos en estos sistemas, comprender por qué cada sistema toma las decisiones que toma.

1.2. Definición del Problema

Un problema abierto que afecta a muchos sistemas, ya sean sistemas conformados por personas o por máquinas, es la falta de transparencia en la forma en cómo funcionan dichos sistemas por parte de personas no expertas. Dicho esto, llevando el problema de la transparencia al área de la inteligencia artificial, un problema abierto en el aprendizaje por refuerzo es la carencia de entendimiento de un usuario final en términos de la toma de decisión por él agente durante el proceso de aprendizaje (Cruz, Dazeley y Vamplew, 2020).

La transparencia se vuelve un elemento fundamental para incorporar estos sistemas a la vida cotidiana (robots hogareños) o delegarlos a procesos productivos donde no existe margen de error. El uso de la inteligencia artificial y de sus algoritmos

se están volviendo cada vez más comunes en diferentes áreas de la industria y no es fácil para una persona no experta encomendar tareas importantes a un sistema basado en AI que no sea capaz de justificar su razonamiento (Cruz, Dazeley y Vamplew, 2020).

Se puede señalar, que la característica de transparencia en los sistemas deriva en lo que se conoce como un problema abierto en Aprendizaje por Refuerzo. Este problema se centra en la carencia de algún mecanismo que permita a los agentes comunicar claramente las razones de por qué tal agente elige cierta acción dado un estado particular (Cruz, Dazeley y Vamplew, 2020).

1.3. Objetivos

1.3.1. Objetivo General

Implementar un agente de aprendizaje por refuerzo con enfoque basado en introspección para la explicación de jugadas en un entorno competitivo.

1.3.2. Objetivos Específicos

En el presente trabajo, se han definido los siguientes objetivos específicos:

- Recopilar información sobre inteligencia artificial y aprendizaje por refuerzo explicable.
- Evaluar diferentes librerías y frameworks enfocados en aprendizaje por refuerzo.
- Desarrollar un agente artificial de aprendizaje por refuerzo mediante el uso del lenguaje de programación Python.
- Desarrollar o utilizar un escenario experimental ya desarrollado de carácter competitivo
- Implementar el método propuesto basado en introspección en un entorno competitivo y analizar los respectivos resultados.

1.4. Hipótesis

Los entornos en los cuales los agentes se encuentran inmersos e interactúan son considerados elementos primordiales en el aprendizaje por refuerzo definen entre

otras cosas como el agente interactúa en sí mismo. La mayoría de las investigaciones solo enfocan sus estudios en el comportamiento de un agente en particular bajo ciertos parámetros del entorno y del agente. A diferencia de lo anterior, la naturaleza del entorno elegido (El sombrero del chef) se constituye por la participación de más de una agente, 3 en total. Esto marca una diferencia sustancial en la mayoría de los entornos utilizados generalmente

Los agentes implementados en este trabajo son los siguientes: un agente basado en DQN (Deep Q-network) con implementación del enfoque basado en introspección, un agente PPO (Proximal Policy Optimization) y dos agentes que realizando acciones aleatorias.

La hipótesis presentada en este trabajo tiene la finalidad de comprobar si los agentes desarrollados e implementados tienen la capacidad de entregar información concreta y comprensible relacionada con el proceso de aprendizaje y principalmente en la toma de decisión.

De otro modo, el agente puede ser capaz de obtener los valores provenientes de su proceso de aprendizaje y de toma de decisiones, sin embargo, los datos obtenidos carecerán de todo entendimiento y lógica, confundiendo más a las personas no expertas.

Finalmente, el agente no es capaz de terminar su proceso de aprendizaje, por lo que no es posible considerarlo como un agente inteligente, ni mucho menos, asignarle tareas que requieran mínimos márgenes de error.

1.5. Metodología de Trabajo

Para poder explicar nuevas propuestas o fenómenos, en la historia de la ciencia han surgido diversos enfoques de pensamiento como el materialismo dialéctico, el positivismo, fenomenología, etc. que propicien el alcance a un nuevo conocimiento. Debido a los avances exponenciales en materia de investigación científica, surgió una polarización en la forma de investigar lo que derivó en lo que hoy conocemos como el enfoque cualitativo y el enfoque cuantitativo (Sampieri, Collado y Lucio, 2006).

De igual modo, muchos procesos y enfoques de investigación están basados en etapas. Cada una de estas etapas tiene un orden y posición. De esta forma, el enfoque cuantitativo de investigación corresponde a un conjunto de procesos secuenciales y probatorios (Sampieri, Collado y Lucio, 2014). Cada una de estas etapas debe

cumplirse rigurosamente en orden secuencial. En otras palabras, la ejecución en paralelo de una etapa no es posible aplicarla a este enfoque.

De modo que, todo lo mencionado sobre los enfoques de investigación desarrollados en el transcurso de la historia, el presente trabajo a realizar tiene características que fundamentan la investigación de carácter cuantitativo.

Según Feynman, el método científico se define como la observación, razonamiento y experimentación (Leó et al, 2015). La figura 1.1 representa el esquema general de las etapas del método científico adaptadas al proyecto.

Basándonos en el contexto científico en el cual se encuentra el área de estudio y en la búsqueda de la validación de la hipótesis, se recurre al método científico como metodología.

- **Planteamiento del problema:**: El problema principal que surge en los métodos basados en aprendizaje por refuerzo es la claridad y transparencia para explicar la toma de decisión en el proceso de aprendizaje por parte del agente artificial (Cruz, Dazeley y Vamplew, 2020). Debido a esta razón, un usuario final, no experto en el área, no confiará ninguna tarea importante a un sistema que no sea capaz de razonar la toma de decisión.

- **Indagar sobre entornos y enfoques existentes**: Se indagará en profundidad sobre información relacionada directamente con entornos competitivos y enfoques actuales de aprendizaje por refuerzo.

- **Desarrollo o implementación del entorno**: El primer paso para comenzar con el experimento es desarrollar el agente artificial basado en aprendizaje por refuerzo implementando el enfoque basado en introspección. Definir los elementos que componen a un agente artificial, como son la política, la señal de recompensa, la función de valor y el respectivo modelo del ambiente. Es motivo de evaluación que tipo de función de valor y política se implementaran al momento de ejecutar el experimento.

  Siguiendo con los pasos del proyecto y después de propuestas sobre entornos, se llego a la conclusión de que el juego del Sombrero del chef, cumplía con los requerimientos básicos necesarios el desarrollo del proyecto que eran ser un entorno competitivo, fácil de jugar, requería poca capacidad de procesamiento y con valores discretos.

- **Resultados del informe**: Los datos, información y gráficos pertenecientes al desarrollo de este proyecto, se verán materializados en un informe escrito

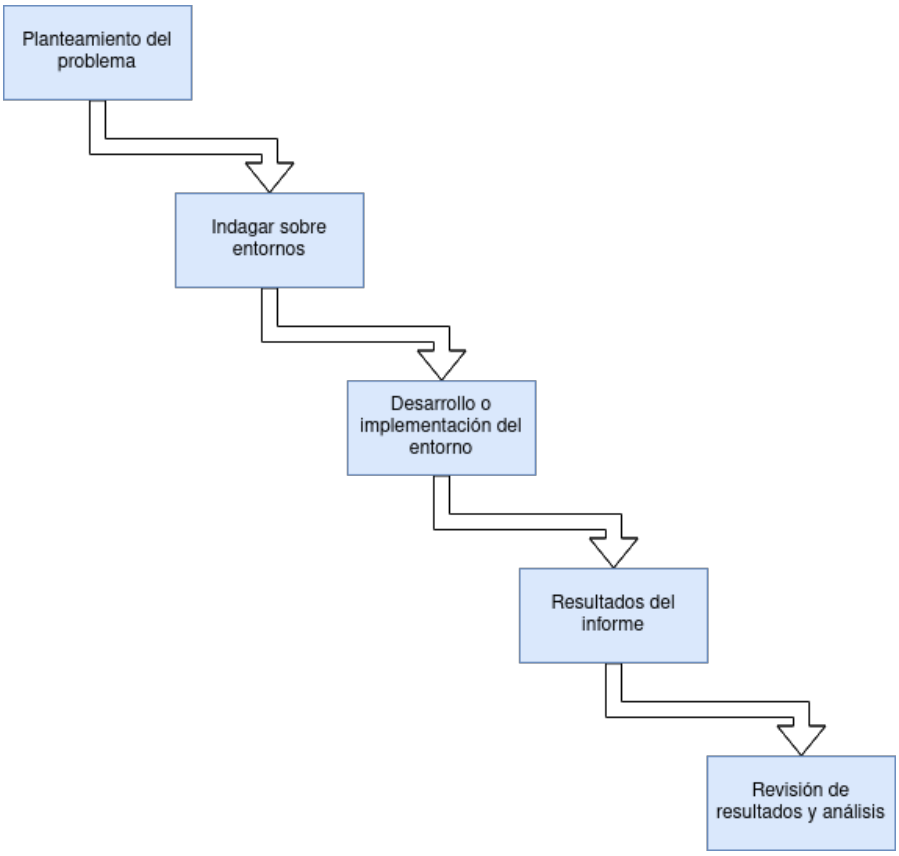


Figura 1.1: Adaptación del método científico a las etapas del proyecto (Elaboración propia)

en formato tesis.

