Universidad Nacional Autónoma de México

Facultad de Psicología

El estudio computacional de la ansiedad con algoritmos de inteligencia artificial

T E S I S

Para obtener el título de Lic. Psicología

Presenta:

Alicia Muñiz Jiménez

418002179

ali.psi.neuro@gmail.com

Celular: 8444433888

Director

Dr. Arturo Bouzas Riaño

Revisor

Dr. Víctor Germán Mijangos de la Cruz

Ciudad Universitaria, Cd. Mx., 2025

Hoja de Datos del Jurado

1. Datos de la alumna

Alicia Muñiz Jiménez

Correo: ali.psi.neuro@gmail.com

Celular: 8444433888

Universidad Nacional Autónoma de México

Facultad de Psicología

Psicología

Número de cuenta: 418002179

2. Datos del director de tesis

Nombre: Dr. Arturo Bouzas Riaño

Correo: abouzasr@gmail.com

3. Datos del revisor de tesis

Nombre: Dr. Víctor Mijangos de la Cruz

Correo: vmijangosc@ciencias.unam.mx

4. Datos del sinodal 1

Nombre: Dr. Juan José Sánchez Sosa

Correo: jujosaso@gmail.com

5. Datos del sinodal 2

Nombre: Dr. Iwein M. Roger Leenen

Correo: iwin.leenen@gmail.com

6. Datos del sinodal 3

Nombre: Dr. Germán Palafox Palafox

Correo: germanpalafox@gmail.com

Agradecimientos

A la Dra. Dora Elia Fantini y mis madrinas que creyeron en mí y me apoyaron para realizar mis estudios universitarios durante toda la carrera, incluso después de cambiarme de carrera. A Mireya que me apoyó con los boletos de avión para realizar una estancia de investigación en Canadá.

Al Dr. Bouzas por dejarme elegir libremente mi tema de tesis, apoyarme y guiarme durante todo el proceso.

Al Dr. Mijangos, por toda su orientación y en especial por las asesorías y paciencia para explicarme conceptos matemáticos o de programación que fueron clave para poder llevar a cabo la tesis.

A mis amigos, en especial mis roomies, que me escucharon todas esas veces en las que estaba tratando de entender ciertas ideas o cuando me surgían dudas, porque fueron un gran apoyo emocional en mi vida universitaria.

Además, esta tesis fue factible gracias al financiamiento recibido por diversos proyectos de investigación o divulgación:

1. **Proyecto Conahcyt A1-S-11703:** Comportamiento adaptable en entornos dinámicos.
2. **Proyecto DGAPA-PAPIIT (UNAM) BG101721:** Modelos computacionales para el comportamiento adaptable de robots de servicio y de humanos.
3. **Proyecto DGAPA-PAPIME (UNAM) PE309624:** Simuladores para la enseñanza de modelos del aprendizaje, la cognición y la decisión.

4. **Programa para el Impulso a la Titulación por Actividades Académicas en el Extranjero (PITAAE) 2024:** Me permitió realizar una estancia de investicación en McGill University bajo la dirección del Dr. Otto Ross y aprender más de modelamiento computacional.

Resumen

La mayoría de las investigaciones sobre ansiedad implican tareas experimentales de un solo paso, a pesar de que la ansiedad surge de la anticipación a una amenaza futura percibida como incontrolable e impredecible, es decir, donde la evaluación secuencial es crucial. Este trabajo utiliza el modelado computacional para estudiar la ansiedad en un entorno más natural al fenómeno de estudio, abordando el debate en la literatura sobre la relevancia de una elavada sensibilidad al castigo o una elevada tasa de aprendizaje a los castigos para simular el comportamiento ansioso. Se desarrollaron dos modelos de aprendizaje por refuerzo: Dyna β-pessimistic SR, con parámetros de sensibilidad diferenciados para castigos y recompensas, y Dyna α-SR, con tasas de aprendizaje distintas para cada uno. Los modelos se evaluaron con diferentes valores para cada parámetro en la tarea Cliff Walking, usando un entorno determinista donde la estocasticidad estaba en la selección de acciones, simulando que el agente no tiene un completo control sobre sus decisiones. Los resultados mostraron que solo el modelo Dyna β-pessimistic SR reprodujo conductas asociadas a la ansiedad, como la evitación, aversión al riesgo, sobreestimación y generalización del peligro. Estos hallazgos subrayan el papel central de una alta sensibilidad a los castigos en el comportamiento ansioso en tareas de evaluación secuencial de múltiples pasos y sugieren que es más relevante cómo impactan los refuerzos estimados en la planificación, que la velocidad con las que se aprenden.

Índice general

1. Introducción

"The mind is a neural computer,
fitted by natural selection with combinational algorithms
for causal and probabilistic reasoning ..."

Steven Pinker

Los desórdenes de ansiedad tienen la mayor incicidencia a nivel mundial en la categoría de las enfermedades mentales, sin embargo, los métodos de tratamiento actuales tienen un porcentaje de efectividad que va del 28 % al 52 %, mostrando que todavía hay una gran incomprensión al respecto (Pike y Robinson, 2022; Stein et al., 2017). Además, se conoce que la ansiedad sesga mútiples procesos cognitivos e impacta en cómo las personas ven y responden a su ambiente, de tal modo que puede llegar a representar un deterioro en la funcionalidad de las personas y una barrera para tener un bienestar económico y una buena calidad de vida (Grant y White, 2016; Wilmer et al., 2021). Por todo ello, es imperativo que sea un tema relevante de investigación, pero que además se aborde desde perspectivas nuevas para entender aquello que se escapa con los protocolos de investigación y métodos de modelamiento tradicionales en la psicología.

El modelamiento computacional ha revolucionado el campo de la psicología dando paso a simular la conducta humana y estudiarla con una precisión matemática, permitiendo entender a mayor profundidad los fenómenos de estudio e ir más

allá de los esquemas descriptivos que abundan en la psicología. (Diederich, 2023; Wilson y Collins, 2019). En particular, la psiquiatría computacional se centra en crear modelos matemáticos de fenómenos neuronales o cognitivos relevantes para los desórdenes psiquiátricos, donde las enfermedades psiquiátricas se conceptualizan como un extremo de una función normal o como una consecuencia de alguna parte alterada de un modelo (Q. Huys, 2013). Tres tipos de modelos representativos derivados desde esta perspectiva teórica son: los modelos de redes neuronales, los modelos de aprendizaje por refuerzo y los modelos Bayesianos (Simmons et al., 2020). Los modelos de aprendizaje por refuerzo son los más evaluados en poblaciones clínicas, donde la teoría detrás de estos describe cómo un agente adapta su comportamiento con base en su experiencia o conocimiento del entorno para maximizar las recompensas y minimizar los castigos (Raymond et al., 2017). El aprendizaje por refuerzo por mucho tiempo ha sido un área de investigación muy prominente en la psicología, pero ahora también es una de las tres principales ramas del aprendizaje de máquina, que a su vez es un subcampo de la inteligencia artificial (Doya, 2023). Esto termina siendo especialmente útil si se usa para hacer simulaciones y predicciones. Por ejemplo, se pueden usar agentes de inteligencia artificial con modelos de aprendizaje por refuerzo detrás, los cuales son construidos a partir de la literatura científica del tema, para luego contrastar la conducta observada del agente artificial (en entornos virtuales) con lo esperado de la literatura. E incluso se pueden llegar a comparar distintos modelos con diferentes hipótesis en su construcción, con el fin de ver cuál hipótesis es la más acertada.

Hasta el momento la mayoría de los paradigmas experimentales para estudiar la ansiedad desde una perspectiva computacional involucran tareas de bandidos multibrazo o del tipo go-no go, es decir, tareas de un solo paso donde el agente llega a realizar múltiples ensayos o episodios de la tarea, pero cada ensayo termina con una sola acción del agente. Estas investigaciones han identificado una elevada tasa de aprendizaje a los castigos como el factor principal para el desarrollo del comportamiento ansioso e incluso algunos estudios indican una falta de evidencia sobre el rol de la sensibilidad a los castigos (Aylward et al., 2019; Katz et al., 2020b; Pike

y Robinson, 2022). Esto puede que sea cierto para paradigmas experimentales como las tareas bandidos multibrazo, pero quizás no sea así para tareas más complejas que impliquen aprendizaje secuencial con múltiples pasos. Por ejemplo, se sabe que los parámetros de un modelo pueden capturar diferentes conductas y procesos cognitivos cuando se aplican a distintas tareas experimentales, lo cual puede explicar en parte la inconsistencia en las investigaciones sobre la relevancia del parámetro de la sensibilidad a los castigos en la ansiedad (Eckstein et al., 2022).

Por otro lado, aunque las investigaciones con tareas de un solo paso han aportado importantes hallazgos sobre la ansiedad, también la simplificación de las tareas empieza a representar una barrera para entenderla, ya que no reflejan contextos más realistas al fenómeno de estudio. En contraparte, plantear experimentos con entornos de múltiples estados, donde el agente tiene que tomar varias acciones antes de obtener una recomensa y por lo tanto realizar una evaluación secuencial, vuelve matemáticamente más complejo el análisis de los datos. Esto ha sido una de las principales razones por el cual se ha pospuesto implementar este tipo de experimentos, aunque recientemente ya se han comenzado a hacer.

En tareas de aprendizaje secuencial con múltiples pasos se ha visto que los modelos híbridos con un componente de aprendizaje directo en el ambiente y un componente de planeación (model-based) son los que más se acercan al comportamiento humano (Momennejad et al., 2017, 2018). En particular, cada vez más se encuentra evidencia sobre que el modelo Succesor Representator (SR) (como componente de aprendizaje directo en el ambiente) es el que explica mejor los datos conductuales y neuronales en los humanos (Gershman, 2018). En contraparte, para el componente de planeación todavía falta mucha más investigación al respecto, ya que no se han comparado estructuras de este tipo en un mismo experimento. Empero, en experimentos ditintos se ha encontrado evidencia para la existencia de algoritmos tipo Dyna y del tipo value-iteration (Momennejad et al., 2017, 2018). Pocos años después de las investigaciones anteriores, llevadas a cabo con población general, se propusieron los primeros modelos mecanicistas de aprendizaje por refuerzo para simular la conducta

ansiosa en contextos similares. El primero de ellos es un algoritmo que implementa el modelo β-pessimistic Q-learning (model-free) o su variación a value-iteration (model-based), logrando capturar características de los desórdenes de ansiedad que involucran aberraciones en los procesos cognitivos orientados al futuro, por ejemplo, la excesiva propagación del miedo y la evitación de situaciones lejos del peligro (Zorowitz et al., 2020). Es importante enfatizar que estos modelos tienen un parámetro de sensibilidad a los castigos, aunque todavía no han sido comparados con otros modelos sobre la ansiedad que no tengan esta variación. Sin embargo, en la literatura se puede encontrar otro modelo que también usa un parámetro de sensibilidad para modelar el miedo y la conducta ansiosa (Dayan y Abbott, 2001). Lo anterior indica que a diferencia de los hallazgos en los experimentos de un solo paso, el factor principal para modelar la conducta ansiosa en las tareas de múltiples pasos es una elevada sensibilidad a los castigos. No obstante, dado que la investigación desde esta perspectiva es relativamente nueva, todavía no se han realizado comparaciones entre modelos equivalentes, ni evaluado el rol de una elevada tasa de aprendizaje a los castigos para simular conductas asociadas a la ansiedad en tareas de evaluación secuencial.

Considerando la disputa en la literatura sobre el rol de los parámetros de la sensibilidad y la tasa de aprendizaje a los castigos en la conducta ansiosa, es importante desarrollar modelos que permitan evaluar el impacto de ambos parámetros de forma independiente en escenarios más naturales a la ansiedad. Para ello, es muy útil considerar que el modelo SR trata el aprendizaje de las recompensas de forma independiente, en contraste con los algoritmos de Q-learning o value iteration, donde el parámetro de aprendizaje impacta tanto en la recompensa recibida como en el valor estimado del estado futuro o el valor Q. De igual forma, por la facilidad de implementar el módulo Dyna y su característica de iterar sobre solo algunos estados y acciones, podiendo ser un simil de imaginar ciertas situaciones en concreto, es que como base para construir los modelos puede se usará el algoritmo DynaSR.

Como resultado, este trabajo se centrará en desarrollar e implementar dos modelos híbridos SR con módulos Dyna, pero con variaciones individuales cada uno:

el primero con la implementación de la variación β-pessimistic, haciendo alusión a la hipótesis de la importancia de la sensibilidad al castigo para la conducta ansiosa; mientras que el segundo modelo solo tendrá distintas tasas de aprendizaje para los castigos y las recompesas, reflejando la hipótesis sobre que una elevada tasa de aprendizaje a los castigos es fundamental para la toma de decisiones ansiosa. Por último, se compararán ambos modelos en entornos de evaluación secuencial, probando diferentes valores de los parámetros centrales de cada modelo, para luego analizar la contribución de cada parámetro a la conducta mostrada por los agentes. Todo con el fin de esclarecer el rol de la sensibilidad y la tasa de aprendizaje a los castigos en contextos más cercanos al mundo real donde surge el comportamiento ansioso.

Primero, en el siguiente apartado de marco teórico se explicarán conceptos importantes para entender qué es la ansiedad, el modelamiento computacional y teoría sobre aprendizaje por refuerzo que es imprescindible conocer para luego pasar a explicar los modelos e ideas relevantes usadas para desarrollar los modelos de este trabajo. Después, en la sección de metodología se dará a conocer la tarea experimental, el pseudocódigo de los algoritmos de los modelos, las condiciones experimentales que se modificaron en los agentes para probar las diferentes hipótesis y cómo se operacionalizaron las variables que se buscaron observar en la conducta de los agentes. Posteriormente, se mostrarán los resultados y la discusión de los mismos. Por último, se abordará la conclusión de la investigación.

2. Marco Teórico

"Our anxiety does not come from thinking about the future,
but from wanting to control it."

