MEMORIA:
PRACTICA 2

GRUPO G

DESARROLLO DE JUEGOS CON INTELIGENCIA ARTIFICIAL

18 de junio de 2025

Jose Ignacio González Vicente,
Cristina González de Lope
& Pablo García García

ÍNDICE

1. INTRODUCCIÓN Y CONTEXTO

En la práctica planteada se parte con una escena en el motor de juego "***Unity***", concretamente en la versión: 2022.3.31f1. Dicha escena se encuentra compuesta por un escenario con muros, un enemigo y el agente principal con el que se trabajará en este proyecto.

1.1. PROBLEMA PLANTEADO

El objetivo principal consta de **entrenar al agente** mediante un algoritmo de **Q-Learning** para que sea capaz de huir del enemigo, que le irá persiguiendo por el mapa, aguantando un número estimado de pasos. El agente, además de huir de él, deberá reconocer las zonas por las que no se puede avanzar, como los muros o las zonas límite, para no caerse del mapa o entorpecer su camino de huida.

Dicho algoritmo de Q-Learning deberá de implementarse junto con una **Tabla Q**, donde se recogerán las acciones y estados del entorno. Por último, habrá que entrenar al agente para que sea capaz de adaptarse a mapas con distinta disposición de los muros y localización inicial del enemigo.

2. ALGORITMO IMPLEMENTADO. CONCEPTOS

Para desarrollar el algoritmo se ha hecho uso de tres clases: la clase "***QMindTrainer***", la clase "***QMindTester***" y la clase "***State***". Además, se ha utilizado un archivo denominado "**TablaQ.csv**", en el que poder guardar los datos del entrenamiento del agente.

Clase "*QMindTrainer*"

Esta clase es la encargada de implementar el algoritmo de **Q-Learning**, en ella, se encuentran las funciones detalladas a continuación:

- **Función "*void Initialize*"** → encargada de inicializar el mapa, los algoritmos y parámetros, así como la tabla Q. En el caso de ser la primera vez que se ejecuta, se creará la Tabla Q vacía, sino, se creará y se cargará la Tabla Q guardada en el archivo .csv.

- **Función "*void DoStep*"** → encargada de realizar los pasos que seguirá el agente. Si el agente es pillado por el enemigo, cruza a una zona inaccesible (como un muro o el vacío) o se lleva a más de 1000 pasos se acabará el primer episodio y se calculará la recompensa media que ha obtenido el agente en dicho margen de pasos. Además, se reseteará el mapa y se guardará la Tabla Q en el archivo .csv.

  Mientras no ocurra ninguna de las tres opciones mencionadas, el agente actualizará su posición constantemente (sumando pasos), calculando la recompensa asociada a cada decisión y guardando los datos de los estados, acciones tomadas, recompensa y próximo estado en la Tabla Q.

- **Función "*int SelectAction*"** → encargada de seleccionar la acción más adecuada que realizará el agente basándose en el valor del "***Epsilon***". Su valor de retorno se toma según el estado recibido, haciendo uso de la función "***GetBestAction***". El valor "***Epsilon***" afectará la probabilidad de que el agente tome una acción aleatoria (exploración).

- **Función "*int GetBestAction*"** → encargada de calcular la mejor acción que podría realizar el agente según el estado recibido. Concretamente, recorre las cuatro opciones disponibles para el agente y toma la acción con el valor Q más alto.

- **Función "*float GetQValue*"** → encargada de acceder a la Tabla Q y devolver el **valor Q** asociado a una acción y un estado concretos.

- **Función "*void UpdateQTable*"** → encargada de actualizar la Tabla Q según la acción tomada desde el estado actual del jugador, añadiendo la recompensa obtenida y su estado resultante.

- **Función "*void SaveQTableToCsv*"** → encargada de guardar la Tabla Q sobre la que se ha trabajado en el archivo .csv. Se recorre el diccionario y se asocian los estados y acciones a los valores Q. Cabe destacar que los estados se representan con un "Id".

- **Función "*Dictionary<(State, int), float> LoadQTable*"** → encargada de cargar los valores asociados a la Tabla Q del archivo .csv. Procesa y divide los datos del archivo en columnas (estado, acción, valor de Q). Por último, devuelve la tabla cargada.

- **Función "*float CalculateReward*"** → encargada de generar un valor de recompensa, basándose en la posición del agente y del enemigo, así como si choca con un muro, sale del mapa o es alcanzado por el enemigo. La mayor penalización se genera cuando le alcanza el enemigo. Concretamente, en el caso de que el enemigo alcance al agente o atraviese una celda inaccesible (como un muro o el vacío), recibirá una penalización de -100. Así mismo, si se acerca al enemigo recibirá otra penalización de -10, mientras que si se aleja, la recompensa será de +50 (por cada paso).

- **Función "*void ResetEnvironment*"** → encargada de reiniciar el entorno, el agente y el enemigo en posiciones aleatorias, además de restablecer todas los parámetros antes de comenzar un nuevo episodio.