- **Revisión de resultados y análisis**: Análisis de los resultados: Los resultados obtenidos mediante pruebas se analizarán de acuerdo con:
  - El agente debe aprender a jugar
  - El agente sea capaz de terminar el juego de forma correcta
  - La probabilidad de éxito obtenida por el agente debe ser acorde a los valores entregados por el entorno.

1.6. Alcances

- El agente que se desarrollará tiene como objetivo aprender, mediante ensayo y error, como jugar y las reglas que componen el juego.
- El entorno elegido tiene la particularidad que está basado en competencias.

Se optó por este entorno debido a que la facilidad para la implementación del enfoque basado en introspección que solo se ha probado en entornos cerrados y limitados. De esta manera, se puede evaluar este tipo de enfoque bajo otra perspectiva.

- Se determinará la interacción entre humano y agente en el proceso del término del juego.

1.6.1. Limitaciones

- Solo se podrá implementar librerías o Framework que estén escritos en el lenguaje de programación Python.
- Las versiones de Python necesarias para la implementación y uso de las librerías tendrán que ser superior a la versión 3.0. No se podrá usar versiones de Python anteriores debido a su inestabilidad e incompatibilidad con las herramientas usadas en este proyecto.
- El enfoque basado en introspección sólo se podrá implementar en un entorno competitivo.

1.6.2. Factibilidad

Puede suceder que, al implementar los agentes artificiales, el medio con el cual se cuenta para realizar los experimentos no posea los recursos físicos necesarios (hardware) y acordes o no posea un rendimiento adecuado a lo esperado. Sin embargo, los recursos físicos con los que se cuentan actualmente hacen posible el desarrollo del proyecto sin mayores dificultades.

1.6.3. Medios

- Como herramienta principal para el desarrollo y la ejecución de los algoritmos se hará uso de un computador personal. La marca del computador personal es Apple, modelo Macbook Pro de mediados del 2012.
- Se utilizará un editor de lenguaje de programación llamado PyCharm de Jetbrains para la codificación de los algoritmos.
- Las librerías elegidas para desarrollar e implementar al agente son Tensor Flow, Numpy, Keras y Matplotlib.
- El lenguaje de programación elegido para el desarrollo del proyecto es Python.

- El entorno elegido fue desarrollado gracias a la ayuda de un Framework de entornos como es OpenIA y su aplicación Gym.
- Para el almacenamiento de datos, imágenes y el formulario de inscripción de proyecto se utilizará Google Drive.

Capítulo 2

Marco Teórico y Estado del Arte

Esta área llamada inteligencia artificial, proviene de las ciencias de la computación, está compuesta por diferentes tópicos, siendo el aprendizaje por refuerzo explicable y el respectivo enfoque basado en introspección a implementar los ejes principales de este trabajo. En esta parte, se abordaran los cimientos teóricos de cada uno de estos temas mencionados anteriormente. En primera instancia, se abordan los fundamentos de la inteligencia artificial y una rama que deriva de ella, la inteligencia artificial explicable. Después de esto, se explicarán conceptos claves como el aprendizaje por refuerzo y el aprendizaje por refuerzo explicable. De esto, se desprende la explicación teórica de un método propuesto basado en introspección y que tiene sus bases en el aprendizaje por refuerzo.

2.1. Inteligencia artificial

La inteligencia artificial es una rama de las ciencias de la computación y tiene por objeto que los ordenadores hagan las misma clase de cosas que puede hacer la mente humana (Boden, 2017). Otra definición, caracterizada por su generalidad en los conceptos que incluye es la de Patrick Henry (1992) que define que la inteligencia artificial es "el estudio de la computación que hace posible que computadores perciban, razones y actúen". Si analizamos bien esta definición, podemos darnos cuenta que la inteligencia artificial extrae conceptos de la psicología y la computación a la misma vez.

La inteligencia artificial no solo posee sus bases en elementos provenientes de la psicología o la computación, aunque a simple vista así lo pareciera. Esta área de las ciencias de la computación considera como ejes fundacionales disciplinas

tales como matemáticas, neurociencias, filosofía, economía, lingüística e ingeniería computacional.

La I.A. o inteligencia artificial es un campo relativamente nuevo. La mayor parte de las investigaciones relacionadas con este campo datan posteriormente a la culminación de la segunda guerra mundial en el año 1945 (Russell y Norvig, 2009), es decir, esta área sólo lleva 8 décadas de vida. Si lo comparamos con otras disciplinas, como la física o las matemáticas, el tiempo de vida es absolutamente mínimo.

El primer trabajo reconocido como inteligencia artificial fue hecho por Warren McCulloch y Walter Pitts (Russell y Norvig, 2009). Su trabajo se basó en una propuesta de modelo de neurona artificial que podía cambiar su estado de encendido o apagado de acuerdo a una señal de entrada. Con este trabajo, los autores pudieron demostrar que cualquier función computable podría ser calculada por alguna red de neuronas interconectadas (Russell y Norvig, 2009). Más tarde, en 1924 Donald Hebb propuso una sencilla regla para actualizar la intensidad (valores) de las conexiones entre neuronas (Russell y Norvig, 2004).

Se han escrito cientos de libros científicos y de ciencia ficción en todos los aspectos desde que este tema surgió. Debido a los avances de las investigaciones realizadas en inteligencia artificial, se ha podido llegar al borde de la era de los robots autónomos y el internet de las cosas (Ertel, 2017), ampliando así los horizontes de las futuras aplicaciones que podría tener este campo en las sociedades del futuro.

2.1.1. Inteligencia artificial explicable

En la sección anterior, se abordó la inteligencia artificial desde un punto de vista general. Partiendo desde sus primeros exponentes, con sus respectivos trabajos y aportes, hasta las implicancias que este campo podría tener en la sociedad del futuro. De acuerdo, al avance en el área, más escepticismo se genera por parte de las personas que no son expertas en esta área, debido al desconocimiento del funcionamiento de todos estos algoritmos de inteligencia artificial y como estos toman las decisiones que toman.

A pesar de que estos algoritmos funcionan poderosamente hablando en término de predicciones y resultados, resulta muy difícil tener información sobre los mecanismos internos de trabajo (Adadi y Berrada, 2018).

La inteligencia artificial explicable o X.A.I (por sus siglas en inglés) viene a aportar elementos de transparencia y confiabilidad a todos los modelos que carecen de ello. Son nuevos algoritmos que establecen la transparencia como nuevo enfoque.

Para abordar estos problemas, la inteligencia artificial explicable propone hacer un cambio hacia una inteligencia artificial más confiable. El objetivo es desarrollar técnicas que produzcan más modelos explicables manteniendo los niveles de rendimiento.

Tal como se suele ver en los sistemas de caja negra, a un usuario no nativo le resultaría casi imposible saber cómo funcionan exactamente estos tipos de sistemas. De todas formas, en la vida diaria, cuando surge algún error en el sistema de reconocimiento facial de la cámara del teléfono o que la predicción de la siguiente palabra no sea lo bastante buena al momento de escribir un mensaje de texto, no son tareas cruciales para las personas. La foto se obtendrá de todas formas o el mensaje se enviará con o sin el sistema predictivo de palabras. Esas tareas se pueden realizar de cualquier manera. Sin embargo, estos márgenes de error no pueden ocurrir así en la industria o en los robots domésticos. Un ejemplo a futuro, las tareas del cuidado de un adulto mayor que se le confían a un robot doméstico carecen totalmente de márgenes de error, debido a la complejidad del escenario. O incluso, una mala decisión de un automóvil de conducción autónoma puede resultar peligroso para la vida o la salud de cualquier ser humano (Samek y Muller, 2019). Cualquier error puede costar vidas que son irremediablemente irreversibles. Debido a esto, es que no se pueden confiar decisiones tan importantes a sistemas que no pueden explicarse por sí mismos (Adadi y Berrada, 2018).

2.1.2. Aprendizaje por refuerzo

El aprendizaje por refuerzo es una sub área perteneciente a la disciplina de la inteligencia artificial y trata principalmente de que un agente aprenda que hacer para maximizar una señal de recompensa numérica obtenida proveniente del entorno (Sutton y Barto, 2018). De forma general, la dinámica que ocurre entre el agente y el entorno en el que se encuentra, se puede apreciar de mejor forma en la figura 2.

A lo largo del último siglo, han surgido muchas teorías sobre el aprendizaje en los animales y seres humanos. Una de ellas es el aprendizaje en teoría conductual. El aprendizaje según esta teoría propuesta por Papalia y Wendkos (Arancibia, Herrera y Strasser, 2007) se define como un cambio relativamente permanente en el comportamiento, que refleja una adquisición de conocimiento y habilidades a través de la experiencia. En otras palabras, estos cambios generados en el comportamiento debe ser razonablemente objetivos, por lo tanto, deben ser posible medirlos.