Kahlil Gibran

2.1. Ansiedad

2.1.1. Incidencia

A nivel mundial la ansiedad es la más prevalente de las enfermedades psicológicas (Clark y Beck, 2012). En México la última Encuesta Nacional de Bienestar Autorreportado (ENBIARE) indicó que el 19.3 % de la población adulta tiene síntomas de ansiedad severa y otro 31.3 % revela síntomas de ansiedad mínima o en algún grado (INEGI, 2021). Sin embargo, este el trastorno mental con mayor brecha en atención, ya que el 85.9 % de la población que padece de ansiedad en México no recibe un tratamiento (Secretaría de Salud, 2022). Aunado a esto los métodos de tratamiento actuales tienen un porcentaje de efectividad que va del 28 % al 52 % (Pike y Robinson, 2022), mostrando que todavía hay una gran incomprensión al respecto.

2.1.2. Concepto

La ansiedad muchas veces es confundida con el miedo o el estrés, pero en realidad estos son términos diferentes, aunque se solapan de cierta manera. Por ejemplo, la ansiedad nace del miedo (Brewer, 2021), pero no es el miedo en sí mismo, a su vez muchos síntomas se comparten entre el estrés y la ansiedad, pero difiren en las causas que los producen (American Psychological Association, 2019). Para entender mejor el concepto de ansiedad a continuación se definirán los tres conceptos.

El miedo es un estado neurofisiológico automático primitivo de alarma que conlleva la valoración cognitiva de una amenaza o peligro inminente para la seguridad física o psicológica de un individuo, el cual está asociado a conductas como la defensa, los pensamientos de peligro inminente y la huida (American Psychiatric Association, 2014; Clark y Beck, 2012). El objeto del miedo puede ser clasificado como real, externo, conocido u objetivo (Steimer, 2002).

Por otra parte, el estrés es un estado de emergencia del organismo en respuesta a un desafío a su homeostasis o demandas ambientales que suponen un peligro, por lo que el organismo inicia una reacción integrada de respuestas fisiológicas y cognitivo-conductuales, pero también implica una respuesta emocional (American Psychological Association, 2019; Daviu et al., 2019). A diferencia del miedo las casusas del estrés pueden ser reales o imaginarias (Ahmed y Çerkez, 2020). Además, el estrés crónico no tratado puede derivar en ansiedad (Khan y Khan, 2017).

Por su lado, la ansiedad se considera una respuesta anticipatoria - conductual, afectiva, cognitiva y fisiológica - a una amenaza futura muy aversiva que se percibe como imprevisible e incontrolable y que potencialmente podrían amenazar los intereses vitales del individuo, pero que además es muy poco probable o incierta (American Psychiatric Association, 2014; Clark y Beck, 2012). Encima, la amenaza es más persistente en el tiempo, a diferencia del estrés o del miedo, la cual es acompañada de otros factores cognitivos como la percepción de vulnerabilidad o indefensión del individuo y conductas de evitación (Clark y Beck, 2012). En particular, se conoce que la

ansiedad sesga mútiples procesos cognitivos, entre ellos el control cognitivo, e impacta en cómo las personas ven y responden a su ambiente, de tal modo que puede llegar a representar un deterioro en la funcionalidad de las personas y una barrera para tener un bienestar económico y una buena calidad de vida (Grant y White, 2016; Wilmer et al., 2021). Sin embargo, niveles moderados de miedo, estrés o ansiedad tienen un valor adaptativo que ayudan a la supervivencia del individuo, pero si estas conductas son sostenidas en el tiempo o exageradas se vuelven desadaptativas (Robinson et al., 2013).

Desde el enfoque clínico el modelo cognitivo de la ansiedad de Beck ha sido un parteaguas en el entendimiento de la ansiedad e incluso modelos posteriores lo toman como base (Gústavsson et al., 2022). De esta manera, sus postulados sobre los factores principales para el surgimiento de la ansiedad: la sobrestimación del peligro, la subestimación de los recursos personales de afrontamiento y la subestimación de las señales de seguridad en el ambiente (Beck y Clark, 1997), siguen siendo considerados como características principales en muchos modelos.

Actualmente, se sabe que la ansiedad como condición clínica puede ser clasificada en varios subtipos, denominados desórdenes de ansiedad, y las diferencias existentes entre ellos son fundamentales por las implicaciones que tienen para su tratamiento (American Psychiatric Association, 2014; Clark y Beck, 2012). A su vez, el índice de comorbilidad entre ellos o con otros desórdenes del ánimo - como la depresión - es muy alta, generando que los síntomas se vuelven más severos y exista una menor efectividad de tratamiento (Rogers et al., 2020). Por ello, si un desórden de ansiedad puede predisponer al desarrollo de otros trastornos, entenderlos a profundidad es fundamental no sólo para tratarlos, sino como prevención al desarrollo de otros desórdenes. No obstante, dada la amplitud de desórdenes de ansiedad y considerando que hay características comunes entre ellos, un primer paso implicaría comprender no sólo a nivel descriptivo, sino con modelos matemáticos cómo se genera la conducta ansiosa de forma general. Como resultado, en este trabajo se abordará la ansiedad desde una perspectiva más global, como en la mayoría de las investigaciones que usan

el modelamiento computacional, ya que han proporcionado una perspectiva nueva para entender los mecanismos que subyacen a ella, a la par de brindar una mayor precisión en los hallazgos.

A continuación, se explicará qué es el modelamiento computacional y la psiquiatría computacional para posteriomente explicar cómo ésta ha contribuido al entendimiento de la ansiedad e identificar los pricipales debates al rededor del tema.

2.2. Modelamiento computacional

Un modelo es una construcción abstracta que captura la estructura de los datos y genera una representación aproximada de la realidad (Farrell y Lewandowsky, 2015). Particularmente, el modelamiento computacional es el uso de las computadoras para simular y estudiar fenómenos o sistemas complejos, con el propósito de entender las interacciones entre los componentes del sistema usando las matemáticas y las ciencias de la computación (National Institute of Biomedical Imaging and Bioengineering, 2020). Los diferentes tipos de modelos se pueden clasificar dependiendo de qué tipo de pregutnas abordan, por ejemplo, preguntas qué, cómo y por qué son abordadas por modelos descriptivos, mecanicistas e interpretativos, respectivamente (Dayan y Abbott, 2001). Los modelos descriptivos explican la relación existente entre variables al resumir grandes cantidades de datos experimentales de manera compacta pero precisa (Dayan y Abbott, 2001), es decir, se fundamentan en suposiciones básicas sobre qué variables observar y cómo relacionarlas (Levenstein et al., 2023). Por ejemplo, indicando el índice de correlación entre variables para explicar un fenómeno o el valor promedio de una variable y su desvición estándar. En comparación, los modelos mecanicistas van más allá y explican cómo se genera el fenómeno de estudio en términos de sus componentes y sus interacciones, tomando en cuenta los procesos que subyacen al fenómeno (Levenstein et al., 2023). Por ello, a este tipo de modelos se les conoce también como "generativos"(Levenstein et al., 2023) y los modelos de aprendizaje por refuerzo son de este tipo. Finalmente, los modelos interpretativos apelan a criterios de optimalidad de acuerdo a las restricciones del fenómeno de estudio y su función

(Levenstein et al., 2023), mismo que sirve para elegir el mejor modelo entre aquellos que puedan llegar a reproducir el fenómeno de estudio pero que tienen implicaciones o hipótesis distintas.

El objetivo del modelamiento computacional en las ciencias del comportamiento, considerando los modelos mecanicistas, es utilizar el poder de cómputo para para desarrollar modelos matemáticos precisos, haciendo simulaciones y/o estimando los parámetros de los modelos con el fin de entender mejor los datos conductuales, como la toma de decisiones, el tiempo de respuesta, etc. (Wilson y Collins, 2019). Es decir, desde esta perspectiva se instancian diferentes "hipótesis algorítmicas" sobre cómo se genera el comportamiento yendo más allá de las diferencias en el comportamiento observado en diferentes agentes que hacen el mismo experimento, revelando regularidades subyacentes (Hunt et al., 2008; Simmons et al., 2020). Asimismo, para la ciencias cognitivas los modelos computacionales permiten establecer supuestos acerca de cómo la cognición podría operar de una manera precisa, permitiendo una prueba rigurosa de esos supuestos (Simmons et al., 2020).

Es importante notar que al probar que los datos que reproduce un modelo concuerdan con los datos conductuales o experimentales existentes, esto sólo demuestra que el modelo es suficiente para explicar los datos, pero es una declaración débil porque no considera explicaciones alternativas, es decir, hacer una comparación con otros modelos (Simmons et al., 2020). Sólo en el caso de que el modelo pueda dar cuenta de resultados diversos en diferentes paradigmas experimentales, entonces por lo menos mostraría que no es un modelo de un solo uso (Simmons et al., 2020). Por último, los cuatro usos de los modelos computacionales que dominan la literatura son: la simulación, la estimación de parámetros, la comparación de modelos y la inferencia de variables latentes (Wilson y Collins, 2019).

2.3. Psiquiatría computacional

La psiquiatría computacional es el estudio de los desórdenes de salud mental usando el modelamiento computacional, por lo que se centra en crear modelos matemáticos de fenómenos neuronales o cognitivos relevantes para estas enfermedades, mismas que se conceptualizan como un extremo de una función normal o como una consecuencia de alguna parte alterada de un modelo (Q. Huys, 2013). Los modelos matemáticos tienen la ventaja de forzar al investigador a ser preciso, consistente y lo más completo posible en derivar las implicaciones del modelo que crean para dar explicación a un fenómeno, dejando atrás inconsistencias que se hayan en los esquemas descriptivos que abundan en la psicología (Simmons et al., 2020). Es así, que se busca proponer cada vez más modelos computacionales para explicar la psicopatología (Simmons et al., 2020). La psiquiatría computacional engloba dos aproximaciones complementarias: la impulsada por los datos o la impulsada por la teoría, donde la primera usa modelos de *machine learning* como clasificadores, mientras que la segunda usa modelos que ejemplifican hipótesis explícitas sobre los mecanismos subyacentes a las enfermedades mentales, incluso a veces en múltiples niveles de análisis y abstracción (Q. J. Huys et al., 2016).

Adicionalmente, tres ejemplares representativos de los modelos derivados desde la perspectiva teórica son: los modelos realistas de redes neuronales, los modelos de aprendizaje por refuerzo y los modelos Bayesianos (Simmons et al., 2020). Actualmente, los modelos de aprendizaje por refuerzo son los más evaluados en poblaciones clínicas, donde la teoría detrás de estos describe cómo un agente adapta su comportamiento basado en su experiencia o conocimiento del entorno para maximizar las recompensas y minimizar los castigos (Pike y Robinson, 2022; Raymond et al., 2017). Como resultado, este trabajo se centrará en este tipo de modelos. Por otro lado, los modelos de inferencia bayesiana resultan muy útiles para describir cómo la probabilidad asociada con una hipótesis se actualiza en función de la nueva evidencia (Raymond et al., 2017), por eso se usan mucho para modelar conductas donde

se implica la estimación de la volatilidad en el ambiente. De igual forma, es común usar aproximaciones que mezclan los modelos de aprendizaje por refuerzo usando la estadística bayesiana para estimar los parámetros de los modelos y así tener mayor información sobre la incertidumbre de los mismos.

2.3.1. Aportes al entendimiento de la ansiedad

Las investigaciones sobre la ansiedad que han empleado el modelamiento computacional se basan en su mayoría en tareas de bandidos multibrazo o del tipo *go-no go*. Estas tareas son consideradas como de un solo paso porque requieren que el agente aprenda a dar la respuesta apropiada para actuar en un único estado del entorno y, así, obtener una recomponsa inmediata (Yamamori y Robinson, 2023). La desventaja de estas tareas es que no reflejan escenarios del mundo real para estudiar la ansiedad, ya que esta al ser una respuesta anticipatoria a una amenaza futura incierta, entonces sería más útil analizarla usando protocolos experimentales de aprendizaje que impliquen una evaluación secuencial de múltiples acciones en diferentes estados del ambiente para obtener una recompensa (Yamamori y Robinson, 2023). Una de las justificaciones al usar experimentos simplificados para estudiar la ansiedad es que en la ciencia los paradigmas experimentales son reduccionistas para poder entender mejor ciertos aspectos del fenómeno de estudio, pero no hay que olvidar que esto puede manipular tanto los resultados que termine llevando a conclusiones sesgadas lejos de la realidad.

Por otra parte, algunos de los hallazgos más reproducibles usando el aprendizaje por refuerzo en distintas investigaciones son los siguientes: la ansiedad está asociada con tasas de aprendizaje elevadas para los castigos y tasas de aprendizaje reducidas para las recompensas (Pike y Robinson, 2022), personas con altos niveles de ansiedad tienen un sesgo exagerado a evitar los estímulos aversivos (Sharp y Eldar, 2019), las personas con ansiedad generalizada tienen una exagerada aversión al riesgo (Sharp y Eldar, 2019), la rumiación en la ansiedad probablemente refleja una interacción entre un mayor control basado en modelos y una simulación sesgada ne-

gativamente (Goldway et al., 2023), al igual que el pesimismo conlleva a una excesiva propagación del miedo, lo cual a su vez explica un amplio rango de comportamientos en la ansiedad como la exagerada evaluación de las amenazas, la generalización del miedo y una persistente evitación del peligro (Zorowitz et al., 2020).

Asimismo, un debate importante en la literatura actual es sobre lo que se considera el componente principal para el desarrollo de la conducta ansiosa. Por un lado, muchos artículos han apoyado por un largo tiempo que las tasas elevadas de aprendizaje para los castigos, en comparación de las tasas de aprendizaje para las recompensas, es lo que conduce al comportamiento ansioso (Pike y Robinson, 2022). Además, que existe una falta de evidencia del efecto de la ansiedad sobre la sensibilidad al castigo (Aylward et al., 2019; Pike y Robinson, 2022). Empero, es importante notar que estas conclusiones se han obtenido de tareas de un solo paso y que en su mayoría usan variaciones del modelo de Bush y Mosteller para modelar la conducta ansiosa. En contraparte, otros artículos cuestionan las posturas anteriores defendiendo que la sensibilidad al castigo o la pérdida es el principal predictor de la ansiedad (Katz et al., 2020a; Zorowitz et al., 2020), ya que los trastornos de ansiedad son fundamentalmente trastornos del aprendizaje de la incertidumbre (Brown et al., 2023; Gagne y Dayan, 2022; Zorowitz et al., 2020). Adicionalmente, esta última postura es la que se ha probado en tareas de evaluación secuencial, donde los algoritmos basados en un modelo del entorno o de aprendizaje directo son usados para tratar de modelar la ansiedad. No obstante, hasta la fecha no se ha hecho una comparación de modelos equivalentes en la misma tarea experimental que permita evaluar las distintas hipótesis en este tipo de entornos.