- **Función "*(CellInfo, CellInfo) UpdateEnvironment*"** → encargada de actualizar el entorno del agente y del enemigo teniendo en cuenta la acción realizada por el agente. Tiene como finalidad devolver las nuevas posiciones de ambos.

Clase "*QMindTester*"

Esta clase está diseñada para simular cómo actuaría el agente dentro de un entorno definido por ***WorldInfo*** y haciendo uso de la Tabla Q generada en la clase "***QMindTrainer***". El agente tomará decisiones sobre sus acciones a realizar según el entrenamiento por refuerzo realizado (reflejado en los valores de la Tabla Q).

Se hace uso de la función "***LoadQTable***" para cargar los valores de la **Tabla Q** desde el archivo QTable.csv. Se inicializa el mundo con las celdas y sus muros, así como las posiciones aleatorias del agente y del enemigo. El agente toma acciones haciendo uso de la función "***GetBestAction***", la cual tiene en cuenta su posición y la del enemigo (además de las celdas inaccesibles). Una vez seleccionada la mejor acción el agente se mueve en función de esta, finalizando en su nuevo estado y recalculando la mejor acción desde la nueva localización.

Clase "*State*"

Esta clase ha sido creada especialmente para manejar los distintos estados en los que se puede encontrar el agente. Se encarga de almacenar la información sobre el entorno y las posiciones del agente y del enemigo.

Almacena la información de las celdas inaccesibles, como es el caso de los muros, según sus direcciones cardinales (**Norte**, **Sur**, **Este** y **Oeste**). Además de guardar la posición del enemigo, es decir, si se encuentra por encima o a la derecha (si alguno resulta negativo significa que el resultado es su opuesto). Por último, incluye la distancia que separa al agente y a su enemigo, dividiéndola en tres valores distintos: cerca, distancia media y lejos respecto a la posición del agente.

Tiene métodos como "***Equals***" o "***GetHashCode***", los cuales permiten usar objetos de la clase "***State***" en el diccionario (estructura de datos) y compararlos entre sí. Por otro lado, existe el método "***StateId***", encargado de generar el id que representa cada estado de forma única para diferenciarlo a la hora de guardarlo en la Tabla Q.

3. DETALLES DEL ENTRENAMIENTO

En cuanto al entrenamiento del agente, se ha optado por lo siguiente:

- **Cambio de estructura del nivel** → durante el entrenamiento se ha variado la disposición del mapa, creando así nueve mapas con combinaciones distintas: zig-zags de obstáculos, pasillos de tamaño justo, etc. De esta forma se consigue mejorar la robustez del agente frente a la posible variabilidad de su entorno. Al cambiar el mapa se obliga al agente a generalizar sus estrategias para que sean válidas en cualquier tipo de mapa y, además, a evitar adaptarse a un tipo de disposición concreta. En general, esto hace que todo el entrenamiento sea más flexible, ya que así se permite que el agente tome decisiones óptimas incluso en contextos nuevos o inesperados. Se considera que esto es especialmente importante teniendo en cuenta el método de evaluación de esta práctica.

- **Parametrización** → teniendo en cuenta que el valor de "***Alpha***" controla la velocidad a la que aprende el agente, mantener este valor constante asegura que el ritmo de aprendizaje sea estable, evitando cambios drásticos. Por otro lado, el parámetro "***Gamma***" determina cuánto peso se asigna a las recompensas futuras frente a las inmediatas; un valor constante en él permite que, el agente, mantenga una visión aún más consistente sobre la importancia de planificar sus acciones a largo plazo. Por último e inicialmente, el valor del parámetro "***Epsilon***" era muy alto (0,9), incentivando la exploración. Cada vez que se entrenaba al agente con dicho valor en todos los mapas, este valor se ha ido reduciendo 0,1 en cada una, para así reducir la probabilidad de exploración y utilizar cada vez más los valores aprendidos en la ***Tabla Q***. Se ha repetido el proceso hasta que ***Epsilon*** valía 0,4.

- **Entrenamiento** → para entrenar al agente en cuestión, se han ejecutado 10.000 episodios para cada uno de los nueve mapas construidos. Este proceso de entrenamiento se ha repetido por cada ocasión que se bajaba el valor del ***Epsilon*** en 0,1. De esta forma, se logra que el agente no se adapte a un entorno concreto y que, a su vez, sea capaz de desenvolverse de forma ágil y correcta en mapas en los que no haya entrenado.

- **Serialización de la Tabla Q (.csv)** → para guardar los nuevos valores aprendidos entre mapas y ejecuciones, es indispensable serializar los datos de la tabla Q. En este caso, los datos (originalmente guardados en un diccionario de tipo **<*(State, int), float*>**), se han guardado en un archivo .csv. Cada fila tiene varias columnas que representan todas las variables que componen un estado completo. Además, en otras columnas se guardan la acción a tomar (Up, Right, Down, Left), y el **valor Q** asociado a esa misma acción. De esta forma, se podrá cargar la tabla Q entre ejecuciones y a la hora de probar el agente en el tester.