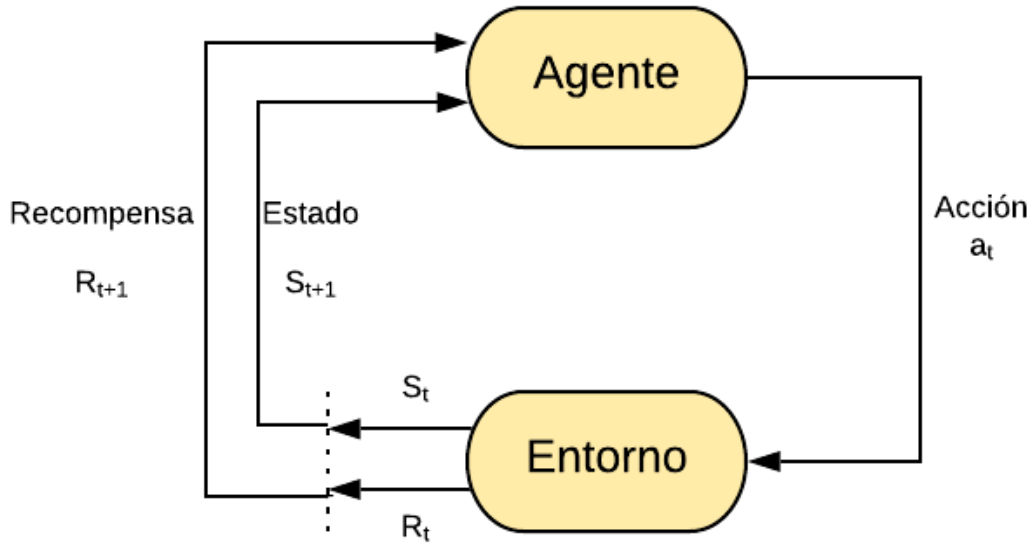


Figura 2.1: Esquema representa la interacción de un agente con su entorno. Un agente toma una acción at, inmediatamente el entorno entrega al agente una recompensa basada en una recompensa pasada y le entrega el siguiente estado al cual el agente llegó. Figura obtenida desde (Sutton y Barto, 2018).

El aprendizaje por refuerzo está basado en intentar una acción y observar qué pasa en el entorno (Cruz, 2017), lo que se asimilaría de forma práctica al enfoque conexionista propuesto por Edward Thorndike. Este autor, plantea que la forma más característica de aprendizaje, tanto en animales como en seres humanos, se produce mediante ensayo y error. Si algún conjunto de acciones tomadas por el agente conducen a mejores situaciones, existe una tendencia a repetir el comportamiento nuevamente, en caso contrario, la tendencia es evitar tales comportamientos en un futuro, por lo tanto, el problema se puede reducir a que un agente aprenda a seleccionar las acciones óptimas en cada caso para poder así alcanzar un objetivo general (Rieser y Lemon, 2011).

Un agente que interactúa con su entorno debe determinar una forma de actuar en una situación determinada (Cruz, Dazeley y Vamplew, 2020). Es decir, un agente debe considerar aprender una política óptima para su ambiente. En primera instancia, se define como política la siguiente forma $\pi : S \rightarrow A$ con S como conjunto inicial de estados y A conjunto de acciones disponibles desde S.

Las funciones de política y acción-valor óptima se define como π y $q*$ respectiva-

mente:

$$q*(s_t, a_t) = q*(s, a) \tag{2.1}$$

La función óptima de acción-valor como se muestra en la ecuación 2.1 Esta ecuación tiene como solución la ecuación de optimalidad de Bellman representada en la ecuación 3.

$$q*(s_t, a_t) = \sum_{s_t+1} p(s_{t+1}|s_t, a_t)[r(s_t, a_t, s_{t+1}) + \gamma q*(s, a)] \tag{2.2}$$

Donde s_t representa el estado actual del agente, a_t la acción realizada, r la recompensa numérica recibida por el agente después de realizar la acción a_t en el estado s_t para alcanzar el estado s_{t+1}, y siendo a_{t+1} una posible acción para el estado s_{t+1}. En el caso de p, se define como la probabilidad de alcanzar s_{t+1} dado el estado actual del agente en s_t y que se haya tomado la acción a_t.

Una forma de resolver la ecuación 2.2, es usar un método llamado SARSA. Este método de aprendizaje actualiza iterativamente los valores de estado-acción Q(s,a) usando la siguiente ecuación (Sutton y Barto, 2018) :

$$q(s, a) = q(s, a) + \alpha[r_{t+1} + \gamma[q(s_{t+1}, a_{t+1}) - q(s, a)]] \tag{2.3}$$

2.1.3. Elementos del aprendizaje por refuerzo

En teoría del aprendizaje por refuerzo, se pueden identificar cuatro elementos principales, además del agente y el ambiente en el que se encuentra inmerso:

1. **Una política**
2. **Señal de recompensa**
3. **Función de valor**
4. **Modelo del entorno**

2.1.4. Política

La política de una agente en aprendizaje por refuerzo determina la forma de actuar del agente en un momento determinado (Sutton y Barto, 2018). La política logra mapear los estados provenientes del ambiente y el conjunto de acciones que un

agente puede tomar, cuando el agente se encuentra en dicho estado. La política es considerada el núcleo de un agente de aprendizaje por refuerzo debido a que solo basta con este elemento para determinar el comportamiento (Sutton y Barto, 2018). La naturaleza de la política puede ser estocástica o determinista.

2.1.5. Señal de recompensa

Este elemento determina la meta de un problema. El objetivo de un agente es maximizar la ganancia total que obtiene a largo plazo. La señal de recompensa se considera el elemento que determina la definición de la política (Sutton y Barto, 2018).

2.1.6. Función de valor

Esta función en simples palabras, indica lo que es bueno para el agente a largo plazo. La ponderación que se le da a un estado, es la cantidad total que un agente puede acumular en el futuro a partir del estado actual. A diferencia de las recompensas provienen de las señales de recompensas, los valores de esta función deben ser estimados desde el ambiente. Es decir, los valores deben ser estimados y reestimados desde la secuencia de acciones que el agente hace mientras se encuentra en el entorno.

2.1.7. Modelo del entorno

El modelo del entorno es algo que imita el comportamiento del ambiente en el cual el agente se encuentra inmerso. Los modelos se utilizan para planificar las posibles acciones, todo esto considerando las probables situaciones futuras que se presenten mucho antes que ocurran (Cruz, 2017).

2.2. Proceso de decisión markoviano finito

El proceso de decisión de Markov o MDP (por sus siglas en inglés) es una formalización clásica de la secuencia de la toma de decisiones, donde las acciones no solo influyen en recompensas inmediatas, sino también en situaciones o estados posteriores (Sutton y Barto, 2018). Este procesos requiere como argumentos un estado s_t actual en donde el agente se encuentra inicialmente, una acción posible a_t para dicho estado y finalmente una recompensa r_{t+1} lograda en función de la acción tomada con anterioridad. Dicho de una manera más formal, el agente y el ambien-

te interactúan en una secuencia discreta de pasos de tiempo $t = 1, 2, 3, \ldots, n$ en cada paso t recibe información sobre el estado del ambiente $S_t \rightarrow S$ y una acción $A_t \rightarrow A(s)$. Posteriormente, el agente recibe una recompensa numérica (positiva o negativa) R_{t+1}, lo que en consecuencia, lograría encontrar el siguiente estado S_{t+1}.

- S es un conjunto finito de estados
- A es un conjunto de acciones
- V es el valor de la función de transición
- r es el valor de la recompensa recibida por el agente

En un proceso de decisión de markov, todos los conjuntos de acciones, recompensas y estados poseen un número finito de elementos. La dinámica que se produce entre todos los finitos elementos se puede modelar en forma de distribución probabilística de acuerdo a la siguiente ecuación:

$$p(s, r|s, a) = pr(S_t = s, R_t = r|S_{t-1} = s, A_{t-1} = a) \tag{2.4}$$

2.3. Aprendizaje por refuerzo explicable

Ya en la sección sobre inteligencia artificial explicable se hizo hincapié en la problemática que se desea abordar con este nuevo enfoque. La transparencia y la confiabilidad en sistemas basados en inteligencia artificial se volvieron un asunto de prioridad para algunos sectores empresariales y académicos. En esta sección se abordará una rama relativamente nueva de la inteligencia artificial explicable llamada aprendizaje por refuerzo explicable o por sus siglas en inglés X.R.L (Explainable Reinforcement Learning).

Algunas personas se preguntarán ¿Por qué es tan importante y crucial la explicabilidad y la transparencia?, y la respuesta es muy sencilla, si un usuario no confía en un sistema o modelo de predicción, simplemente no usa ese sistema. Esto responde a una reacción psicológica humana. Es por esto que en la actualidad se considera un problema abierto la carencia de algoritmos que les permitan comunicar claramente las razones de por que toman ciertas decisiones en determinados momentos (Adadi y Berrada, 2018). Es por esta sencilla razón que elementos como la transparencia, confiabilidad y explicabilidad se vuelven tan importantes a la hora de desarrollar un sistema basado en inteligencia artificial.

2.4. Enfoque basado en introspección

Debido a que la inteligencia artificial explicable, y más precisamente el aprendizaje por refuerzo explicable son áreas de estudio prometedores, y que significarán principalmente un marco regulatorio para el desarrollo de futuros algoritmos de I.A, es que se han desarrollados diferentes enfoques para abordar esta problemática. Es por eso, que en este trabajo, se incluirá como parte del desarrollo de un agente de aprendizaje por refuerzo un enfoque basado en introspección.