A continuación, en el siguiente apartado se profundizará en la teoría sobre aprendizaje por refuerzo y los algoritmos que se usarán como base para construir un modelo que permita evaluar en un entorno más cercano a la realidad las hipótesis del debate actual sobre el factor principal que conlleva a la conducta ansiosa.

2.4. Aprendizaje por refuerzo (RL)

El aprendizaje por refuerzo o mejor conocido por su nombre en inglés Reinforcement Learning (RL) es un metodología de aprendizaje que nace en la psicología y que con el tiempo se extiende a la ciencias computacionales para ser estudiado como parte de la inteligencia artificial (Subramanian et al., 2022; Wu, 2023). Desde las ciencias cognitivas el aprendizaje se define como cambios en el comportmaiento en un organismo que resultan de las regularidades en su ambiente (De Houwer et al., 2013). Esto es muy similar a la definición de aprendizaje de máquina en las ciencias computacionales, donde se considera el aprendizaje como la adaptación a nuevas circunstancias, identificando y extrapolando patrones (Russell y Norvig, 2022).

En particular, el RL es una de las tres principales áreas del *machine learning* (aprendizaje de máquina), el cual se subdivide en: aprendizaje supervisado, aprendizaje no supervizado y aprendizaje por refuerzo, mismo que termina siendo un punto intermedio de los dos anteriores (Doya, 2023). El RL es aprender qué hacer mapeando situaciones a acciones para maximizar las recompenzas a través de la experiencia, ya que el agente puede interactuar con su entorno y recibir refuerzos - recompenzas o castigos - que reflejan qué tan bien está eligiendo sus acciones, o hacer simulaciones de sus interacciones si tiene un modelo del entorno (Russell y Norvig, 2022; Sutton y Barto, 2018). Las caracterísicas más importantes del RL es la búsqueda a través de la prueba y el error, así como la recompenza demorada, ya que a través de diferentes intentos el agente debe descubrir qué acciones le llevan a obtener una mayor recompenza y, además, puede que no se obtenga una recompenza de forma inmediata, sino a futuro al realizar una secuencia de acciones (Russell y Norvig, 2022; Sutton y Barto, 2018).

Uno de los retos que surgen en el RL es el dilema de exploración-explotación, es decir, cuándo conviene elegir la mejor acción estimada (explotar) o intentar diferentes acciones a la óptima para descubrir más del entorno (explorar) (Sutton y Barto,

2018). Esto particularmente cuando el ambiente es desconocido, no es totalmente observable y/o es estocástico. El dilema es que no se puede perseguir el explotar o explorar de forma exclusiva sin fallar en la tarea, sino que se tiene que encontrar un punto adecuado de compensación para elegir entre las dos opciones (Sutton y Barto, 2018). Por ejemplo, en una tarea estocástica cada acción se debe intentar muchas veces hasta ganar un estimado confiable de las recompenzas esperadas (Sutton y Barto, 2018).

Elementos de un sistema de aprendizaje por refuerzo:

- **Modelo del entorno:** Imita el comportamiento del ambiente y permite hacer inferencias sobre cómo el entorno se comportará en el futuro (Sutton y Barto, 2018). Conocer el modelo del entorno implica conocer las probabilidades de transición entre los diferentes estados, así como la función de recompensa, es decir, el valor de las recompensas y cuándo se pueden obtener (Sutton y Barto, 2018). Los métodos para resolver los problemas de RL que utilizan un modelo del entorno y la planificación se denomina métodos basados en modelos (Swazinna et al., 2022), a diferencia de los métodos más simples que no usan un modelo del entorno y, en cambio, los agentes aprenden explícitamente a través de la prueba y el error por su experiencia directa en el ambiente, aunque también hay métodos híbridos.
- **Política:** Define la forma en cómo el agente elige sus acciones (Doya, 2023). Las políticas pueden ser estocásticas al especificar probabilidades por cada acción (Sutton y Barto, 2018).
- **Señal de recompensa:** En cada paso temporal el ambiente le proporciona al agente un refuerzo y la señal de la recompensa define cuáles son los buenos y malos eventos para el agente (Sutton y Barto, 2018). Por ejemplo, valores bajos indican un castigo y valores altos un reforzamiento positivo.
- **Función de valor:** Sirve para estimar el valor de cada estado a largo plazo, ya

que toma en cuenta las recompenzas acumuladas esperadas por el agente en el futuro partiendo de ese estado Russell y Norvig, 2022. El componete principal de todos los algoritmos de RL es considerar un método eficiente para estimar los valores de los estados, ya que el agente busca acciones que lo lleven a los estados de mayor valor, no a la mayor recompenza inmediata (Sutton y Barto, 2018).

Es importante destacar que los elementos de un sistema de aprendizaje por refuerzo son muy similares al de un Proceso de Decisión de Markov (MDP), el cual es un problema de decisión secuencial en un ambiente totalmente observable y estocástico que cumple la propiedad de Markov - la probabilidad de alcanzar un estado futuro depende sólo del estado actual - y se define como una tupla, $\mu = (S, A, P, \gamma, R)$, que consiste en lo siguiente (Ng et al., 1999; Russell y Norvig, 2022):

- Un conjunto de estados finito, $S = \{s_1, s_2, ..., s_n\}$.
- Un conjunto de al menos 2 acciones, $A = \{a_1, a_2, ..., a_n\}$.
- Una función de recompensa, $R(s, a, s')$.
- Un factor de descuento, γ.
- Probabilidades de transición, $P(s'|s, a)$.

Nota: por facilidad de notación se usa $s' = s_{t+1}$ y $s = s_t$.

Esta simulitud entre el RL y el MDP se debe a que la teoría del RL se desarrolla en un Proceso de Decisión de Markov, aunque difiere de simplemente resolver un MDP porque puede que el agente no conozca el modelo de transición ni la función de recompensa del MDP, pero debe tomar acciones para aprender más (Russell y Norvig, 2022). Eso sí, el agente debe ser capaz hasta cierto punto de identificar el estado en el que está y de tomar acciones que afecten ese estado, así como tener una meta (Sutton y Barto, 2018). Finalmente, por la relevancia de este concepto y para enterderlo mejor, en el siguiente subapartado se explicará a mayor profundidad qué es un MDP.

2.4.1. Proceso de Decisión Markoviano (MDP)

Los MDP son una forma de idealización matemática para problemas de aprendizaje por reforzamiento donde un agente aprende de la interacción con su ambiente para alcanzar una meta (Singh et al., 2022). En un MDP un agente y el ambiente interactúan en una secuencia de pasos temporales discretos y en cada paso t el agente recibe una representación del estado en el ambiente en el que está, luego con base en ello el agente selecciona una acción y como resultado recibe una recompensa y transita a un nuevo estado (Doya, 2023). Esto se puede apreciar de forma gráfica en el diagrama de abajo de los autores Sutton y Barto (2018). Además, en la figura se observa un caso práctico de ejemplo.

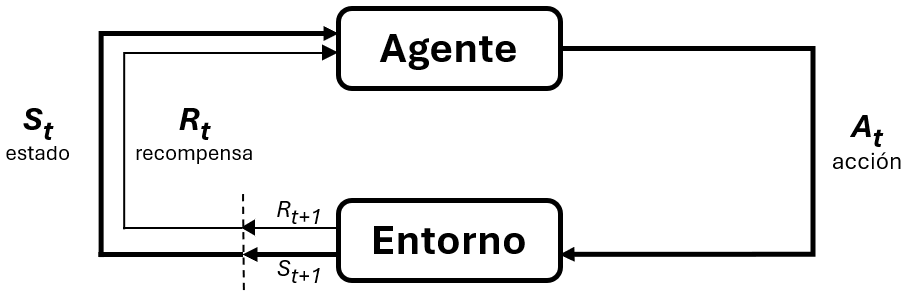


Figura 2.1: Diagrama de la interacción agente-entorno en un MDP.

Es importante notar que en un MDP en cada paso temporal el agente implementa un mapeo de la probabilidad en el estado actual para elegir una acción, a lo cual se le conoce como la política del agente (π_t o π), donde $\pi(a|s)$ es la probabilidad de elegir la acción a en el estado s, mientras que $\pi(s)$ es la acción recomendada por la política π en el estado s (Russell y Norvig, 2022; Sutton y Barto, 2018). Es decir, la política es la solución que especifica lo que el agente debe hacer en cada estado que alcance. Dada la propiedad estocástica del entorno en un MDP, entonces cada vez que una política es ejecutada, ésta termina en una historia del ambiente diferente y, por ello, la calidad de una politica es medidad con la utilidad esperada de las posibles historias de los entornos generados por esa política (Doya, 2023; Russell y Norvig,

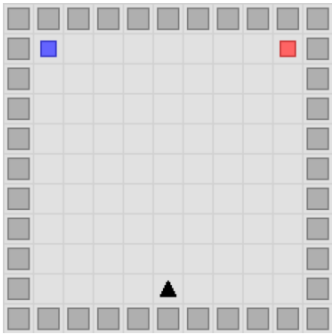


Figura 2.2: Ejemplo de un MDP. Cada cuadrado de color gris claro en el ambiente representa un estado distinto, la posición inicial del agente está representado por el triángulo negro, el cuadrado azul representa la meta y el cuadrado rojo representa un estado aversivo que el agente debe evitar. El agente a través de la interacción con su entorno evaluará el resultado de sus acciones, considerando la recompensa obtenida y el siguiente estado al que transite, para encontrar cuál es la ruta óptima para llegar a la meta.

2022). Es decir, esta función (2.1) permite calcular la utilidad esperada obtenida de ejecutar la política π empezando en el estado s. Además, la función de utilidad se basa en la ecuación de Bellman.

$$U^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R(s, \pi(s), s')\right] \tag{2.1}$$

Donde:

- E = Valor ponderado o esperanza matemática.
- γ = factor de descuento que puede tomar valores entre 0 y 1. Entre más cercano sea su valor a 1 significa que el agente está más dispuesto a esperar por las recompensas a largo plazo.

Por otro lado, la función de valor (2.2) es la función de utilidad condicionada al estado actual, por lo que esta pequeña modificación permite estimar qué tan bueno es para el agente estar en un estado determinado (Sutton y Barto, 2018).

$$v_{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R(s, \pi(s), s') \middle| S_t = s\right] \tag{2.2}$$

Adicionalmente, si lo que se quiere estimar es el valor de tomar la acción a en el estado s bajo la política π, entonces se usa la función de valor-acción (2.3).

$$q(s,a) = E\left[\sum_{t=0}^{\infty} \gamma^t R(s,\pi(s),s') \middle| S_t = s, A_t = a\right] \tag{2.3}$$

De igual forman, hay que recalcar que una política óptima (π^*) es aquella que tiene el valor más alto de la utilidad esperada (2.4). Por ende, el valor de un estado bajo una política óptima es el rendimiento esperado de la mejor acción (2.5), al igual que para el valor de un estado-acción bajo una política óptima (2.6).

$$\pi^* =_\pi E\left[\sum_{t=0}^{\infty} \gamma^t R(s,\pi(s),s') \middle| \pi\right] \tag{2.4}$$

$$v_{\pi^*}(s) = \max_\pi v_\pi(s) \tag{2.5}$$

$$q_{\pi^*}(s,a) = \max_\pi q_\pi(s,a) \tag{2.6}$$

Por otra parte, un tipo de distinción entre los tipos de políticas es si son deterministas o estocásticas, donde la primera siempre elige la misma acción para un estado determinado (por ejemplo elegir la mejor acción), mientras que en el segundo caso dependiendo de una probabilidad se seleccionará una acción para el mismo estado. Los ejemplos más comunes para estos tipos de políticas son la política greedy (siempre maximiza) y la política ϵ-greedy, que dependiendo de la probabilidad ϵ se elije una acción al azar y con probabilidad $1-\epsilon$ se maximiza. La elección del tipo de política a usar depende en gran medida del problema de estudio y el algoritmo que se utiliza, aunque es común usar la política ϵ-greedy o variaciones de la misma por la ventaja que conlleva que el agente tenga la posibilidad de explorar y explotar.

Finalmente, la clave para resolver un MDP o problemas de aprendizaje por refuerzo (RL) es encontrar la política óptima (π) de la función de valor o de valor-acción

para el problema específico. No obstante, estas funciones pueden variar dependiendo del algoritmo de aprendizaje del agente y, en consecuencia, también la política óptima. En el siguiente apartado se explicarán a detalle los algoritmos de Q-learning, β-pessimistic, Dyna-Q y Sucessor Representator, ya que cada uno es relevante para luego entender los modelos que se propondrán para contrastar la relevancia del parámentro de la sensibilidad a los castigos respecto con el peso de la diferencia de las tasas de aprendizaje para refuerzos y castigos, para el desarrollo de la conducta ansiosa.

