Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Rol del aprendizaje operante en la cooperación entre animales evaluado con el dilema del prisionero iterado: una teoría computacional.

Tesis presentada para obtener el título de Licenciado en Ciencias de la Computación

Autor: Daniel Alejandro Jercog

BUENOS AIRES, ARGENTINA
AGOSTO, 2007

UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE CIENCIAS EXACTAS Y NATURALES
DEPARTAMENTO DE COMPUTACIÓN

Fecha: _____

Alumno: Daniel Alejandro Jercog (L.U. 576/00)

Director de Tesis: _____
Dr. Ing. B. Silvano Zanutto

Co Director: _____
Ing. Sergio Lew

Jurado: _____
Dr. Juan Santos

Jurado: _____
Dr. Rubén Muzio

Tabla de Contenidos

Resumen

En diversos ambientes biológicos, los sistemas adaptativos han evolucionado con la característica de garantizar su propia estabilidad replicativa. Bajo esta condición, una de las más importantes paradojas de la teoría de la evolución resulta ser la cooperación entre individuos.

Por otra parte, ante la alternativa de elegir un refuerzo apetitivo pequeño e inmediato o uno mayor pero retardado en el tiempo, los animales optan en general por la primera opción. Este fenómeno, denominado "impulsividad", ha sido observado en distintas especies, algunas de ellas filogenéticamente muy distantes.

Este efecto trae importantes consecuencias en el paradigma de cooperación del dilema del prisionero iterado (DPI). Si bien, a nivel teórico la mejor opción bajo condiciones de reciprocidad resulta ser la opción de cooperar (puesto que a largo plazo es la que otorga mayor beneficio mutuo), experimentalmente los animales muestran una fuerte tendencia a la deserción. La devaluación del refuerzo retardado con respecto al inmediato parece jugar en este fenómeno un papel preponderante.

A su vez existe evidencia experimental en el estudio del comportamiento de urrracas azules (Cyanocitta Cristata) en el marco del DPI, indicando que la acumulación de alimento durante un número sucesivo de jugadas es un mecanismo válido para lograr niveles sostenidos de cooperación.

En este trabajo se propone la construcción de una teoría computacional utilizando redes neuronales para la formalización de hipótesis neurofisiológicas y conductuales. Utilizando las reglas de aprendizaje Hebbiano y de Rescorla-Wagner para la computación de pesos sinápticos, el modelo propuesto es capaz de predecir los resultados experimentales acerca de los efectos en la devaluación de refuerzos y la acumulación de alimento, permitiendo explicar el fenómeno del altruismo recíproco en un ambiente de condicionamiento operante.

Capítulo 1

Introducción

Las ciencias biológicas han descrito sistemas adaptativos altamente sociables. Aunque a veces compiten entre ellos, muchas veces comparten esfuerzos en tareas cooperativas. Esta estructura social es quizás el más importante y difícil problema para poder entender la historia natural de los sistemas adaptativos. Tal como apunta Dawkins, en su libro "El gen egoísta" (1976), en un mundo de competición los sistemas adaptativos han evolucionado con la característica única de garantizar su propia estabilidad replicativa. Desde esta visión, una aparente paradoja de la teoría evolutiva es la cooperación entre individuos.

Darwin, en su libro "El origen de las especies" (1859), notó que la cooperación observada en insectos no parecía poder explicarse mediante la teoría Evolutiva. De cualquier manera, esbozó una posible solución al proponer que la selección natural podría actuar no sólo a nivel de individuo sino también a nivel de familia. Existen teorías que resuelven la paradoja evolutiva, explicando diversos mecanismos mediante los cuales es posible que los comportamientos de cooperación hayan podido evolucionar. Uno de los mecanismos mediante el cual es posible observar cooperación hasta incluso en individuos de diferentes especies es la reciprocidad directa, propuesto por Trivers (1971) [39], en el cual los comportamientos altruistas pueden generar futuras reciprocidades. El dilema del prisionero es un problema de teoría de juegos que permite estudiar de manera concreta la reciprocidad directa entre individuos. Sin embargo, el paso interesante lo dan Axelrod y Hamilton al estudiar cómo es la evolución de la cooperación en el marco de sucesivos encuentros del dilema del prisionero [2]. Este nuevo juego estratégico es denominado el dilema del prisionero iterado y ha sido abordado por diversas disciplinas como la

psicología, la economía y las neurociencias, con el fin de poder entender conductas de cooperación.