El enfoque abordado, basado en introspección, permite estimar la probabilidad de éxito P_s directamente de los valores provenientes de $Q - values$ usando una transformación numérica. La idea principal que sostiene este enfoque es poder relacionar los valores de $Q - values$ hacia la probabilidad de éxito como medio de introspección desde la auto motivación del agente (Cruz, Dazeley y Vamplew, 2020).

Como se pudo entender en la sección de aprendizaje por refuerzo y considerando el enfoque de aprendizaje basado en diferencia temporal, los valores entregados por la ecuación 4, representan las futuras recompensas. Por lo tanto, si un agente alcanza un estado terminal en una tarea episódica al obtener una recompensa R^T, el Q-values asociado se aproxima a esta recompensa (Cruz, Dazeley y Vamplew, 2020).

Esto cualquier caso, una manera simple es considerando cualquier valor de $Q(s, a)$ como una recompensa final R^T, multiplicado n por el factor de descuento , mostrado en la ecuación 5:

$$Q(s, a) \approx R^T * \gamma^n \tag{2.5}$$

Gracias a la ecuación 2.5, se puede aplicar una transformación logarítmica, llegando a la siguiente expresión:

$$n \log Q(s, a) * R^T \tag{2.6}$$

Donde $Q(s, a)$ son los valores para $Q - values$, R^T es la recompensa obtenida cuando la tarea se ha completado exitosamente y n es la distancia estimada, en número de acción, hasta la recompensa (Cruz, Dazeley y Vamplew, 2020).

2.5. Estado del arte

Los avances en investigaciones en el área del aprendizaje por refuerzo han crecido de forma exponencial en estos últimos años. Y con ello, los ambientes en donde son aplicados todos estos avances. Un ejemplo de ello es el desarrollo y la implementación de un entorno de simulación virtual para el juego El sombrero del chef. Juego completamente centrado en promover la interacción humano-robot y donde es posible incluir dentro del juego agentes de aprendizaje por refuerzo (Barros et al., 2020).

Otro trabajo atractivo realizado es el de extraer "elementos interesantes" del aprendizaje por refuerzo mediante introspección. Gracias a un conjunto de interacciones almacenadas en la memoria del agente, se puede aplicar introspección de 3 niveles (análisis del ambiente, análisis de las interacciones y los meta análisis). Gracias a este proceso, se pueden extraer elementos que faciliten la explicación del comportamiento del agente (Pedro Sequeira, Eric Yeh y Melinda Gervasio, 2019).

Finalmente, algunos trabajos están centrados en entender el mundo por medio de relaciones causales. Estos modelos ofrecen a un agente la capacidad de considerar eventos no ocurridos. En este trabajo, los autores presentan un enfoque de aprendizaje que aprende del mundo exterior modelos estructurales causales durante el proceso de aprendizaje del agente (Prashan Madumal et al., 2019).

Capítulo 3

Escenario experimental

3.1. Recopilación de información

Desde un inicio, el primer problema que aqueja a realizar un trabajo de investigación surge desde el origen de las ideas. Las ideas no son claras desde un principio, otras de alguna manera no son viables, y así se va develando a medida que se trabaja en la recopilación del material inicial (Sampieri, Collado y Lucio, 2014). Todo esto se debe también al escaso conocimiento sobre el tema a tratar. De este modo, cuando una persona ya concibe una idea clara de investigación, debe familiarizarse completamente con el campo de conocimiento en el que se centra la idea (Sampieri, Collado y Lucio, 2014).

Siguiendo los pasos planteados en la metodología de trabajo y en los objetivos específicos, en primera instancia se hizo efectivo la recopilación de una variada literatura del área en cuestión. Primero se acudió a obtener la literatura considerada base teórica para ahondar en el campo de la inteligencia artificial y el aprendizaje por refuerzo como son los libros de inteligencia artificial: un enfoque moderno y aprendizaje por refuerzo: Una introducción.

Ya conociendo la estructura base de los temas a tratar, se procedió a la recopilación y análisis de un conjunto de trabajos de investigación relacionados con una nueva tendencia en estas áreas llamada explicabilidad.

3.2. Modelo del entorno

El entorno elegido para la implementación del agente de aprendizaje por refuerzo con su respectivo enfoque basado en introspección reúne todas las cualidades necesarias de los que otros entornos carecen. La elección de este entorno se debe principalmente a que fue desarrollado tomando en cuenta ciertos requerimientos que algunos de juego no contaban como por ejemplo contar con reglas fáciles de seguir y entender o que el número de acciones sea pequeña y fácil de procesar. Este juego provee de una plataforma donde diferentes algoritmos de aprendizaje para la toma de decisiones puedan ser integrados, desplegados y evaluados correctamente (Barros, C. Bloem y Barakova, 2020).

3.2.1. Entorno *El sombrero del chef*

El juego del sombrero del chef es un juego basado en cartas. Solo está permitido cuatro jugadores como máximo por mesa. La elección de este juego nos asegura situaciones controlables de acuerdo con su mecánica de turnos. De lo anterior, el agente cuenta con un turno cada cierto momento en el cual, puede tomar una cierta acción y de esta forma interactuar no solo con su entorno, sino que también con los otros jugadores. Este es un entorno donde cada acción tomada por un jugador afecta directa o indirectamente a los demás jugadores en las etapas posteriores del juego.

3.2.2. Elementos del juego

3.2.2.1. Tablero

El campo de juego (entorno) es representado por una tabla de madera y una base de pizza (masa y salsa de tomate). Este campo cuenta con una capacidad de 11 espacios libres donde cada jugador, en su respectivo turno, deposita n cantidad de cartas permitidas por las reglas del juego (movimientos y acciones posibles se ven con más detalle en la sección "Mecánica del juego").

En la figura 3.1 se puede apreciar de manera gráfica el campo de juego. Estos campos se van llenado en la medida que cada jugador deposita sus cartas. Al momento de que cada jugador no posea más opciones para jugar, el área de juego es despejada, lo que significa que la pizza logro completarse con los ingredientes elegidos hasta ese punto de la ronda.

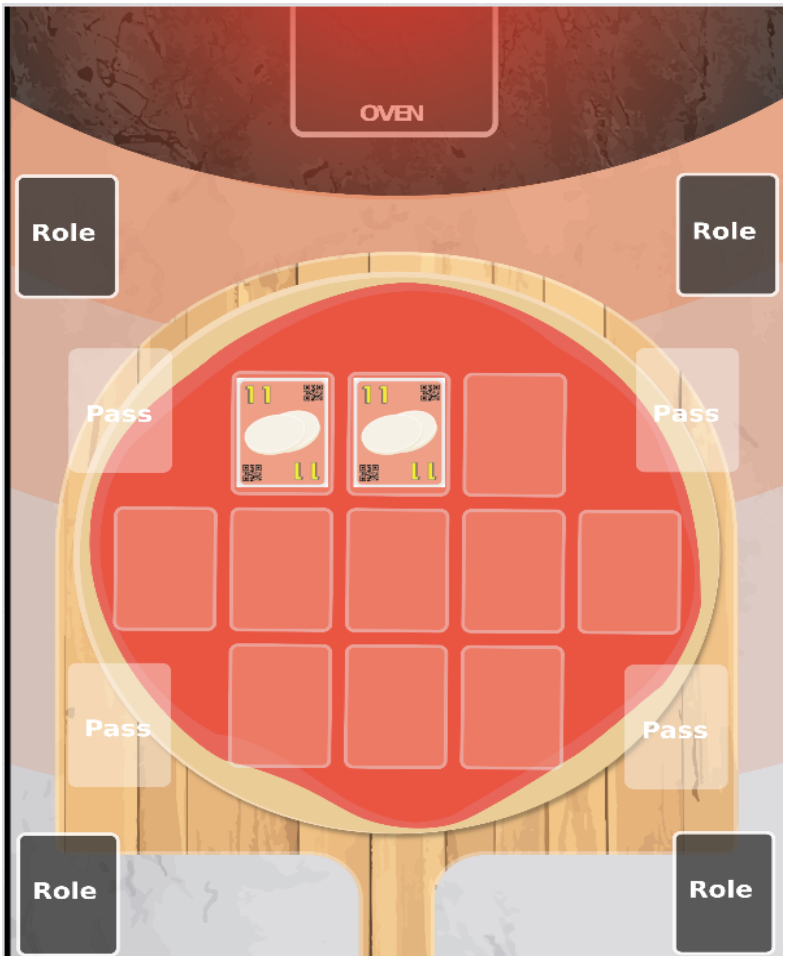


Figura 3.1: Tablero principal donde se lleva a cabo el juego. Los cuadros blancos donde depositan las cartas los jugadores. El recuadro Role es donde aparecerá después del primer turno, el rol de cada uno de los jugadores dentro del juego (Chef, Sous-Chef, camarero y el lavaplatos) (Barros, Tanevska y Sciutti, 2020).

3.2.2.2. Cartas

Adicional a al tablero de juego en donde lo que se buscar es completar una pizza, también existen ingredientes que completan la pizza. En el juego existen 68 cartas, de las cuales, existen 11 diferentes tipos. En la tabla 3.1 se puede apreciar los valores para cada una de las cartas, el ingrediente asociado a ese valor de carta y su cantidad dentro del juego (número copias de cada ingrediente).