2.4.2. Algoritmos de aprendizaje por refuerzo

Existen diferentes algoritmos para resolver un problema de RL, pero una forma de distinguirlos es si son basados en un modelo, libres de modelo o híbridos. Los primeros corresponden a los métodos de Programación Dinámica (DP), los cuales son más costosos computacionalmente debido a que el agente usa el modelo del entorno que conoce perfectamente para hacer simulaciones de la interacción agente-ambiente y, así, aprender la política óptima (Sutton y Barto, 2018). Por ello, a este tipo de algoritmos se les asocia más con la planeación *offline* (a través de simulaciones) (Sutton y Barto, 2018). En comparación, los algoritmos libres de modelo al no contar con un modelo del entorno requieren que el agente aprenda por experiencia directa en el ambiente, como con los métodos de diferencias temporales (TD) (Sutton y Barto, 2018. A esta familia de algoritmos pertenecen el Q-learning y el B-pessimistic Q-learning, mientras que el Dyna-Q y el SR pertenecen a los algoritmos híbridos, es decir, combinan la planeación con el aprendizaje directo. No obstante, el módulo Dyna implica una planeación *offline* y el SR una planeación *online*. A continuación, se profundizará en cada algoritmo empezando por los libres de modelo, para luego seguir con los híbridos.

Q-learning

El Q-learning es un algoritmo libre de modelo que se basa en estimar la función de valor-acción (valores Q) a través del método de diferencias temporales (Russell y Norvig, 2022), mismo que se apoya en la diferencia entre predicciones sucesivas en el tiempo considerando el estado actual y el estado siguiente (Sutton, 1988). La ecuación de actualización de los valores Q es la que sigue (Sutton y Barto, 2018):

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r(s,a) + \gamma \max_{a'} Q'(s',a') - Q(s,a)] \tag{2.7}$$

De la ecuación anterior, la variable $r(s,a)$ es la recompensa obetenida en el estado actual dada la acción a que se tomó. El parámetro γ corresponde al factor de descuento, su valor es fijo y puede tomar el rango de valores entre [0,1]. El símbolo α es el parámetro de aprendizaje, mientras que $Q(s,a)$ es el valor Q del estado actual dada la acción a. Por último, $\max_{a'} Q'(s',a')$ representa el máximo valor Q estimado del siguiente estado.

Por otro lado, el Q-learning se considera un *policy off-learner* debido a que aprende el valor de la política óptima independientemente del tipo de acciones que tome el agente durante su fase de aprendizaje, es decir, que las acciones que elige el agente en la fase de entrenamiento pueden ser distintas a la política óptima final (Jang et al., 2019). Esto queda más claro al considerar que en Q-learning el agente trabaja con una política ϵ-greedy, lo cual es muy beneficioso para el aprendizaje del agente. Sin embargo, el algoritmo de Q-learning se considera optimista respecto al criterio de selección de las acciones, ya que asume que la acción con el valor esperado más alto siempre se va a ejecutar en el futuro (Gaskett, 2003), aunque sabemos que en el mundo real no siempre es así, ya que las acciones del agente se pueden ver influenciadas por el ambiente u otros factores. Esto se puede apreciar mejor con el ejemplo de la siguiente imagen.

La figura anterior ilustra el comportamiento generado por un algoritmo de Q-learning con $\alpha = 0.5$ $\gamma = 1$ en la tarea *Cliff-walking*. La posición inicial del agente

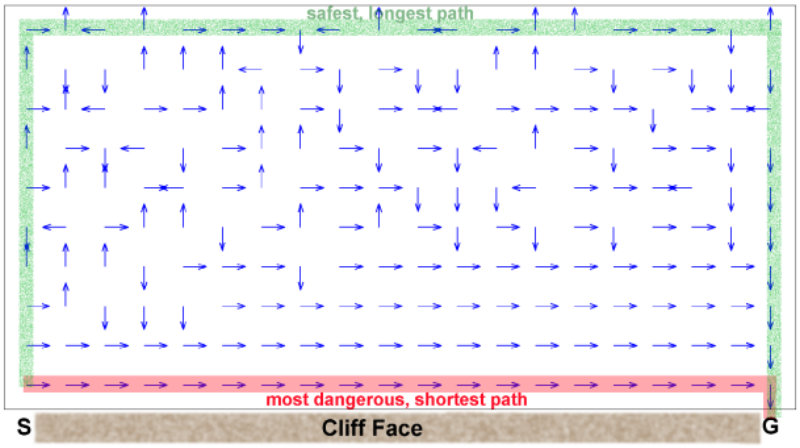


Figura 2.3: Agente Q-learning con incertidumbre en la selección de acciones encuentra una solución riesgosa en la tarea *cliff-walking* (Gaskett, 2003).

es la **S** y la meta es llegar a la **G**, mientras que en la parte de abajo - en color café - se encuentra el precipicio (*the cliff*) y si el agente cae en él le cuesta 100 puntos y se acaba el episodio, haciendo un símil con morir. En cambio, cada paso temporal le cuesta al agente 1 punto y llegar a la meta concluye el episodio, además de no restarle puntos. Por lo tanto, aunque lo óptimo es llegar a la meta con la menor cantidad de pasos temporales, hay que considerar que hay un 10 % de incertidumbre en la selección de acciones y por ende puede ser peligroso pasar justo al lado del precipicio (por la probabilidad de elegir una acción que mate al agente). Las flechas representan la acción estimada más alta por el agente en cada estado después de 10,000 ensayos, lo cual se puede obtener eligiendo el máximo valor Q por cada estado, aunque también hay que considedar que al comienzo estos valores se inicializan de forma aleatoria y por eso en la parte de arriba se ven flechas en diferentes direcciones sin un patrón particular, reflejando que no se llegaron a explorar esos estado. Además, la ruta marcada en verde es la más segura, mientras que la marcada en rojo es la más peligrosa. También, para hacer la simulación más justa, en la simulación realizada por Gaskett (2003), el agente no podía caer al precipicio en desde el estado de inicio.

Como se aprecia en la imagen, el algoritmo de Q-learning elige la ruta roja como la más óptima, ya que todas las flechas están hacia la izquierda. Esto es así porque el modelo siempre considera que en el futuro se legirá la mejor opción, ignorando

la incertidumbre a poder seleccionar otra opción y caer en el precipicio. Por ende, el Q-learning no es un algoritmo tan robusto para problemas bajo incertidumbre en la selección de acciones.

B-pessimistic Q-learning

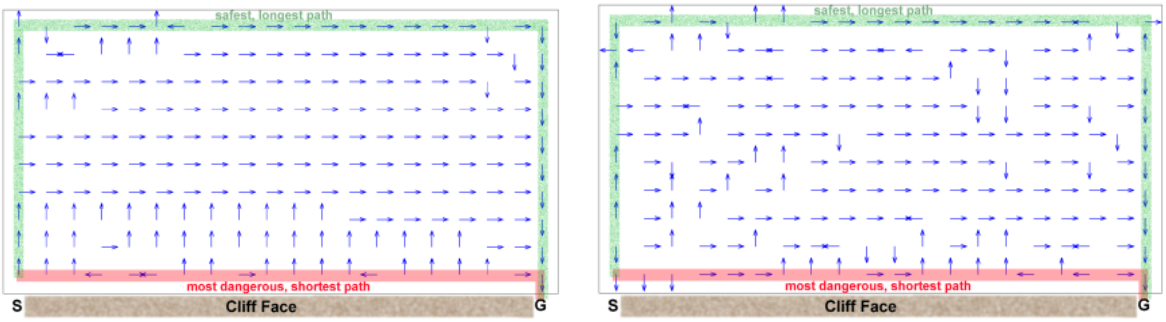
Varios de los algoritmos de RL abordan la incertidumbre de la transición entre estados, sin embargo, la incertidumbre respecto a la selección de las fututras acciones también es importante considerarla, en especial para las tareas de decisión secuencial donde las acciones se deben elegir considerando el futuro (Gaskett, 2003). Es decir, se ha de considerar que un agente no siempre tiene un completo control de sus propias acciones, como en el algoritmo B-pessimistic Q-learning (2.9) propuesto por Gaskett (2003) y luego usado por Zorowitz et al. para modelar la ansiedad (2020). Esto es se vuelve crucial sobretodo al recordar que los agentes pueden llegar a tener conductas subóptimas en el mundo real al contemplar un peligro potencial, por ejemplo, como esperar llegar a fallar en tomar una acción protectora en alguno de los futuros estados para evitar el peligro (Zorowitz et al., 2020).

$$\beta = \omega \max_{a'} Q'(s', a') + (1 - \omega) \min_{a'} Q'(s', a') \tag{2.8}$$

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[r(s, a) + \gamma\beta - Q(s, a) \right] \tag{2.9}$$

La función 2.9 es una extensión del algorimto de Q-learning (2.7) que adiciona el módulo β-pessimistic (2.8) con el parámetro de peso ω para controlar el grado de pesimismo u optimismo en el agente. Adicionalmente, ω sólo puede tomar los valores del intervalo [0,1] y es un parámetro de sensibilidad a la incertidumbre en la selección de las acciones (Gaskett, 2003). Por lo tanto, ω puede considerarse que refleja el nivel de autoeficacia del agente, es decir, su nivel de confianza para ejecutar los comportamientos necesarios con el fin de producir logros de rendimiento específicos (Bandura, 1978). Como resultado, un agente pesimista ($\omega = 0$) espera actuar en

contra de sus preferencias y por ende estima transitar al siguiente estado con el menor valor esperado, mientras que un agente optimista ($\omega = 1$) considera que tiene un control absoluto de sus acciones para maximizar la ganancia esperada (Zorowitz et al., 2020), como en el modelo estándar de Q-learning. En cambio, si ω tiene un valor intermedio entre 0 y 1 como 0.3, entonces se hace una ponderación del valor esperado del siguiente estado entre el nivel de optimismo (0.3) y pesimismo (0.7) del agente, haciendo que un área segura en el espacio de estados no se vea fuertemente afectada por cuál acción es elegida (Gaskett, 2003). Asimismo, por cómo se utiliza ω en el módulo β-pessimistic es que se puede considerar a ω como un parámetro de sensibilidad a las recompensas, al igual que su complemento $(1-\omega)$ un parámetro de sensibilidad a los castigos. Esto se puede apreciar mejor en el ejemplo de la siguiente imagen.



(a) $\omega = 0.8$ (b) $\omega = 0.5$

Figura 2.4: Agente B-pessimistic Q-learning con incertidumbre en la selección de acciones en la tarea *cliff-walking* (Gaskett, 2003).

Usando la misma tarea experimental que en el subapartado de Q-learning, en la figura 2.4 se puede ver que en el caso de la izquierda donde el agente tiene un mayor grado de optimismo ($\omega = 0.8$) que de pesimismo ($1 - \omega = 0.2$), este elige un camino seguro tomando varios pasos lejos del precipicio para llegar a la meta, sobretodo cuanto más lejos está de ella. Esto debido a que considera que no siempre tiene el control de sus acciones y el riesgo que esto conlleva. En cambio, en el caso de la izquierda cuando $\omega = 0.5$, el agente salta al precipicio desde el estado inicial, lo cual se debe a que considera tan aversivo el ambiente que prefiere terminar el episidio lo antes posible. Sin embargo, si logra acercarse a la meta entonces elige un camino un

poco más seguro. Comparando ambos casos, se puede apreciar que valores bajos de ω y por ende valores altos de su complemento $(1 - \omega)$ generan una mayor sensibilidad a los castigos.

Este algoritmo es mucho más robusto que el Q-learning, sin embargo, no incluye un módulo de planeación como los algoritmos basados en modelos. Esto es esencial para modelar la ansiedad, ya que esta es una respuesta anticipatoria a una amenaza futura y, por ende, además del aprendizaje directo en el ambiente las simulaciones sobre escenarios futuros son cruciales. Por ello, un algoritmo híbrido sería lo más apropiado para modelar la ansiedad.

Dyna-Q

La familia de algoritmos Dyna implementa la planeación *off-line*, la cual logra el agente a partir de aprender de la repetición de su conducta registrada al experimentar en el ambiente (Momennejad et al., 2018). Por ello esta familia de algoritmos se considera híbrida, ya que al inicio el agente no cuenta con un modelo del entorno, sino que lo va contruyendo a partir de lo que experimenta y, luego, lo usa para hacer simulaciones de las interacciones agente-entorno. El módulo Dyna puede integrarse a diferentes tipos de algoritmos, por ejemplo, Dyna-Q es una adaptación de Q-learning adicionando este módulo. A continuación, se muestra el pseudocódigo del agoritmo Dyna-Q (Sutton y Barto, 2018), en el cual se puede observar que las simulaciones de las experiencias pasadas en Dyna-Q se eligen aleatoriamente y, también, anticipadamente se especifica el número de simulaciones (n) a realizar. Si bien esto puede no ser lo ideal, es útil como una aproximación a la conducta humana y lo más usando en una versión sencilla de un módulo Dyna, empero ya hay variaciones más complejas que no se abordarán en este trabajo debido a la falta de literatura al respecto con la ansiedad.

Se inicializan $Q(s,a)$ y $Model(s,a)$ $\forall$ $s \in S$ y $a \in A(s)$.

Loop infinito (o hasta que termine el número de episodios programados):

a) **Se obtiene el estado actual:**

$s \longleftarrow$ estado actual (no terminal).

b) **Se selecciona una acción:**

$a \longleftarrow$ ϵ-greddy(s,Q)

c) **Ejecuta a; observa la recompensa recibida, r, y el siguiente estado, s'.**

d) **Se actualiza el valor Q para el estado s y la acción a:**

$$Q(s,a) \longleftarrow Q(s,a) + \alpha[r(s,a) + \gamma \max_{a'} Q'(s',a') - Q(s,a)]$$

e) **Se actualiza el modelo del entorno guardando la transición observada del siguiente estado, s', y la recompensa obtenida, r, para el estado s y la acción a:**

$Model(s,a) \longleftarrow r, s'$ (asumiendo un ambiente determinista).

f) **Planificación *off-line* basada en el modelo del entorno.**

Repite n veces:

$s \longleftarrow$ estado aleatorio observado previamente.

$a \longleftarrow$ acción aleatoria ejecutada con anterioridad en s.

$r, s' \longleftarrow Model(s,a)$

$$Q(s,a) \longleftarrow Q(s,a) + \alpha[r(s,a) + \gamma \max_{a'} Q'(s',a') - Q(s,a)]$$

*En psicología es común inicializar Q y $Model$ con todos sus valores en 0.

* Hiperparámetros: γ, ϵ, α y n.