Valor	Cantidad	Ingredientes
1	1x	Ají
2	2x	Camarón
3	3x	Albahaca
4	4x	Gorgonzola
5	5x	Pimentón
6	6x	Aceitunas
7	7x	Salame
8	8x	Champiñón
9	9x	Cebolla
10	10x	Tomate
11	11x	Mozarela
Joker	2x	Joker

Tabla 3.1: Descripción de los valores y las cantidades de cartas que componen el juego (Elaboración propia).

Cada ingrediente posee un valor diferente al otro. Los valores parten con el valor mínimo representado por el ají, valor 1 y terminan con el valor mayor representado por la mozzarela llegando a un valor de 11. Para cada valor de carta, existe un número de copias. Ejemplo, para el camarón con valor 2, solo existen 2 copias dentro del juego. Adicionalmente a los ingredientes, se reparten 2 cartas mas llamadas joker. Cada joker tiene la función de reemplazar cualquier carta sobre el campo de juego.

Figura 3.2: Cartas del juego representadas por ingredientes y sus respectivos valores (Barros, Tanevska y Sciutti, 2020)

3.2.2.3. Roles

Adicional a las cartas y al tablero de juego, existe una jerarquía que es por la que todos los jugadores compiten. Alcanzar el sombrero del chef es considerado el objetivo máximo del juego. El ganador del juego, es decir, el jugador que consiga el puntaje máximo, se consagrará ganando el sombrero del chef. En segundo lugar, se encuentra el premio del sombrero del Sous-chef. En tercer lugar, se encuentra el rol de camarero y finalmente, el jugador con menos puntos de los 4, se lleva el rol de lavar los platos de la cocina. Los roles no son distribuidos entre los jugadores si no hasta que se gana el primer juego. El ganador del primero juego, se le asigna el sombrero del chef.

3.2.3. Mecánica del juego

El juego busca simular la dinámica que ocurre en una cocina de un restaurante donde se preparan pizzas. Dicha cocina posee una jerarquía base donde cada jugador puede ser chef, sous-Chef, camarero o la persona encargada de lavar los platos (Barros et al, 2020). Los jugadores buscan descartarse todas las cartas de su mano y así poder conseguir la distinción de chef (al conseguirlo, se lleva la puntuación máxima de la ronda). El juego termina cuando uno de los 4 jugadores logra alcanzar un total de 15 puntos. Esta dinámica general se puede apreciar con más detalle el flujo del juego en el algoritmo 1.

Como se puede ver en el algoritmo 3.1, el juego comienza con la distribución de las cartas de forma aleatoria. Se reparten la misma cantidad de cartas para cada jugador (un total de 17 por jugador). El jugador que cuente con los dos jokers o un mozzarela de oro al finalizar la distribución de cartas, comienza con la ronda. De lo anterior, es considerado acciones especiales.

Cada jugador tiene dos posibilidades al momento de iniciar su turno, o se descarta las cartas compatibles con una acción valida o pasa. Cuando un jugador pasa, se entiende que no quiere agregar más ingredientes a la pizza o no posee cartas para cometer una jugada válida.

Una acción permitida es aquella donde un jugador se descarta n cantidad cartas dentro del tablero, siempre y cuando las cartas en el tablero de juego sean mayores en valor y en número de cartas. Por ejemplo, al comenzar un turno de un jugador, el tablero cuenta con 3 copias de tomate (valor 10), por consiguiente, el jugador actual no podrá descartarse cartas mayores o iguales al valor de las cartas del tablero. Una acción válida seria descartarse cualquier carta menor al valor del

tomate (cartas con valores 1,2,3,...,9). Si en el tablero hay 4 copias de la carta tomate, el jugador se puede descartar 4 o más copias de una carta de valor más bajo.

Algorithm 3.1. Flujo de juego del *El sombrero del chef* (Barros et al, 2020)

```
Barajar las cartas
Repartir una cantidad igual de cartas por jugador
Intercambiar roles
Intercambiar cartas
if Es ejecutada la acción especial then
    Hacer acción especial
end if
Primer jugador descarta sus cartas
while No sea el final del turno do
    for Cada jugador do
        if El jugador quiere y puede descartar then
        else
            Pasa
        end if
        if Todos los jugadores pasan then
            Limpiar el tablero
        end if
        if Todos los jugadores terminaron then
            Termino de la ronda
        end if
    end for
end while
```

3.2.4. Entorno Gym de OpenIA

Gym es considerado un framework que busca ser el lugar donde los desarrolladores e investigadores puedan utilizar y evaluar sus algoritmos en ambientes completamente simulados. El núcleo del juego del sombrero del chef implementa herramientas desarrolladas por este framework. Esta librería es de distribución gratis y de libre uso para quien desea utilizarla en sus experimentos.

3.2.5. Configuraciones del entorno

Como se explicó y se profundizó en en el capítulo *escenario experimental*, el entorno seleccionado cumple con ciertas características que lo hacían idóneo para la implementación de la enfoque propuesto, ya que el enfoque basado en introspección no ha sido probado en entornos competitivos. Las características requeridas y necesarias determinadas por el entorno para poder implementar el agente basado en introspección son las siguientes (Barros, Tanevska y Sciutti, 2020):

- El juego cuenta con la posibilidad de crear estrategias adaptables en base a la acción del contrincante anterior. La interacción con los otros jugadores es parte del flujo natural del juego
- El juego debe tener una mecánica específica
- El juego debe ser fácil de entender, lo que implica que los turnos sean claros entre cada agente. El número de acciones es reducido, lo que deriva en un procesamiento ligero de los datos
- El juego otorga la oportunidad suficiente para interactuar con otros jugadores por medio de la mecánica del juego. Las acciones tomadas deben limitar el accionar del siguiente jugador
- Los estados y las acciones son discretas. Esto facilita en mayor medida el procesamiento de los datos.

El juego de cartas del sombrero del chef está implementado a través de un entorno de simulación basado en OpenAI (Barros, Tanevska y Sciutti, 2020). Gracias al uso de este framework, es posible utilizar funciones pre desarrolladas que ayudan a agilizar la configuración del entorno.

La mecánica del juego esta representada por los estados y acciones de cada jugador. Los estados son representados por las cartas que el jugador posee en la mano y las cartas representadas en el tablero, lo que en total suman 28 cartas (11 cartas en el tablero de juego más las 17 cartas de la mano del jugador).

Para el caso de las acciones posibles de cada jugador, estas están representadas por un total de 200 acciones. Las acciones permitidas se encuentran directamente relacionadas con las cartas en el tablero. El jugador solo puede hacer una acción por turno.

Las acciones y los estados del juego los suministra directamente el entorno en forma de lista o arreglo en python. La lista tiene un nombre de *observación*. Esta

lista esta compuesta por un total de 228 posiciones. Las primeras 11 posiciones representan los estados del tablero del juego (0 a 10). Las siguientes 17 posiciones, corresponden a las cartas en la mano de cada jugador (11 al 28) y finalmente, las 200 posiciones restantes, corresponden a las acciones posibles por el jugador.

Las acciones aceptables por cada jugador en cada uno de sus turnos son procesadas en forma de función, donde la función recibe como parámetro principal la observación entregada por el entorno. La visualización de las jugadas permitidas se pueden ver en formato de matriz, donde las columnas equivalen a los valores que se les asigna a cada carta y las filas representan la cantidad de cartas que se pueden dejar en el tablero más las 2 cartas joker. (ver tabla 3.1)

Las recompensas entregadas por el entorno están configuradas de la siguiente manera:

- Por cada jugada, el agente recibe una recompensa negativa de *-0.001.*
- Cuando el jugador logra descartarse el total de cartas de la mano, gana la ronda y pon ende, recibe una recompensa de *1.*

Capítulo 4

Diseño e implementación de los agentes

En la presente sección del trabajo de tesis, se encontrarán con información general y detallada de las configuraciones realizadas en los agentes implementados y el entorno. Se conocerán los parámetros usados por cada agente y los respectivos tipos de algoritmos utilizados para cada uno de ellos.

4.1. Agentes implementados

Los agentes artificiales son parte del esquema general del aprendizaje por refuerzo. Juegan un rol fundamental. El aprendizaje por medio de la interacción con el entorno se encuentra arraigado en áreas como la psicología o la neurociencia. Estas disciplinas generan un relato sobre el comportamiento de los animales y de como estos poseen la capacidad de optimizar su control general en un entorno (Mnih1 et al, 2015).

Para la validación del entorno y del método propuesto en este trabajo (enfoque basado en introspección), se implementaron 2 tipos de agentes diferentes basados en TD (temporal difference). Los otros dos agentes adicionales están desarrollados para que funcionen de manera aleatoria, lo que se le conoce como agentes dummy. Los agentes (DQN y PPO)fueron implementados usando una librería de código abierto llamada Keras para crear redes neuronales. Esta librería se uso generalmente como interfaz entre tensorflow y el usuario para la configuración de redes neuronales.

4.1.1. Agente DQN

Deep Q-learning fue una revelación por el año 2015, año de su publicación, ya que los autores tenían una tarea ambiciosa y era desarrollar un algoritmo que pudiera desarrollar una amplia forma de competencia para una amplia variedad de tareas (Mnih1 et al, 2015). El algoritmo de Deep q-learning es considerado una extensión al ya clásico Q-learning (Fan et al, 2020). El algoritmo de Deep Q-learning usa redes neuronales profundas para poder aproximar mediante iteraciones, los valores de la función acción-valor del agente (Fan et al, 2020). Adicional a lo anterior, este algoritmo introduce además dos características principales las cuales son: experience replay y target model. El experiencce replay brinda estabilidad a la red neuronal profunda. Para verlo de otra forma, este tipo de redes neuronales usan lotes de memoria donde almacenan la trayectoria que toman los valores del proceso de decisión markoviano (Fan et al, 2020). Por otra parte, el target model contribuye a estabilizar el aprendizaje de los valores Q, entregando una estimación de los Q-values estable durante el proceso de entrenamiento (Barros, Tanevska y Sciutti, 2020).

Se puede apreciar en la figura 4.1 la arquitectura correspondiente al agente DQN. En el presente trabajo, se implementó un algoritmo de Deep Q-Network uno de los agentes propuestos debido a la complejidad en las transiciones del juego y al costo computacional que tendría el procesar el agente para obtener los Q-values para cada par estado-acción. La adaptación del algoritmo DQN al entorno de Gym El sombrero del chef se debe al trabajo realizado por Pablo barros en el paper **The Chef's Hat Simulation Environment for Reinforcement-Learning-Based Agents** (Barros, Tanevska y Sciutti, 2020).

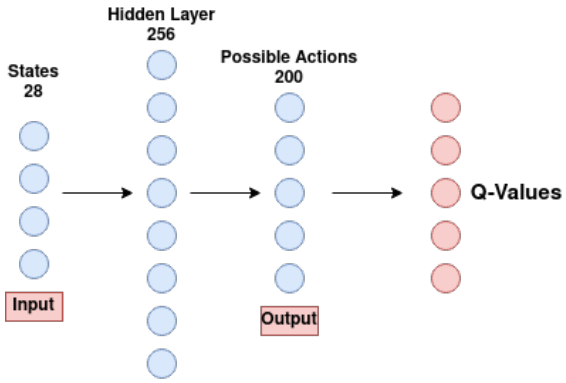


Figura 4.1: Arquitectura del agente DQN (Elaboración propia)

A continuación, se detalla la configuración de los parámetros de la red neuronal del agente DQN:

Tabla 4.1: Configuración agente DQN

Capas	**Neuronas**	**Función de activación**
Capa entrada	28	ReLU
Capa dense	256	ReLU
Capa salida	200	Softmax

4.1.2. Agente PPO

Para el caso del algoritmo PPO, este hace uso de una familia de métodos de optimización de políticas que utilizan múltiples épocas de ascenso del gradiente estocástico para realizar cada actualización de la política (Schulman et al, 2017)

En términos generales, este nuevo algoritmo implementa un control adaptativo de penalizaciones, basado en la divergencia KullbackLeibler, para impulsar las actualizaciones del agente en cada interacción (Barros, Tanevska y Sciutti, 2020), permitiendo que el modelo cree una región de actualización que funcione de manera similar a la optimización del descenso del gradiente estocástico, simplificando el uso y evitando el uso de ciertas estructuras de memorias que complejizan la actualización de reglas.

4.1.3. Agente Random

El agente que viene por defecto contenido en el entorno es el agente llamado random (aleatorio) o dummy. Este agente no resiste mayor análisis. La principal cualidad del agente es realiza acciones dentro del entorno de forma aleatoria, independiente si el agente se encuentra ganando o perdiendo, si tiene mas cartas en su mano que los contrincantes.

4.2. Configuraciones

Los agentes deben interactuar con su entorno para poder recibir una recompensa. Para el proceso de entrenamiento del agente, debe existir una configuración de parámetros y número de épocas definas con antelación. Parámetros como gamma,

alpha, numero de épocas de entrenamientos, o en este caso en particular, número de juegos en el que el agente participa, son sumamente importantes a la hora de definir con anterioridad antes de llevar a cabo el proceso de entrenamiento.

Parámetros	**Valores**
alfa	0.5
gamma	0.9
épsilon	0.7

Tabla 4.2: Configuración de parámetros para los agentes DQN y PPO

En el caso de la distribución del número de juegos, se estableció como cantidad de juegos un número inicial de 10, luego seguido de 30, 50 y hasta llegar a los 100 juegos. El juego poseía una mecánica que permitía jugarlo de dos formas, una por puntos y otro por juegos. El experimento solo contemplo el término de cada juego por agente, ya que para efectos de los resultados, era necesario si el agente llegaba a completar y ganar un juego, más que el puntaje entregado.

Capítulo 5

Resultados Experimentales

En esta sección del presente trabajo de tesis, se abordarán los resultados obtenidos gracias a diferentes pruebas hechas a los agentes en el juego. Se configuraron diferentes tipos de pruebas, todas ellas relacionadas con el número de juegos y épocas de cada prueba.

Las pruebas realizadas corresponden a una configuración del entorno definida. El número de épocas se decidió mantener en el mínimo debido a que el flujo del juego es dinámico, ya que esta compuesto por un número grande de jugadores (4), y las rondas pueden extenderse o pueden ser breves, dependiendo de la estrategia que adopte cada jugador. El juego puede, al finalizar, terminar con una enorme cantidad de rondas. Es por esta razón, que se decidió sólo en esta primera parte incrementar los valores de la cantidad de juegos. Es decir, cada juego se establece como unidad de medida temporal para el agente. Todos los agentes implementados en el entorno compiten entre si.

5.1. Enfoque basado en introspección

En esta sección se abordará el enfoque basado en introspección. Conocer sus principales implicancias en el proceso de aprendizaje del agente durante la interacción con el entorno y sus ventajas en relación al costo computacional necesario para calcular la probabilidad de éxito. La probabilidad de éxito se define como la *probabilidad de realizar una tarea siguiendo un criterio particulares relacionados con el escenario* (Cruz Dazeley y Vamplew, 2020).

Para ejemplificar la probabilidad de éxito, se puede considerar un automóvil autóno-

mo que mediante aprendizaje por refuerzo, logro aprender a manejar el vehículo y moverlo de un punto a otro. Si en un trayecto de ida o de vuelta, el vehículo toma ciertas rutas no muy comunes, el agente (el vehículo) estará completamente facultado para poder responder antes las preguntas del piloto. Preguntas como ¿Por qué tomaste esta ruta y no te fuiste por la carretera? Para este tipo de preguntas, el agente contará eventualmente con la capacidad de emitir una respuesta: *Elegí tomar esta ruta debido a que existe un 85 % de probabilidad de que por esta vía lleguemos a destino en menos tiempo.*

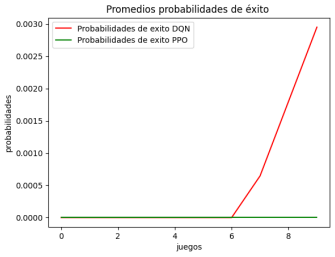
El enfoque basado en introspección, a diferencia del memory-based o learning-based (Cruz Dazeley y Vamplew, 2020), posee la capacidad de calcular la probabilidad de éxito proveniente de los q-values. Es decir, no requiere de una estructura adicional como una tabla para almacenar los p-values como si lo requiere el enfoque learning-based o alguna estructura de dato. He ahí el nombre de introspección. Usar los valores Q como medio introspectivo de la auto motivación del agente.

5.2. Probabilidades de éxito

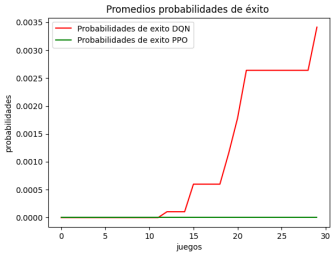
El eje principal de este proyecto es conocer el rendimiento de un enfoque de aprendizaje por refuerzo basado en introspección, aplicado directamente en un entorno competitivo. La tarea principal de este trabajo es evaluar que tan eficiente es esta propuesta de enfoque en entornos competitivos, pero además a esto, cómo responde el enfoque a influencias o interacciones directas con otros competidores inmersos en el mismo entorno (contrincantes en el juego). Los trabajos anteriores donde se implemento este algoritmo, carecían de todas las cualidades mencionadas anteriormente.

Los valores que se pueden apreciar en la figura 5.1 dan cuenta de un crecimiento del promedio de las probabilidades de éxito del agente basado en introspección. En la medida que avanzaban los juegos (cosa que se puede notar en el eje x donde se definen los total de juegos transcurridos) el agente DQN con la implementación del enfoque basado en introspección, responde a un crecimiento en la medida que los juegos transcurrían. Para los entrenamientos con más de 50 juegos realizados es posible apreciar que las probabilidades de éxito poseen un crecimiento hasta transcurrido los juegos 20 y 30 y que en su valor máximo alcanza un valor cercano al 0.016 %. Ya después de haber transcurrido los 50 juegos completados, se puede apreciar en las figura 5.1c y figura 5.1d que las probabilidades de éxito para el agente DQN se mantienen constantes en la medida que transcurren los juegos, sin

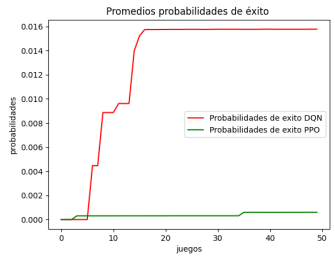
la más mínima variación en su valor.