También, en el pseudocódigo se puede apreciar que el modelo del entorno ($Model$) se actualiza solo con experiencia directa en el ambiente, mientras que un valor Q se puede actualizar tanto con la expriencia directa como con la planeación *offline*. Estas características hacen a Dyna-Q y a los modelos derivados mucho más cercanos a cómo opera un agente humano, ya que incluso hay evidencia de que las personas usan mecanismos de planeación *offline* al estar bajo incertidumbre en tareas de decisión secuencial (Momennejad et al., 2018), consideración especialmente importante para estudiar la ansiedad.

Por otra parte, la desventaja de las variantes de Q-learning con un módulo Dyna es que solo consideran un paso a futuro (planeación *online*) en su experiencia directa en el ambiente. Esto termina siendo un limitante importante para las tareas de decisión secuencial de múltiples estados que no tienen un contexto muy complejo, ya que es común que las personas puedan planear más allá de un paso temporal.

Succesor Representator (SR)

Sucessor Representator (SR) es un algoritmo híbrido que se basa principalmente en la planeación de varios pasos temporales sobre los estados que se estima visitar a futuro (representaciones sucesoras) durante su experiencia directa en el ambiente (Dayan, 1993), por lo que integra un tipo de la planeación *online*. A la par, este algoritmo aprende la función de recompesa de forma independiente, una diferencia esencial de los algoritmos anteriores (Dayan, 1993; Gershman, 2018). Estas dos aspectos en conjunto permiten modelar decisiones semi-flexibles en el agente (Momennejad et al., 2017), ya que aprender las representaciones sucesoras junto con la función de recompensa es similar a construir un modelo del entorno y, por ende, el SR tiene una mayor flexibilidad a los algoritmos *model-free*, aunque menor a los *model-based* (Momennejad et al., 2017). Por ejemplo, el SR puede captar de forma más rápida que el Q-learning los cambios en las recompensas, en contraste, aunque el SR también es sensible a los cambios entre las transiciones entre estados, no lo es tanto como un algoritmo *model-based*, ya que el SR no considera la propabilidad de transitar a un estado, sino solo que se transitará a un estado con mayor frecuencia (Gershman, 2018).

Asimismo, el SR se puede considerar como un método de rápida generalización que se basa en estimar las relaciones de ocupación entre los estados de acuerdo a la política que se sigue (Ducarouge y Sigaud, 2017; Momennejad et al., 2017). Estas relaciones se guardan en la matriz M, por lo que una entrada de esta matriz, $M(s, s'))$, codifica la ocupancia descontanda del estado futuro s' promediando sobre las trajectorias iniciadas en el estado s, como se muestra en la ecuación 2.10 (Ducarouge y

Sigaud, 2017; Gershman, 2018).

$$M(s, s') = E\left[\sum_{t=0}^{\infty} \gamma^t I_{[s_t=s']} \middle| s_0 = s\right] \tag{2.10}$$

Donde $I_{[s_t=s']}$ es un función indicadora, por lo que $I_{[s_t=s']} = 1$ si el argumento de la función es verdadero ($s_t = s'$) o $I_{[s_t=s']} = 0$ si no se cumple la condición de la función. Por otra parte, en la función de actualización de lo los estados succesores se puede usar el método de diferencias temporales como se muestra en la función 2.11, donde $M(s, \mathbf{s}')$ es un vector que guarda la relación descontada de ocupación entre el estado actual con los diferentes estados futuros (s'), asignando valores más altos a los estados que se podría transitar en un futuro más cercano y valores más bajos a los que se puede llegar a transitar en un futuro lejano.

$$M(s_t, \mathbf{s}') \longleftarrow M(s_t, \mathbf{s}') + \alpha\left(I_{[s_t=s']} + \gamma M(s_{t+1}, \mathbf{s}') - M(s_t, \mathbf{s}')\right) \tag{2.11}$$

Simplificando la notación, otra forma de escribir la ecuación anterior se muestra en la ecuación 2.12. Donde s representa el estado actual, s' el estado siguiente considerando un paso temporal a futuro y s'' el estado siguiente considerando dos pasos temporales a futuro respecto del estado actual.

$$M(s, \mathbf{s}') \longleftarrow M(s, \mathbf{s}') + \alpha\left(I_{[s=s']} + \gamma M(s', \mathbf{s}'') - M(s, \mathbf{s}')\right) \tag{2.12}$$

Finalmente, el SR usa dos estructuras que se combinan de forma lineal para computar el valor de un estado (Gershman, 2018; Momennejad et al., 2017): un vector R (codifica la recompensa inmediata esperada en cada estado de acuerdo a la función de recompensa) y una matriz M (con las representaciones sucesoras de los estados), ya que así se obtiene una evaluación de la trajectoria óptima a la recompensa. Cada entrada del vector R corresponde al refuerzo esperado de un estado del entorno y se actualiza con la ecuación 2.13, mientras que para obtener el valor estimado de cada

estado se usa la ecuación 2.14.

$$R(s) = R(s) + \alpha(r - R(s)) \tag{2.13}$$

$$V(s) = \sum_{s'} M(s, s')R(s') \tag{2.14}$$

Por otro lado, se pueden construir las representaciones sucesoras de los estados considerando la acción inicial, a, en el estado inicial (s - estado en el que se encuentra el agente) como se muestra en la ecuación 2.15 (Ducarouge y Sigaud, 2017). En consecuencia, ya no sería una matriz M la que guarde esta información, sino un tensor M (objeto algebraico de tres dimensiones) y, por ende, su regla de actualización también cambiaría (2.16). Además, con esta modificación en vez de estimar los valores de los estados, $V(s)$, se pueden obtener los valores Q(s,a) (2.17). La ventaja de trabajar con los valores Q es que permiten capturar mejor la incertidumbre asociada con las acciones individuales, lo cual es importante para modelar la ansiedad. Por otra parte, como las recompensas se aprenden de forma independiente a las representaciones sucesoras, entonces no es necesario que se cambie la función de recompensa y, por lo tanto, se puede seguir usando la ecuación 2.13 para actualizar cada entrada del vector R.

$$M(s, \mathbf{s}', a) = E\left[\sum_{t=0}^{\infty} \gamma^t I_{[s_t = s']} \middle| s, a\right] \tag{2.15}$$

$$M(s, \mathbf{s}', a) \longleftarrow M(s, \mathbf{s}', a) + \alpha\left(I_{[s=s']} + \gamma M(s', \mathbf{s}'', a') - M(s, \mathbf{s}', a)\right) \tag{2.16}$$

$$Q(s, a) = \sum_{s'} M(s, s', a)R(s') \tag{2.17}$$

Para seleccionar a' en la regla de actualización $M(s, s', a)$ (2.16), se usa la

ecuación 2.18. Es decir, se elige la acción que maximiza las ganacias esperadas considerando los valores Q estimados hasta el momento por el agente, como en el algoritmo de Q-learning. Por último, para conocer con mayor claridad el orden de los diferentes cálculos implementados en algormitmo SR, en el anexo 1 se encuentra su pseudocódigo.

$$a' = \arg\max_{a} Q(s', a') \tag{2.18}$$

Adicionalmente, evidecia reciente sugiere que en la familia de algoritmos de reinforcement learning, el SR sería el más similar a la conducta humana para tareas de evaluación secuencial de múltiples pasos, en especial al combinarlo con estructuras de planeación *offline*, es decir, con un módulo *model-base* (Momennejad et al., 2017). No obstante, actualmente no hay investigaciones que contrasten las diferentes combinaciones posibles que se podrían hacer entre los distintos tipos de módulos *model-base* con SR, por lo que las diferencias entre ellas serían especulativas, aunque en cuestión de practicidad un módulo Dyna es fácil de adaptar y menos costoso computacionalmente que otras opciones. Por último, integrar un módulo *offline* es relevante al trabajar en tareas bajo incertidumbre, ya que la sensibilidad a la incertidumbre predice la repetición y replanación de escenarios pasados (Momennejad et al., 2018), permitiendo además una integración o generalización más rápida de la información. En especial, esto es importante considerarlo al estudiar la ansiedad, por el impacto que tiene la incertidumbre sobre esta.

Por otra parte, hacer adaptaciones al SR resulta ideal para modelar y contrastar las hipótesis debatidas en la literatura de la ansiedad, específicamente sobre la importancia de la diferencia en las tasas de aprendizaje de los castigos y las recompensas, en comparación con la relevancia de la sensibilidad a los castigos para generar la conducta ansiosa. Esto se debe principalmente a cuatro cosas:

1. Adaptar un módulo Dyna a SR es muy práctico para cubrir la parte de planea-

ción *offline* en un agente.

2. El SR aprende el refuerzo esperado por cada estado de forma independiente a las representaciones sucesoras, por lo que es factible generar una variación del modelo Dyna-SR con tasas de aprendizaje diferentes para las recompensas (α_r) y los castigos (α_c), posibilitando medir la contribución de esta diferencia a la conducta generada.

3. Se puede adaptar al Dyna-SR una variación β-pessimistic usando como referencia la ecuación 2.18 para crear una ecuación similar que minimice las ganancias esperadas ($a^- = \arg\min_{a'} Q(s', a')$), para luego usar la mejor y peor acción del siguiente estado a futuro en una variación del módulo B-pessimistic en la función de actualización 2.16. Esto con el fin de medir el impacto de los parámetros de sensibilidad a los castigos y las recompensas a la conducta del agente.

4. Al ser variaciones de un mismo modelo, los dos modelos generados permiten una comparación más sencilla del impacto de los diferentes parámetros en el comportamiento del agente.

Considerando lo anterior y que hasta la fecha no se han contrastado ambas hipótesis en experimentos de evaluación secuencial de múltiples pasos temporales, en este trabajo se realizará el modelamiento computacional de la conducta ansiosa a partir de los modelos propuestos y se compararán los resultados. Esto con el objetivo de esclarecer el nivel de contribución de cada hipótesis al desarrollo de la conducta ansiosa en ambientes más naturales al objeto de estudio.

3. Metodología

"The important thing in science is not so much to obtain new facts as to discover new ways of thinking about them."

William Henry Bragg

Se modeló la toma de decisiones ansiosa en un proceso de decisión Markoviano (MDP) utilizando dos tipos de agentes de aprendizaje por refuerzo en un ambiente virtual. La principal diferencia entre los agentes radica en los modelos propuestos que utiliza cada uno: el primero emplea el modelo Dyna β-pessimistic SR, el cual incorpora parámetros de sensibilidad para los castigos y las recompensas, mientras que el segundo tipo de agente usa el modelo Dyna α-SR que utiliza diferentes tasas de aprendizaje para las recompensas y los castigos. El experimento se diseñó adaptando la tarea *cliff walking* a un entorno determinista, de tiempo discreto y con horizonte finito, en la que la estocasticidad se presenta en la selección de acciones del agente. La implementación de ambos tipos de agentes en la tarea permitió evaluar el impacto de los distintos parámetros y sus diferentes valores para el desarrollo de la conducta ansiosa, así como realizar una comparación entre los modelos. A continuación, se describen los modelos, los agentes, la tarea experimental y el experimento.

3.1. Modelos

Se propusieron dos modelos para simular la conducta ansiosa implementando modificaciones al modelo DynaSR con el método de diferencias temporales: Dyna β-pessimistic SR y Dyna α-SR. El primero se basa en la importancia de los parámetros de sensibilidad para los castigos y las recompensas, mientras que el segundo se enfoca en la relevancia de la diferencia en las tasas de aprendizaje de los castigos y las recompensas. A continuación, se presenta el pseudocódigo de cada modelo.

Dyna β-pessimistic SR

Se inicializa $R(s)\ \forall\ s \in S$.

Se inicializa $Q(s,a)$ y $Model(s,a)\ \forall\ s \in S$ y $a \in A(s)$.

Loop hasta que episodio = 200:

a) $s \longleftarrow$ estado actual (no terminal).

b) $a \longleftarrow \epsilon$-greddy(s,Q)

c) Ejecuta a; observa la recompensa recibida, r, y el siguiente estado, s'.

d) $a^{+} \longleftarrow \arg\max_{a} Q(s',a')$

e) $a^{-} \longleftarrow \arg\min_{a} Q(s',a')$

f) $\delta^{M} \longleftarrow \omega M(s',\mathbf{s''},a^{+}) + (1-\omega)M(s',\mathbf{s''},a^{-})$

g) M(s,s' ,a) $\longleftarrow M(s,\mathbf{s'},a) + \alpha\left(I_{[s=s']} + \gamma\delta^{M} - M(s,\mathbf{s'},a)\right)$

h) R(s) $\longleftarrow R(s) + \alpha(r - R(s))$

i) $Q(s,a) \longleftarrow \sum_{s'} M(s,s',a)R(s')$

j) $Model(s,a) \longleftarrow r, s'$

k) Repite n veces:

$s \longleftarrow$ estado aleatorio observado previamente.

$a \longleftarrow$ acción aleatoria ejecutada con anterioridad en s.

$r, s' \longleftarrow Modelo(s,a)$

Ejecuta los pasos del **d** al **i** usando los argumentos s, a, r y s'.

Hiperparámetros: ϵ, α, ω, n y el número de episodios programados.

La diferencia entre el algoritmo Dyna β-pessimistic SR y un algortimo tradicio-

nal DynaSR es la implementación del módulo $\beta - pessimistic$. Este cambio impacta directamente en la actualización de cada entrada del tensor M, es decir, en cómo se aprenden las representaciones sucesoras. Para ello, se implementan los pasos e y f, así como las modificaciones pertinentes al paso g (en el algoritmo anterior). En comparación, para el algoritmo Dyna α-SR el cambio es en cómo se aprenden los refuerzos de cada estado, es decir, la regla de actualización para cada entrada del vector R (el paso g en el siguiente algoritmo).

Dyna α-SR

Se inicializa $R(s)\ \forall\ s \in S$.

Se inicializa $Q(s,a)$ y $Model(s,a)\ \forall\ s \in S$ y $a \in A(s)$.

Loop hasta que episodio = 200:

a) $s \longleftarrow$ estado actual (no terminal).

b) $a \longleftarrow \epsilon$-greddy(s,Q)

c) Ejecuta a; observa la recompensa recibida, r, y el siguiente estado, s'.

d) $a^+ \longleftarrow \arg\max_a Q(s',a')$

f) $M(s,\mathbf{s}',a) \longleftarrow M(s,\mathbf{s}',a) + \alpha\left(I_{[s=s']} + \gamma M(s',\mathbf{s}'',a^+) - M(s,\mathbf{s}',a)\right)$

g) Se actualiza R(s):

Si r ≥ 0 : R(s) $\longleftarrow R(s) + \alpha^+(r - R(s))$

Si r <0: $R(s) \longleftarrow R(s) + \alpha^-(r - R(s))$

h) $Q(s,a) \longleftarrow \sum_{s'} M(s,s',a)R(s')$

i) $Model(s,a) \longleftarrow r, s'$

j) Repite n veces:

$s \longleftarrow$ estado aleatorio observado previamente.

$a \longleftarrow$ acción aleatoria ejecutada con anterioridad en s.

$r, s' \longleftarrow Modelo(s,a)$

Ejecuta los pasos del **d** al **h** usando los argumentos s, a, r y s'.

Hiperparámetros: ϵ, α, α^+, α^-, n y el número de episodios programados.

3.2. Agentes

Se diseñaron dos tipos de agentes de aprendizaje por refuerzo que implementan detrás modelos dintintos: Dyna β-pessimistic SR para el agente tipo A y Dyna α-SR para el agente tipo B. En los agentes tipo B lo que varía son las tasa de aprendizaje para las recompensas (α^+) y los castigos (α^-). En cambio, para los agentes tipo A lo que cambia entre ellos es el parámetro de sensibilidad a los refuerzos, ω, y su complemento que es el parámetro de sensibilidad a los castigos, $1-\omega$. Los hiperparámetros compartidos entre los modelos se fijaron de la misma manera para ambos tipos de agentes con $\epsilon = 0.2$ (probabilidad de que se elija una acción de forma aleatoria), $\alpha = 0.1$ (tasa de aprendizaje de las representaciones sucesoras), $\gamma = 0.9$ (factor de descuento), $n = 3$ (recall o simulaciones del módulo Dyna) y número de episodios = 200.

3.3. Tarea experimental

Se implementó una variación de la tarea *cliff walking* en un *gridworld* de 9x9 con un entorno de tiempo discreto, horizonte finito y determinista, debido a que los refuerzos y las transiciones entre estados son fijos. En este ambiente la estocasticidad está en la selección de acciones del agente, ya que este usa una política $\epsilon - greedy$ para elegir la acción que implementará en cada estado. Esta política se usa tanto en la fase de entrenamiento del agente (200 episodios) como en la prueba final (1 *rollout*), con el fin de simular que el agente no tiene un completo control de sus acciones.

Como se muestra en la figura 3.1, el ambiente tiene solo 8 estados que dan al agente un refuerzo diferente de cero: 6 estados dan un refuerzo negativo igual a -1.0 (en color rojo) y uno da un refuerzo positivo igual a 1.0 (en color azul). El objetivo de un agente en este entorno es llegar a la meta sin caer en un estado aversivo, aunque en un inicio el agente no conocerá el ambiente y, por lo tanto, necesitará explorarlo para encontrar la ruta óptima. Asimismo, las acciones que puede realizar un agente en el ambiente son 4: arriba, abajo, derecha e izquierda, siempre y cuando no se choque con

una pared (cuadros en gris oscuro). Por último, las condiciones que pueden terminar un episodio son las siguientes: el agente llega a la meta, el agente cae en un estado aversivo (cuadros rojos) o se alcanza el límite de pasos temporales por episodio (100).

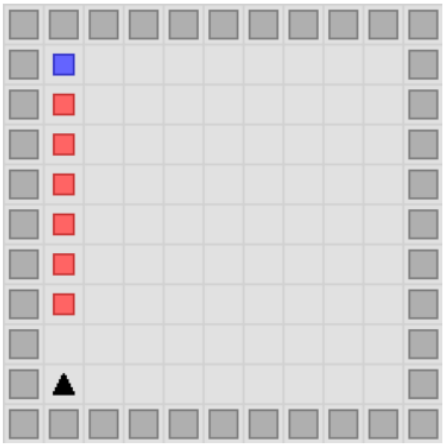


Figura 3.1: Entorno virtual de la tarea *cliff-walking*. El cuadro azul representa la meta, los cuadros rojos el acantilado y el triángulo la posición inicial del agente.

3.4. Experimento

Se comparó el desempeño de los modelos Dyna β-pessimistic-SR y Dyna α-SR para simular la conducta ansiosa en la tarea experimental de evaluación secuencial *cliff walking*. Para ello se realizaron 3 simulaciones de agentes tipo A con valores de $\omega = [1.0, 0.5, 0.0]$, con el objetivo de contrastar el impacto de los distintos valores de ω en la conducta de los agentes. Adicionalmente, se realizaron 3 simulaciones de agentes tipo B con valores de $\alpha^- = [0.05, 0.075, 0.1]$, manteniendo en todos los casos constante el valor de $\alpha^+ = 0.05$. Los valores de α^+ y α^- se eligieron considerando que en la literatura el rango normal de una tasa de aprendizaje para que converjan adecuadamente los algoritmos de *reinforcement learning* en tareas de decisión sequencial es de [0,0.1] (Aylward et al., 2019). Estas últimas simulaciones permitieron evaluar el cambio que generan los distintos rangos de diferencia entre las tasas de aprendizaje de los refuerzos positivos y negativos para la toma de decisiones del agente: cuando α^- y α^+ valen lo mismo, cuando α^-es 50 % mayor que α^+ y cuando α^- es 200 % mayor que α^+. Finalmente, considerando todas las simulaciones se pudo analizar la contribución

de la diferencia en las tasas de aprendizaje y los parámetros de sensibilidad de las recompensas y los castigos para el desarrollo de la conducta ansiosa.

Condición	**Agentes tipo A**	**Agentes tipo B**
1	$\omega = 1.0$	$\alpha^- = 0.050$
2	$\omega = 0.5$	$\alpha^- = 0.075$
3	$\omega = 0.0$	$\alpha^- = 0.100$

Tabla 3.1: Variaciones en las diferentes condiciones experimentales para cada tipo de agente.

Por otro lado, las conductas que se buscaron observar relacionadas con la ansiedad se operacionalizaron de la siguiente manera:

- **Sobreestimación del peligro:** Asignación a los estados valores de peligro mayores al real.
- **Generalización del peligro:** Consideración como aversivos a los estados neutros que están lejanos de los estados aversivos.
- **Evitación:** El agente se aleja de los estados aversivos.
- **Aversión al riesgo:** El agente en vez de explorar más el ambiente, explota una zona de estados segura.

Por último, todas las simulaciones se hicieron implementando el lenguaje de programación de Python con la versión 3.11.11, usando como base la librería neuronav para progamar los agentes y la tarea experimental (Juliani et al., 2022). Todo el código está disponible públicamente en el siguiente link: https://github.com/Alicia-MJ/tesis-modelamiento-computacional-de-la-ansiedad

4. Resultados

"Now is the time to understand more,
so that we may fear less."

Marie Curie

A continuación, se presentan los resultados de tres análisis distintos para cada agente en las diferentes condiciones de la tarea experimental. Primero, se generó un mapa de calor del entorno por cada agente usando el cálculo del valor estimado de cada estado, el cual se obtuvo promediando los valores Q por estado. Los estados en los mapas varían de color de acuerdo a su valor estimado: entre más intenso es el color azul significa que su valor es más cercano a 1.0 (valor máximo); valores cercanos a 0 se aprecian en color blanco; mientras que los estados con valores cercanos a -1.0 (valor mínimo) tienen un color rojo oscuro. De igual forma, en estos mapas las flechas indican la acción de mayor valor en cada estado y el rollout correspondiente a la fase de evaluación está marcado por las flechas más oscuras (en color negro). Además, se graficó el número de pasos temporales que realizó cada uno de los agentes por episodio durante la tarea experimental, así como las recompensas totales obtenidas por episodio. Los resultados de estos análisis en conjunto permiten evaluar el aprendizaje de los distintos agentes, así como analizar su conducta. Finalmente, la comparación entre ambos grupos de agentes permitirá contestar la pregunta de investigación sobre la importancia de las tasa de aprendizaje y los parámetros de sensibilidad para los

castigos y las recompensas para el desarrollo de la conducta ansiosa.

Como se muestra en la figura 4.1 el agente más optimista ($\omega = 0.1$) toma una ruta riesgosa para llegar a la meta, pasando al lado del precipicio, y estima estados cercanos al mismo de forma positiva. Es decir, subestima el peligro de estos estados. Por ello, aunque es el agente que menos pasos temporales usa en la mayoría de los episodios e incluso en muchos de ellos llega a la meta (70 % de las ocasiones), también muchas otras veces cae en el precipicio (30 % de las veces). En cambio, el agente con un nivel intermedio de optimismo y pesimismo ($\omega = 0.5$) toma una ruta más segura para llegar a la meta al alejarse del precipicio. También, los estados cercanos a él los considera peligrosos, ya que están en tonalidades rojas. Además, es el agente que llega más veces a la meta (en el 97 % de los episodios) y después del episodio 50 se estabiliza el número que usa de pasos temporales. De igual forma, este agente solo en el 2 % de los episodios cae en un estado aversivo y se queda en estados neutros hasta el final solo el 1 % de los episodios. Por último, el agente pesimista ($\omega = 0.0$) estima gran parte del ambiente como aversivo y se aleja del precipicio, pero casi nunca consique llegar a la meta (el 87 % de los episodios se queda en estados neutros, alrededor del 1 % llega a la meta y alrededor del 0.02 % cae en un estado aversivo). Es decir, la estimación negativa de los estados sesga la estimación de las recompensas y en vez de buscar llegar a la meta, el agente pesimista prioriza evitar el peligro y es por eso que en la mayoría de los casos su recompensa total por episodio es igual a 0, sobretodo a partir desde alrededor del episodio 90. Esto último, se puede considerar como una conducta de aversión al riesgo, debido a que el agente explota estados de una zona segura en vez de explorar más el ambiente para encontrar la meta. Adicionalmente, los agentes con ($\omega = 0.5$) y ($\omega = 0.0$) presentan conductas de evitación, pero es el agente pesimista el que sobreestima y generaliza el peligro de forma excesiva.

Por otro lado, en la figura 4.2 se aprecia que los tres agentes tipo B con el modelo Dyna α-SR tienen una conducta y estimación del entorno muy similar, subestimando el peligro en el ambiente y eligiendo una ruta a la meta poco segura. Además,

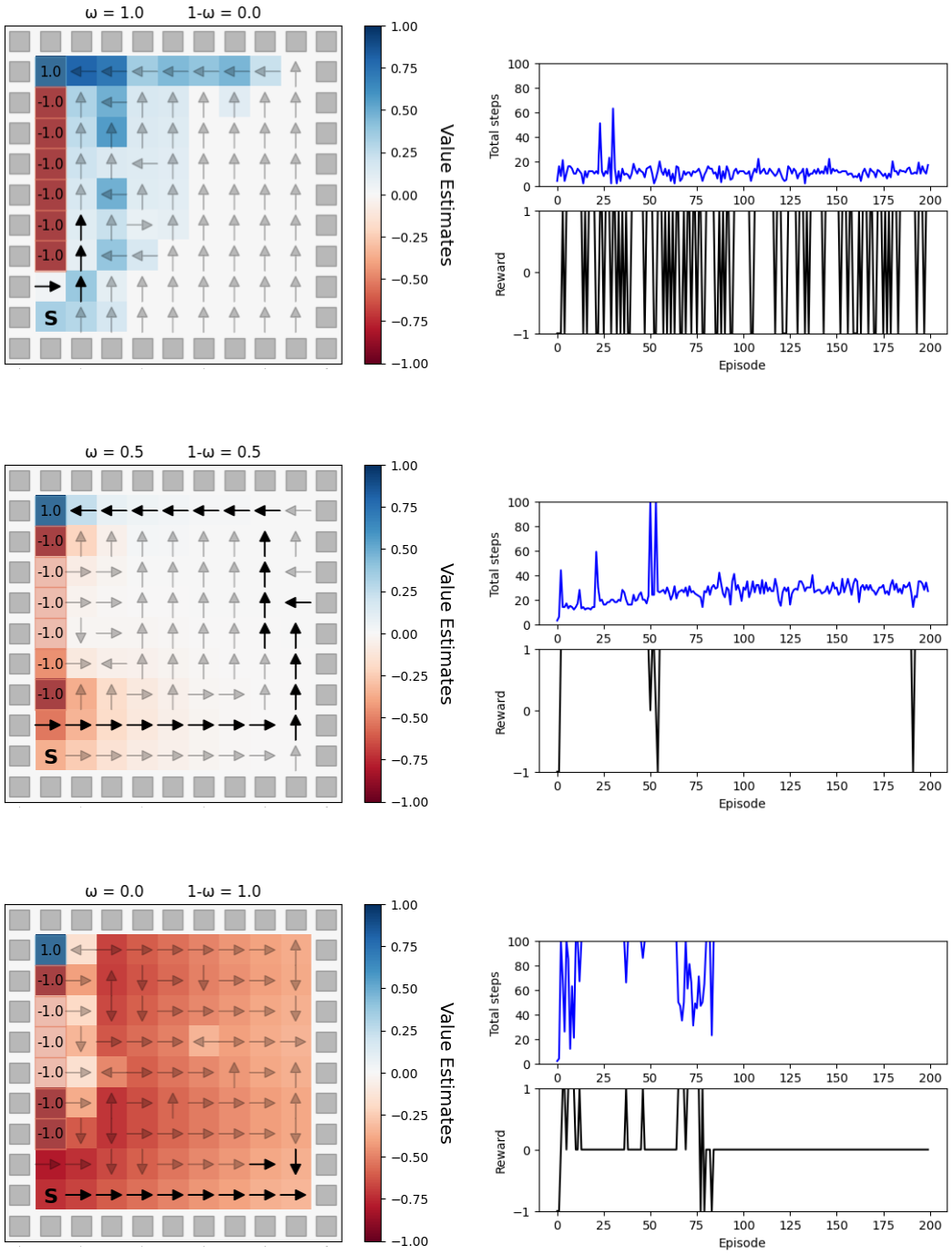


Figura 4.1: **Resultados de los agentes tipo A - Dyna β-pessimistic SR.** El bloque superior corresponde a los resultados de agente optimista ($\omega = 1.0$), el bloque intermedio del agente con $\omega = 0.5$, y el bloque inferior del agente pesimista ($\omega = 0.0$). A la izquierda de cada bloque se muestran los mapas de las estimaciones del entorno; en la parte superior derecha, la gráfica del número de pasos temporales por episodio; y en la parte inferior derecha, la gráfica de la recompensa obtenida por episodio.