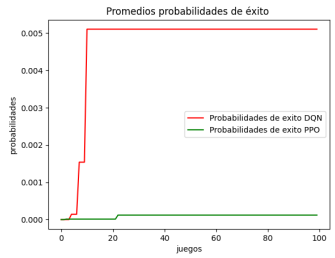
(a) Promedio de probabilidad de éxito para 10 juegos



(b) Promedio de probabilidad de éxito para 30 juegos



(c) Promedio de probabilidad de éxito para 50 juegos



(d) Promedio de probabilidad de éxito para 100 juegos

Figura 5.1: Valores pertenecientes a los promedios de las probabilidades de éxito para los agentes DQN y PPO. Los cálculos de los promedios de las probabilidades de éxito corresponden en su totalidad a los 10, 30, 50 y 100 juegos respectivamente.

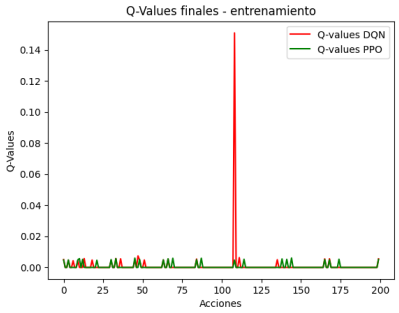
Por otra parte, el agente PPO (proximal policy optimization) mantuvo su probabilidad de éxito casi constante en la medida que transcurrían los juegos. Se puede notar un leve crecimiento después del juego 20 y 35 en las figuras 5.1c y figura 5.1d, pero luego de esto, mantuvo su comportamiento en el tiempo sin grandes diferencias . Es más, las probabilidades nunca alcanzaron un valor por sobre el 0.02 %. Solo se pueden apreciar pequeñas alzas de probabilidad en las figuras 5.1c y 5.1d.

5.3. Q-Values y recompensas

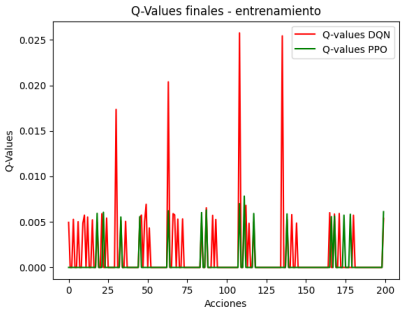
5.3.1. Q-values

En relación de los Q-values, el gráfico 5.2 representa los Q-values para un estado final. En este gráfico se puede apreciar que en general, los Q-values para el agente

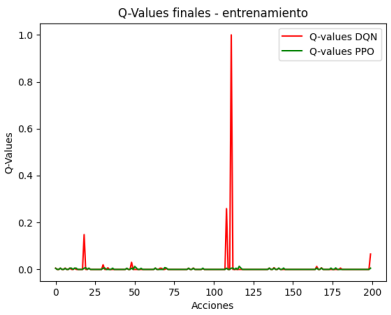
DQN son mayores que los Q-values para el agente PPO. La figura 5.2b indica que solo para la configuración de juegos, al terminar el juego 30, los Q-values, para ciertas acciones son similares. Para los casos correspondientes a las primeras acciones, los Q-values del agente DQN son en su mayoría mas altos, no así para las acciones con numero 80 y 90 dentro de la lista de posibles acciones. Siguiendo con el ejemplo de esta imagen, se puede deducir que el agente fue capaz de aprender un conjunto de acciones que le garantizaban posiblemente ganar los juegos, en general, el agente DQN obtiene mayores Q-values, incluso llegando al valor máximo a obtener al seleccionar una acción, que es el 1. Esto da cuenta de que el agente a llevando a cabo el proceso de aprendizaje. El juego está estructurado de manera tal que el agente devuelve una lista de valores para cada posible acción. Si la acción es válida, el valor de la posición de la lista es 1, si la acción no es permitida, el valor es cero. Con más de 50 juegos en la fase de entrenamiento, el agente en algunos pasajes, puede llegar al valor máximo.



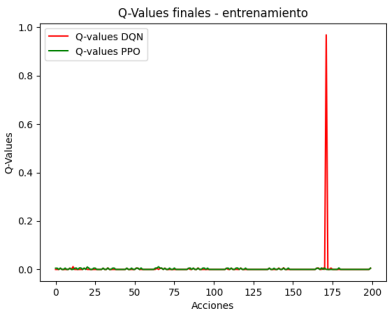
(a) Q-values finales para 10 juegos



(b) Q-values finales para 30 juegos



(c) Q-values finales para 50 juegos



(d) Q-values finales para 100 juegos

Figura 5.2: Valores pertenecientes a los Q-values para estados finales pertenecientes a los agentes DQN y PPO. Los Q-values corresponden a la cantidad de 10, 30 50 y 100 juegos respectivamente.

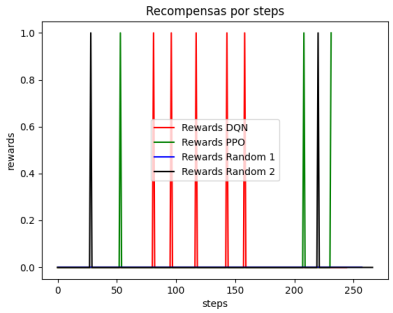
Otro factor que influye en el progreso de los valores Q, es el actuar de cada contrincante, es decir, cada jugada realizada por el jugador que antecede al actual jugador, tiene una implicancia directa en la selección de la actual acción. Esto origina una gran limitación para el agente, ya que lo impide de elegir una acción libremente, o una mejor acción en desmedro de una acción quizás no tan buena. Al final de esto, los valores Q se ven afectados de forma directa para cada estado específico.

5.3.2. Recompensas

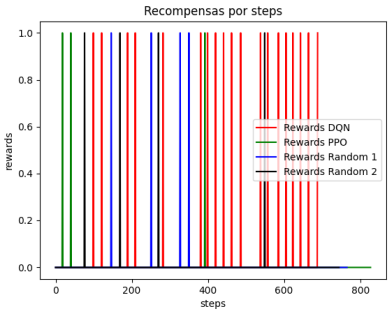
Las recompensas son el objetivo principal del agente. Maximizar las recompensas es por lo que el agente interactúa con su entorno episodio por episodio. Se puede apreciar en la tabla 5.1 que el agente propuesto en este trabajo tiene un muy buen rendimiento en los primeros 100 juegos aproximadamente. Gana fácilmente la mayoría de los juegos. En los primeros 10 y 30 juegos, el agente DQN saca casi el doble de puntaje que sus contrincantes. Esto se puede apreciar de manera clara en la figura 5.3b donde el agente DQN en los últimos pasajes del juegos, es capaz de mantenerse ganando. Sin embargo, a partir del juego 70 en adelante aproximadamente, se puede ver un incremento en el rendimiento del agente PPO. Este agente comienza a ganar juegos de forma consecutiva, incluso llegando a ganar 11 juegos de forma continua. En los entrenamientos que involucro 100 juegos de pruebas, el agente PPO doblego al agente DQN en más de la mitad del puntaje.

Tabla 5.1: Cantidad de juegos ganados por agente

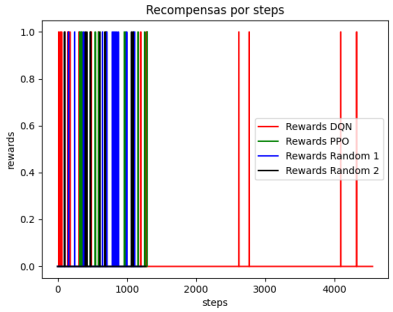
Agentes	**10 juegos**	**30 juegos**	**50 juegos**	**100 juegos**
DQN	5	20	17	20
PPO	3	3	12	54
Dummy 1	0	5	13	20
Dummy 2	2	2	8	6



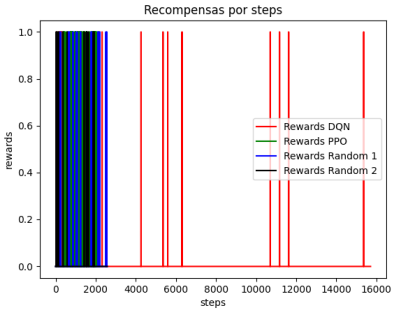
(a) Recompensas por steps para 10 juegos



(b) Recompensas por steps para 30 juegos



(c) Recompensas por steps para 50 juegos



(d) Recompensas por steps 100 juegos

Figura 5.3: Valores pertenecientes a las recompensas obtenidas por steps por cada uno de los agentes implementados. Las recompensas tienen un valor de 1 para el agente que gane el juego y -0.001 para cada jugada donde el agente no gane. Las imágenes corresponden a la cantidad de 10, 30, 50 y 100 juegos respectivamente.