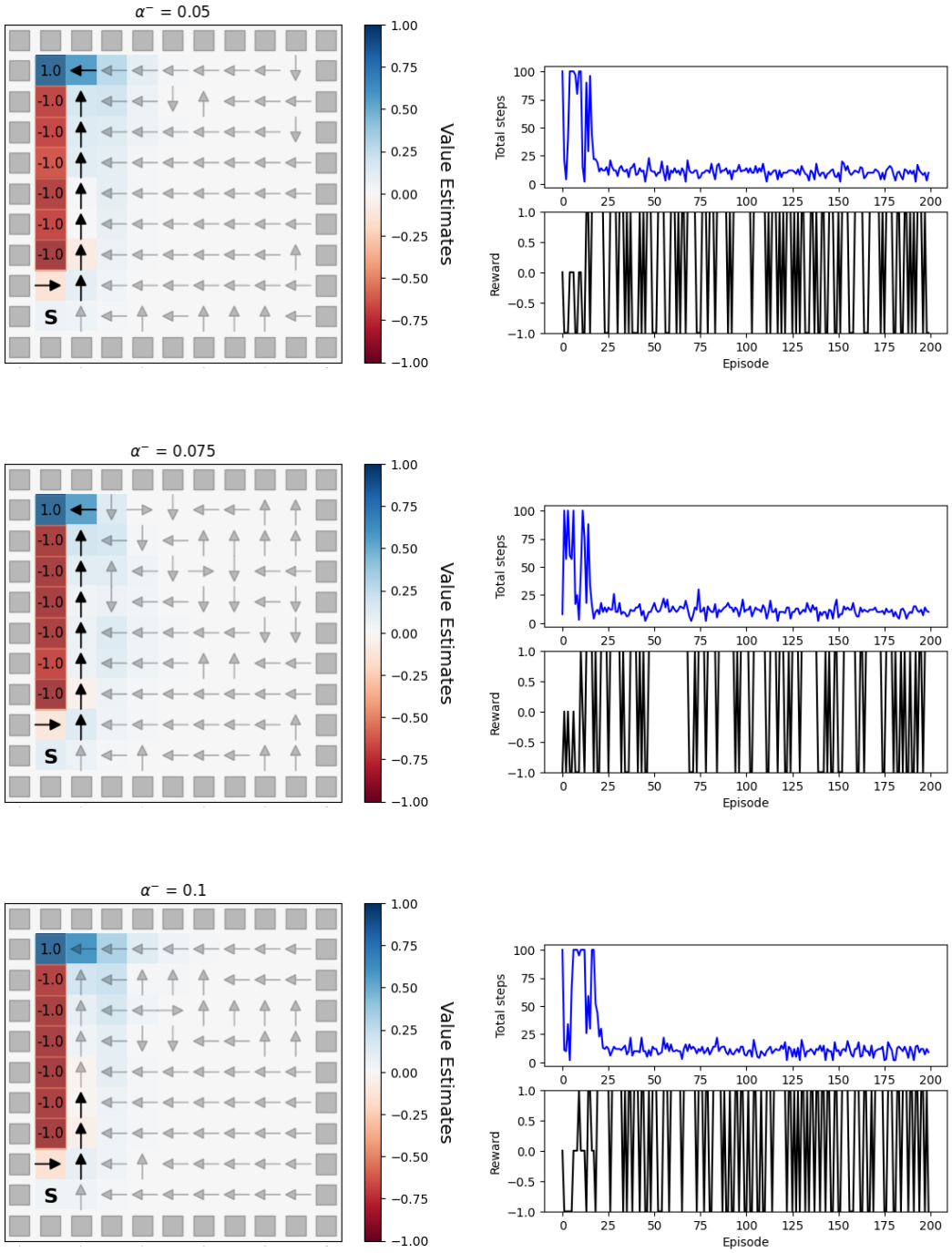


Figura 4.2: **Resultados de los agentes tipo B - Dyna α-SR.** El bloque superior corresponde a los resultados del agente con $\alpha^- = 0.05$ (ambas tasas de aprendizaje tienen el mismo valor), el bloque intermedio del agente con $\alpha^- = 0.075$ (150 % mayor que α^+), y el bloque inferior del agente con $\alpha^- = 0.1$ (200 % mayor que α^+). A la izquierda de cada bloque se muestran los mapas de las estimaciones del entorno; en la parte superior derecha, la gráfica del número de pasos temporales por episodio; y en la parte inferior derecha, la gráfica de la recompensa obtenida por episodio.

los tres agentes antes del episodio 25 han terminado de explorar el entorno, puesto que después su número de pasos temporales por episodio disminuye radicalmente y se estabiliza. De igual forma, el rollout de cada agente indica que están eligiendo la ruta más corta a la meta, pasando directamente al lado del precipicio y es el tercer agente el que no alcanza a llegar a la meta, quedándose a 6 pasos de la misma. Por último, es importante notar que estos resultados se parecen mucho a los del agente optimista con el modelo Dyna β-pessimistic SR.

5. Discusión

"Science will always be a quest...
It is a journey, never a destination."

Karl Popper

Estudiar la ansiedad en contextos más naturales al fenómeno de estudio, como son las tareas de evaluación secuencial de múltiples pasos, permitió vislumbrar la relevancia de la diferencia en los parámetros de sensibilidad para los castigos y recompensas para el comportamiento ansios. Especialmente, la diferencia en el desempeño de los agentes tipo A en la tarea experimental muestra como una mayor sensibilidad a los castigos (respecto a las recompensas) impacta en una expectativa más alta de actuar en contra de las propias preferencias para maximizar el refuerzo total a futuro. Es decir, que ser más sensible a la falta de control sobre las propias acciones lleva al agente a esperar un futuro más pesimista y al final promueve un mayor comportamiento asociado a la ansiedad como la evitación, la sobrestimación y generalización del peligro, así como aversión al riesgo. Esto puede traducirse a cuando las personas presentan desórdenes de ansiedad orientados al futuro, como la ansiedad generalizada y la ansiedad social, al tener una percepción sesgada de la amenaza y una subestimación de sus recursos de afrontamiento, lo cual puede operacionalizarse como un bajo nivel de autoeficacia que refleja la falta de creencia en su capacidad para ejecutar las conductas necesarias para alcanzar objetivos específicos de rendimiento.

Estos hallazgos concuerdan con las investigaciones de Katz et. al (2020) y Zorowitz et. al (2020), donde la sensibilidad al castigo es el principal predictor de la ansiedad. Así como con las ideas de Brown et. al (2023), Gagne y Dayan (2022) que consideran que los desórdenes de ansiedad son principalmente trastornos del aprendizaje de la incertidumbre.

En contraste, la similitud en la conducta de los agentes tipo B indica que la diferencia en la tasas de aprendizaje para los castigos y las recompensas no marca una diferencia significativa para reproducir una conducta ansiosa en la tarea experimental. Particulamente, considerando que ni siquiera sucede esto en el caso donde el agente tiene un parámetro de aprendizaje para los castigos que vale el doble que la tasa de aprendizaje de las recompensas. Esto va en contra de los hallazgos de investigaciones que han usado el modelamiento computacional para estudiar la ansiedad en tareas de un solo paso, como Aylward et. al (2019) y Pike y Robinson (2022), que consideran que la rapidez con la que se integra la información sobre el castigo a lo largo del tiempo es lo que fomenta la conducta ansiosas, además de considerar que existe una falta de evidencia del efecto de la ansiedad sobre la sensibilidad al castigo.

Los resultados llevan a cuestionar si los experimentos de un solo paso son adecuados para estudiar la ansiedad o si más bien están sesgando lo que entendemos de ella. Otra perspectiva lleva a pensar que los resultados de los diferentes tipos de tareas, ya sea de un solo paso o de evaluación secuencial, en realidad lo que hacen es esclarecer el papel que pueden jugar las tasas elevadas a los castigos en la ansiedad. Es decir, una tasa de aprendizaje elevada para los castigos se vuelve más relevante en el corto plazo, como en los casos que estudian las tareas de un solo paso y, en comparación, las tasas elevadas para los castigos pierden relevancia a largo plazo, donde un elevado parámetro de sensibilidad a los castigos es mucho más relevante para la conducta ansiosa, como en las tareas de múltiples pasos temporales. Sin embargo, considerando que la ansiedad es principalmente una respuesta anticipatoria ante una amenaza futura que se percibe como impredecible o incontrolable (American Psychological Association, 2019; Daviu et al., 2019), los resultados de las tareas

de evaluación secuencial se podrían considerar más acertados para entender los factores que impactan en la conducta ansiosa. En consecuencia, para los tratamientos terapéuticos de la ansiedad es recomendable enfatizar el trabajar con la sensibilidad al castigo del paciente.

Por otro lado, los resultados muestran que la planeación pesimista, tanto *offline* como *online*, tiene un papel central en la conducta ansiosa. Esto se debe a que es más relevante cómo impactan los castigos estimados en la actualización de las representaciones sucesoras que la velocidad con la que se aprenden en sí. Esto se debe a que hay una aberración progresiva en el mapa cognitivo que genera el agente de las predicciones sucesoras o rutas que se pueden seguir. En otras palabras, se enfatiza el rol del pensamiento castastrófico o la rumiación en el comportamiento ansioso, que sería el símil al considerar las rutas más desventajosas en la planeación *online* y *offline*, respectivamente.

Finalmente, el presente trabajo es teórico y se hizo usando métodos de modelamiento computacional, aprovechando lo que ya se conocía de la literatura sobre el tema. La tarea experimental permitió evaluar la simulación de la conducta ansiosa en un ambiente no determinista donde el agente no tiene un completo control de sus acciones. Así como determinar que entre los dos modelos, Dyna β-pessimsitic fue el mejor para modelar el comportamiento ansioso. No obstante, hay muchas cosas que se pueden estudiar en el futuro. Por ejemplo, desde la parte metodológica se podrían hacer experimentos donde el ambiente también es estocástico y, para ello, primero se tendrían que adaptar un módulo de planeación que contemple esto. Adicionalmente, se podría considerar tener una variable de recall que de acuerdo con el nivel de sorpesa en la recompensa recibida y el nivel de pesismismo del agente. De igual forma, sería importante reproducir el experimento con agentes humanos para evaluar el ajuste de los datos al modelo y hacer los cambios necesarios, por ejemplo, quizás haya relaciones entre variables que en vez de ser lineales son no-lineales. Por último, cabe recalcar que el estudio de la ansiedad desde una perspectiva computacional es relativamente reciente, por lo que seguramente en el futuro se desarrollarán modelos más completos

y precisos.

6. Conclusión

"To dare is to momentarily lose your footing,
to not dare is to lose yourself."

Søren Kierkegaard

Se extendió la investigación sobre la ansiedad en tareas de evaluación secuencial, considerando modelos híbridos de aprendizaje por refuerzo, para encontrar los aspectos de la adquisición del aprendizaje que aporta información adicional sobre el comportamiento ansioso. Se encontró que el parámetro de sensibilidad al castigo para estimar las representaciones sucesoras es esencial para el desarrollo de la conducta ansiosa, pero no una elavada tasa de aprendizaje a los castigos. Sin embargo, se reconoce el impacto que tienen los castigos en la fase de planeación. Los hallazgos van en contra de gran parte de la literatura de modelamiento computacional de la ansiedad que se enfoca en experimentos de un solo paso, lo cual enfatiza la importancia de estudiar la ansiedad en tareas más cercanas al fenómeno de estudio, como son las tareas de múltiples pasos temporales. A futuro entre las principales consideraciones está comparar los resultados de los modelos desarrollados con otras tareas experimentales donde se contemple estocasticidad en el ambiente y para ello será necesario extender el módulo de planación *offline* a ambientes no deterministas, así como hacer la validación del modelo Dyna β-pessimistic SR con datos humanos.

7. Anexos

7.0.1. Anexo 1: Pseudocódigo del algoritmo SR

Se incializa $R(s)$ $\forall$ $s \in S$.

Se inicializa $Q(s,a)$ y $Model(s,a)$ $\forall$ $s \in S$ y $a \in A(s)$.

Loop hasta que termine el número de episodios programados:

a) **Se obtiene el estado actual:**

$s \longleftarrow$ estado actual (no terminal).

b) **Se selecciona una acción:**

$a \longleftarrow$ ϵ-greddy(s,Q)

c) **Ejecuta a; observa la recompensa recibida, r, y el siguiente estado, s'.**

d) **Se selecciona la mejor acción del siguiente estado:**

$a^{+} \longleftarrow \arg\max_{a} Q(s', a')$

g) **Se actualiza la función de las representaciones sucessoras:**

M(s, : ,a) $\longleftarrow$ $M(s,:,a) + \alpha \left(I_{[s=s']} + \gamma M(s',:,a^{+}) - M(s,:,a)\right)$

h) **Se actualiza la función de recompensa:**

R(s) $\longleftarrow$ $R(s) + \alpha(r - R(s))$

i) **Se estiman los valores Q:**

$Q(s,a) \longleftarrow \sum_{s'} M(s,s',a)R(s')$

*Hiperparámetros: ϵ, α y γ

Referencias

Ahmed, S. A., & Çerkez, Y. (2020). The Impact of Anxiety, Depression, and Stress on Emotional Stability among the University Students. *Propósitos y Representaciones*, *8*(3). https://doi.org/10.20511/pyr2020.v8n3.520

American Psychiatric Association. (2014). Manual diagnóstico y estadístico de los trastornos mentales (DSM-5) [Traducción al español] (CIBERSARM, Trad.). En. Editorial Médica Panamericana.