En un futuro no muy lejano, se espera que los algoritmos de aprendizaje por refuerzo, y más específicamente, los algoritmos basados en explicabilidad, puedan entregar información de forma entendible y clara a personas no expertas. La principal característica de este enfoque basado en introspección, es la capacidad de generar probabilidades de éxito a partir de los Q-values, las recompensas entregadas por el entorno y un parámetro gamma constante. Estas probabilidades de éxito permiten mapear de cierta manera la forma de actuar del agente, en otras palabras, el por qué el agente tomo tal acción y no otra. En este caso, las probabilidades de éxito se pueden relacionar a las estrategias tomadas por nuestro agente. Si una persona no experta quisiera saber por qué el agente tomó en el episodio número 50, la decisión de jugar 3 cartas de una valor 5 y adicionar un joker, a sabiendas de que poseía 4 cartas del mismo valor, sabiendo que también era una

jugada permitida y que de igual forma se podían jugar esas 4 cartas para, el joker guárdalo para una futura ocasiones, podría recurrir a las probabilidades de éxito del agente y consultar. La posible explicación del agente sería que *existía un 25 % de posibilidades de ganar la ronda si arrojaba el joker junto con las otras 3 cartas, debido a que si solo dejaba el joker en la mano, tenia una probabilidad de 0.005 % de poder ganar la siguiente ronda, lo cuál, es una probabilidad muy baja en comparación con la primera.* De esta manera, sería más entendible para una persona no experta, entender el funcionamiento del proceso de la toma de decisión del agente y por qué actuó como actuó.

Capítulo 6

Conclusiones

Gracias al cumplimiento de los objetivos específicos, actualmente se cuenta con un entorno idóneo para la elaboración e implementación del agente basado en introspección. Se pudo conocer los fundamentos teóricos y las directrices principales que sustentan temas relativamente nuevos y que derivan principalmente de una nueva forma de ver la inteligencia artificial y el aprendizaje por refuerzo, que es la explicabilidad. Siguiendo este argumento, se cumple con el entendimiento a cabalidad de la problemática que este nuevo campo de la inteligencia artificial viene a afrontar en las siguientes décadas.

La implementación del nuevo enfoque propuesto trajo consigo muchos desafíos. Uno de los principales desafíos era poder aplicar el enfoque en un entorno competitivo, en otras palabras, el objetivo principal era deseado no por un agente, sino que por 4. Esto hacia que cada jugador influyera de manera directa o indirecta en nuestro agente. Otro punto a evaluar era la capacidad de aprender a jugar y a generar estrategias por parte de nuestro agente.

En términos generales, el agente (DQN) y el enfoque basado en introspección propuesto se pudieron implementar de forma correcta. Las probabilidades probabilidades otorgadas por el enfoque en general eran muy pequeños. En el caso del agente DQN, las probabilidades de éxito para cada acción tomada, no superaba el 0.016 %, siendo este valor muy bajo para considerarlo como una medida de confianza en la toma de decisiones. Para el caso del agente PPO, este tampoco pudo destacar con las probabilidades de éxito, ya que su mayor registro no fue mayor a 0.002 %.

En cuanto a las recompensas obtenidas por los agentes, el agente DQN comenzaba de muy buena forma ganando cierta cantidad de juegos durante los primeros juegos, pero transcurrido los 100 o mas juegos, su rendimiento disminuía de forma paulatina, quedando siempre por debajo del agente PPO. En otras palabras, el agente DQN rendía de forma positiva, pero no destacaba por sobre el resto. Es más, en alguno pasajes del juego, la cantidad de puntos obtenidos por DQN, se igualaban a los obtenidos por el agente Dummy. Los agentes Dummys o randoms lograban ganar juegos y obtener recompensar, pero su rendimiento no decía mucho.

Se pudo comprobar que de forma experimental, y validando la hipótesis planteada en este trabajo, el enfoque basado en introspección rinde de forma positiva en el juego, superando a los otros agentes en ciertas partes del juegos, sin embargo durante el proceso de entrenamiento y validación, nunca fue determinante. Nunca logro obtener una diferencia apabullante sobre los otros agentes. Los resultados obtenidos generan sensaciones positivas ya que el agente pudo completar una cantidad aceptable de juegos, pero no sería viable el poder establecer mecanismos de transparencia debido a las bajas probabilidades.

Otro factor que influyó en en el rendimiento del agente, fue el entorno basado en competencia. La competitividad de alguna u otra manera, influyo de forma negativa en la toma de decisión, ya que cada acción de los oponentes, repercutía en la estrategia adoptada por agente. Cada acción de los contrincantes significaba generar una planificación nueva, es decir, no existía la forma del agente en mirar hacia el futuro y poder tomar la mejor decisión.

En conclusión, los objetivos planteados de forma preliminar al desarrollo del experimento, son cumplidos. Se pudo corroborar el uso de este enfoque respondió a las expectativas y a un entorno sumamente complejo en cuanto a su dinámica y no a sus reglas. Se logró implementar el método propuesto, con resultados aceptables.

6.1. Trabajos Futuros

Para los posteriores trabajos relacionados con el área de explicabilidad, se espera proponer una extensión de esta tesis con un nuevo enfoque que tome como base el enfoque basado en introspección pero que sea capaz de poder generar estrategias de corto, mediano o largo plazo en entornos basados en competencias. De esta manera, los futuros agentes de aprendizaje por refuerzo se puedan insertan en entornos competitivos de mejor manera, adoptando mejores estrategias para la toma de decisión y utilizando las probabilidades de éxito para entregar explicaciones sobre

sus acciones a personas no técnicas.

Capítulo 7

Lista de Acrónimos

AI – Artificial Intelligence.

XIA – Explainable Artificial Intelligence.

ANN – Artificial Neural Network.

HRI – Human-robot Interaction.

MDP – Markov Decision Process.

NN – Neural Network.

RL – Reinforcement Learning.

XRL – Explainable Reinforcement Learning.

DNN – Deep Neural Network.

DQN – Deep Q-Network.

PPO – Proximal Policy Optimization.

TD – Temporal Difference.

HRI – Human-Robot Interaction

Bibliografía

1. Margaret A. Boden. 2017. Inteligencia Artificial, Editorial Turner.
2. Stuart Russell and Peter Norvig. 2010. Inteligencia Artificial, Un Enfoque Moderno, Editorial Pearson.
3. Pablo Barros, Anne C. Bloem and Emilia Barakova. 2020. It's a Food Fight! Introducing the Chef's Hat Card Game for Affective-Aware HRI.
4. Francisco Cruz, Richard Dazeley and Peter Vamplew. 2020. Explainable robotic systems: Understanding goal-driven actions in a reinforcement learning scenario.
5. Roberto Hernández Sampieri, Carlos Fernández-Collado y Pilar Baptista Lucio. 2006. Metodología de la investigación. Cuarta edición. Editorial McGraw Hill.
6. C. León, G. Miranda, C. Rodríguez, E. Segredo y Carlos Segura. 2015. El método científico en la era de los ordenadores.
7. Tamara Otzen, Carlos Manterola, Iván Rodríguez-Núñez y Maricela García-Domínguez, La Necesidad de Aplicar el Método Científico en Investigación Clínica. Problemas, Beneficios y Factibilidad del Desarrollo de Protocolos de Investigación, International Journal of Morphology. Extraído de https://scielo.conicyt.cl
8. Richard S. Sutton and Andrew G. Barto. 2018. Reinforcement Learning an introduction, Second edition. The MIT Press.
9. Roberto Hernández Sampieri, Carlos Fernández Collado y Pilar Baptista Lucio. 2014. Metodología de la investigación. Sexta edición. McGraw-Hill
10. Pablo barros, Anne C. Bloem, Inge M. Hootsmans, Lena M. Opheij, Roman H.A. Toebosh, Emilia Barakova and Alessandra Sciutti. 2020. The Chef's Hat Simulation Environment for Reinforcement-Learning-Based Agents. Extraído de https://arxiv.org/
11. Pedro Sequeira, Eric Yeh and Melinda Gervasio. 2019. Interestingness Elements

for Explainable Reinforcement Learning through Introspection. Extraído de http://ceur-ws.org

12. Prashan Madumal, Tim Miller, Liz Sonenberg and Frank Vetere. 2019. Explainable Reinforcement Learning Through a Causal Lens. Extraído de https://arxiv.org

13. Pablo Barros, Ana Tanevska and Alessandra Sciutti. 2020. Learning from Learners: Adapting Reinforcement Learning Agents to be Competitive in a Card Game. Extraído de https://arxiv.org/abs/2004.04000

14. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. 2017. Proximal policy optimization algorithms. Extraído de https://arxiv.org/abs/1707.06347

15. Jianqing Fan, Zhaoran Wang, Yuchen Xie and Zhuoran Yang. 2020. A Theoretical Analysis of Deep Q-Learning. 2nd Annual Conference on Learning for Dynamics and Control. Extraído de http://proceedings.mlr.press/v120/yang20a.html

16. Volodymyr Mnih1, Koray Kavukcuoglu1, David Silver1, Andrei A. Rusu1, Joel Veness1, Marc G. Bellemare1, Alex Graves1, Martin Riedmiller1, Andreas K. Fidjeland 1, Georg Ostrovski1, Stig Petersen1, Charles Beattie 1, Amir Sadik1, Ioannis Antonoglou1, Helen King1, Dharshan Kumaran1, Daan Wierstra1, Shane Legg1 and Demis Hassabis1. 2015. Human-level control through deep reinforcement learning.

17. Violeta ArancibiaC.,Paulina Herrera P. and Katherine Strasser S. 2007. Manual de psicología educacional. Séptima edición