American Psychological Association. (2019). What's the difference between stress and anxiety? [Accessed: 2024-05-28]. https://www.apa.org/topics/stress/anxiety-difference

Aylward, J., Valton, V., Ahn, W.-Y., Bond, R. L., Dayan, P., Roiser, J. P., & Robinson, O. J. (2019). Altered learning under uncertainty in unmedicated mood and anxiety disorders. *Nature human behaviour*, *3*(10), 1116-1123. https://doi.org/10.1038/s41562-019-0628-0

Bandura, A. (1978). Self-efficacy: Toward a unifying theory of behavioral change [Perceived Self-Efficacy: Analyses of Bandura's Theory of Behavioural Change]. *Advances in Behaviour Research and Therapy*, *1*(4), 139-161. https://doi.org/https://doi.org/10.1016/0146-6402(78)90002-4

Beck, A. T., & Clark, D. A. (1997). An information processing model of anxiety: automatic and strategic processes. *Behaviour research and therapy*, *35*(1), 49-58. https://doi.org/10.1016/s0005-7967(96)00069-1

Brewer, J. (2021). *Unwinding Anxiety: New Science Shows How to Break the Cycles of Worry and Fear to Heal Your Mind*. Avery, an imprint of Penguin Random House.

Brown, V. M., Price, R., & Dombrovski, A. Y. (2023). Anxiety as a disorder of uncertainty: implications for understanding maladaptive anxiety, anxious avoidance, and exposure therapy. *Cognitive, Affective, Behavioral Neuroscience*, *23*(3), 844-868. https://doi.org/10.3758/s13415-023-01080-w

Clark, D. A., & Beck, A. T. (2012). *Terapia Cognitiva para Trastornos de Ansiedad: Ciencia y Práctica* (J. Aldekoa, Trad.) [Título de la edición original: *Cognitive Therapy of Anxiety Disorders: Science and Practice*, © 2010, The Guilford Press, New York, USA]. Desclée de Brouwer.

Daviu, N., Bruchas, M. R., Moghaddam, B., Sandi, C., & Beyeler, A. (2019). Neurobiological links between stress and anxiety. *Neurobiology of Stress*, *11*, 100191. https://doi.org/10.1016/j.ynstr.2019.100191

Dayan, P. (1993). Improving Generalization for Temporal Difference Learning: The Successor Representation. *Neural Computation*, *5*(4), 613-624. https://doi.org/10.1162/neco.1993.5.4.613

Dayan, P., & Abbott, L. (2001). *Theoretical Neuroscience*. Massachusetts Institute of Technology.

De Houwer, J., Barnes-Holmes, D., & Moors, A. (2013). What is learning? On the nature and merits of a functional definition of learning. *Psychonomic Bulletin Review*, *20*(4), 631-642. https://doi.org/10.3758/s13423-013-0386-3

Diederich, A. (2023). Computational modeling. En H. Cooper, M. N. Coutanche, L. M. McMullen, A. T. Panter, D. Rindskopf & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Research designs: Quantitative, qualitative, neuropsychological, and biological* (2nd ed., pp. 515-536). American Psychological Association. https://doi.org/10.1037/0000319-023

Doya, K. (2023). Reinforcement Learning. En R. Sun (Ed.), *The Cambridge Handbook of Computational Cognitive Sciences* (2nd, pp. 350-380). Cambridge University Press. https://doi.org/10.1017/9781108755610

Ducarouge, A., & Sigaud, O. (2017). The Successor Representation as a model of behavioural flexibility. *Journées Francophones sur la Planification, la Décision et l'Apprentissage pour la conduite de systèmes (JFPDA 2017)*. https://hal.science/hal-01576352

Eckstein, M. K., Master, S. L., Xia, L., Dahl, R. E., Wilbrecht, L., & Collins, A. G. (2022). The interpretation of computational model parameters depends on the context (C. Hartley, T. E. Behrens & A. Radulescu, Eds.). *eLife*, *11*, e75474. https://doi.org/10.7554/eLife.75474

Farrell, S., & Lewandowsky, S. (2015). An Introduction to Cognitive Modeling. En B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 3-24). Springer Science + Business Media. https://link.springer.com/book/10.1007/978-1-4939-2236-9

Gagne, C., & Dayan, P. (2022). Peril, prudence and planning as risk, avoidance and worry. *Journal of Mathematical Psychology*, *106*, 102617. https://doi.org/https://doi.org/10.1016/j.jmp.2021.102617

Gaskett, C. (2003). Reinforcement learning under circumstances beyond its control. *Proceedings of the International Conference on Computational Intelligence, Robotics and Autonomous Systems*. https://researchonline.jcu.edu.au/632/1/cimca2003.pdf

Gershman, S. J. (2018). The Successor Representation: Its Computational Logic and Neural Substrates. *Journal of Neuroscience*, *38*(33), 7193-7200. https://doi.org/10.1523/JNEUROSCI.0151-18.2018

Goldway, N., Eldar, E., Shoval, G., & Hartley, C. A. (2023). Computational Mechanisms of Addiction and Anxiety: A Developmental Perspective. *Biological Psychiatry*, *93*(8), 739-750. https://doi.org/10.1016/j.biopsych.2023.02.004

Grant, D. M., & White, E. J. (2016, diciembre). Influence of Anxiety on Cognitive Control Processes. https://doi.org/10.1093/acrefore/9780190236557.013.74

Gústavsson, S. M., Salkovskis, P. M., & Sigurðsson, J. F. (2022). Revised Beckian cognitive therapy for generalised anxiety disorder. *The Cognitive Behaviour Therapist*, *15*, e58. https://doi.org/10.1017/S1754470X22000563

Hunt, C. A., Ropella, G. E., Park, S., & Engelberg, J. (2008). Dichotomies between computational and mathematical models. *Nature Biotechnology*, *26*(7), 737-739. https://doi.org/10.1038/nbt0708-737

Huys, Q. (2013). Computational Psychiatry. En D. Jaeger & R. Jung (Eds.), *Encyclopedia of Computational Neuroscience* (pp. 1-10). Springer New York. https://doi.org/10.1007/978-1-4614-7320-6_501-1

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience*, *19*(3), 404-413. https://doi.org/10.1038/nn.4238

INEGI. (2021). Presenta INEGI Resultados de la Primera Encuesta Nacional de Bienestar Autorreportado (ENBIARE) 2021. https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2021/EstSociodemo/ENBIARE_2021.pdf

Jang, B., Kim, M., Harerimana, G., & Kim, J. W. (2019). Q-Learning Algorithms: A Comprehensive Classification and Applications. *IEEE Access*, *7*, 133653-133667. https://doi.org/10.1109/ACCESS.2019.2941229

Juliani, A., Barnett, S., Davis, B., Sereno, M., & Momennejad, I. (2022). Neuro-Nav: A Library for Neurally-Plausible Reinforcement Learning. *The 5th Multidisciplinary Conference on Reinforcement Learning and Decision Making*.

Katz, B. A., Matanky, K., Aviram, G., & Yovel, I. (2020a). Reinforcement sensitivity, depression and anxiety: A meta-analysis and meta-analytic structural equation model. *Clinical Psychology Review*, *77*, 101842. https://doi.org/https://doi.org/10.1016/j.cpr.2020.101842

Katz, B. A., Matanky, K., Aviram, G., & Yovel, I. (2020b). Reinforcement sensitivity, depression and anxiety: A meta-analysis and meta-analytic structural equation model [Epub 2020 Mar 9]. *Clinical Psychology Review*, *77*, 101842. https://doi.org/10.1016/j.cpr.2020.101842

Khan, S., & Khan, R. A. (2017). Chronic Stress Leads to Anxiety and Depression. *Annals of Psychiatry and Mental Health*, *5*(1), 1091. https://www.jscimedcentral.com/public/assets/articles/psychiatry-5-1091.pdf

Levenstein, D., Alvarez, V. A., Amarasingham, A., Azab, H., Chen, Z. S., Gerkin, R. C., Hasenstaub, A., Iyer, R., Jolivet, R. B., Marzen, S., Monaco, J. D., Prinz, A. A., Quraishi, S., Santamaria, F., Shivkumar, S., Singh, M. F., Traub, R., Nadim, F., Rotstein, H. G., & Redish, A. D. (2023). On the Role of Theory and Modeling in Neuroscience. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *43*(7), 1074-1088. https://doi.org/10.1523/JNEUROSCI.1179-22.2022

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *eLife*, *7*. https://doi.org/10.7554/eLife.32548

Momennejad, I., Russek, E. M., Cheong, J. K., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680-692. https://doi.org/10.1038/s41562-017-0180-8

National Institute of Biomedical Imaging and Bioengineering. (2020, mayo). Computational Modeling. https://www.nibib.nih.gov/science-education/science-topics/computational-modeling

Ng, A., Harada, D., & Russell, S. J. (1999). Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:5730166

Pike, A. C., & Robinson, O. J. (2022). Reinforcement Learning in Patients With Mood and Anxiety Disorders vs Control Individuals: A Systematic Review and Meta-analysis. *JAMA psychiatry*, *79*(4), 313-322. https://doi.org/10.1001/jamapsychiatry.2022.0051

Raymond, J. G., Steele, J. D., & Seriès, P. (2017). Modeling Trait Anxiety: From Computational Processes to Personality. *Frontiers in Psychiatry*, *8*. https://doi.org/10.3389/fpsyt.2017.00001

Robinson, O. J., Vytal, K., Cornwell, B. R., & Grillon, C. (2013). The impact of anxiety upon cognition: perspectives from human threat of shock studies [eCo-

llection 2013]. *Frontiers in Human Neuroscience*, *7*, 203. https://doi.org/10.3389/fnhum.2013.00203

Rogers, A. H., Wieman, S. T., & Baker, A. W. (2020). Anxiety comorbidities: Mood disorders, substance use disorders, and chronic medical illness. En E. Bui, M. E. Charney & A. W. Baker (Eds.), *Clinical handbook of anxiety disorders: From theory to practice* (pp. 77-103). Humana Press/Springer Nature. https://doi.org/10.1007/978-3-030-30687-8_5

Russell, S., & Norvig, P. (2022). Making Complex Decisions. En *Artificial Intelligence: A Modern Approach* (2.ª ed., pp. 552-586). Pearson.

Secretaría de Salud. (2022). Segundo Diagnóstico Operativo de Salud Mental y Adicciones. https://www.gob.mx/cms/uploads/attachment/file/730678/SAP-DxSMA-Informe-2022-rev07jun2022.pdf

Sharp, P. B., & Eldar, E. (2019). Computational models of anxiety: Nascent efforts and future directions. *Current Directions in Psychological Science*, *28*(2), 170-176. https://doi.org/10.1177/0963721418818441

Simmons, J. M., Cuthbert, B., Gordon, J. A., & Ferrante, M. (2020). Introduction: Toward a Computational Approach to Psychiatry. En P. Seriès (Ed.), *Computational Psychiatry* (pp. 1-20). The MIT Press.

Singh, V., Chen, S.-S., Singhania, M., Nanavati, B., kumar kar, A., & Gupta, A. (2022). How are reinforcement learning and deep learning algorithms used for big data based decision making in financial industries–A review and research agenda. *International Journal of Information Management Data Insights*, *2*(2), 100094. https://doi.org/https://doi.org/10.1016/j.jjimei.2022.100094

Steimer, T. (2002). The biology of fear- and anxiety-related behaviors. *Dialogues in clinical neuroscience*, *4*(3), 231-249. https://doi.org/10.31887/DCNS.2002.4.3/tsteimer

Stein, D. J., Scott, K. M., de Jonge, P., & Kessler, R. C. (2017). Epidemiology of anxiety disorders: from surveys to nosology and back. *Dialogues in clinical*

neuroscience, *19*(2), 127-136. https://doi.org/10.31887/DCNS.2017.19.2/dstein

Subramanian, A., Chitlangia, S., & Baths, V. (2022). Reinforcement learning and its connections with neuroscience and psychology. *Neural Networks*, *145*, 271-287. https://doi.org/https://doi.org/10.1016/j.neunet.2021.10.003

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*(1), 9-44. https://doi.org/10.1007/BF00115009

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2.ª ed.). The MIT Press.

Swazinna, P., Udluft, S., Hein, D., & Runkler, T. (2022). Comparing Model-free and Model-based Algorithms for Offline Reinforcement Learning [6th IFAC Conference on Intelligent Control and Automation Sciences ICONS 2022]. *IFAC-PapersOnLine*, *55*(15), 19-26. https://doi.org/https://doi.org/10.1016/j.ifacol.2022.07.602

Wilmer, M., Anderson, K., & Reynolds, M. (2021). Correlates of Quality of Life in Anxiety Disorders: Review of Recent Research. *Current Psychiatry Reports*, *23*(11), 77. https://doi.org/10.1007/s11920-021-01290-4

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data (T. E. Behrens, Ed.). *eLife*, *8*, e49547. https://doi.org/10.7554/eLife.49547

Wu, H. (2023). Reinforcement Learning Inspired by Psychology and Neuroscience. *Journal of Education, Humanities and Social Sciences*, *8*, 2164-2170. https://doi.org/10.54097/ehss.v8i.4673

Yamamori, Y., & Robinson, O. J. (2023). Computational perspectives on human fear and anxiety. *Neuroscience and Biobehavioral Reviews*, *144*, 104959. https://doi.org/10.1016/j.neubiorev.2022.104959

Zorowitz, S., Momennejad, I., & Daw, N. D. (2020). Anxiety, avoidance, and sequential evaluation [Epub 2020 Mar 1]. *Computational Psychiatry (Cambridge, Mass.)*, *4*. https://doi.org/10.1162/cpsy_a_00026