Si bien, analizando a nivel teórico al dilema del prisionero iterado, la estrategia de cooperación mutua permite tener un mayor beneficio a largo plazo en condiciones de reciprocidad, resulta difícil sostener niveles de cooperación significativos de manera experimental. Esto, entre otras razones, se debe a que los animales realizan una devaluación de los beneficios a largo plazo de cooperar. Stephens et al [36] proponen, mediante un estudio en urracas azules (Cyanocitta cristata), que juegan al dilema del prisionero iterado, que el hecho de hacer visibles los beneficios a largo plazo hace posible que se logren niveles significativos y estables de cooperación. Asimismo, ponen en evidencia que el problema de auto-control, que consiste en elegir entre dos alternativas que producen un beneficio: uno menor pero inmediato y otro mayor pero retrasado (similar en algún aspecto al problema de cooperar o no cooperar en el dilema del prisionero), puede verse influenciado por este mismo hecho.

La adaptación, además de consistir en tratar con los cambios del entorno, se refiere a la posibilidad de realizar acciones que introduzcan cambios que modifiquen al comportamiento posterior de un individuo. En este trabajo interesa estudiar cuáles son los mecanismos conductuales básicos que explican al altruismo reciproco.

Tal como lo define Skinner, en el condicionamiento operante o instrumental los animales cambian su comportamiento en función de lo experimentado. Se lo denomina operante, dado que el comportamiento es capaz de generar modificaciones en el entorno. A su vez, se lo denomina instrumental porque el comportamiento es una herramienta capaz de introducir cambios para obtener recompensas.

Aunque existen modelos poblacionales de cooperación para el dilema del prisionero para analizar comportamientos económicos, la única evidencia de modelos que abordan al problema desde un punto de vista de condicionamiento operante es el trabajo de Gutnisky y Zanutto (2002) [14]. Ellos muestran que un modelo de aprendizaje operante basado en evidencias neurofisiológicas y conductuales se desempeña de manera robusta y satisfactoria al ser expuesto a la situación del dilema del prisionero, aún jugando frente a estrategias de lo más variadas. Sin embargo, este modelo no estudia diversos aspectos como, por ejemplo, la influencia de retrasos en los refuerzos y los efectos de acumulación de los mismos.

En este trabajo se desea estudiar cuál es el rol del condicionamiento operante en el problema de cooperación del dilema del prisionero iterado bajo los efectos del retraso y la acumulación utilizando la evidencia experimental provista por Stephens [36]. Por

este motivo, se propone la construcción de un modelo computacional basado en trabajos previos sobre modelos de condicionamiento operante [14] [20], construyendo un modelo capaz de representar dichos efectos anteriormente enunciados. De esa forma se podría analizar si el condicionamiento operante, bajo estas restricciones propuestas, es capaz de explicar dichos resultados obtenidos sobre cooperación entre animales, dando así más información acerca del posible rol del condicionamiento operante en la evolución de la cooperación.

Capítulo 2

Comportamiento adaptativo

En el presente capítulo se introducen los conceptos básicos de los paradigmas de aprendizaje abordados por la psicología experimental: el condicionamiento clásico y el condicionamiento operante. De estos principios nace la motivación de construir un modelo neuronal como teoría formal que permita explicar resultados experimentales del aprendizaje en animales, y posteriormente si por medio del mismo es posible plantear al condicionamiento operante como un posible mecanismo para lograr cooperación en animales.

2.1. Biología y comportamiento

Existe una división dentro de los comportamientos entre aquellos considerados innatos, que son independientes de la experiencia previa, y aquellos aprendidos que son enteramente dependientes de ella. Algunos critican esta dicotomía de resultar poco clara, diciendo que en realidad no hay aprendizaje completamente innato en el sentido de la total independencia de la experiencia previa. Aunque la diferencia resulte poco clara, el mecanismo que permite el aprendizaje es innato, aún cuando aquello que es aprendido no lo sea [35].

Staddon define al nicho de un organismo como al rol de dicho organismo en conjunto con su entorno. Lo que un animal aprende y la manera en que lo hace está intrínsecamente relacionado con su nicho. Dado que los nichos difieren en muchos aspectos, también

difieren en los mecanismos de aprendizaje. Sin embargo, hay aspectos que se aplican a todos los nichos, tales como el espacio y el tiempo. A su vez, existen reglas generales que se aplican a todos los nichos. Por ejemplo, la información vieja suele ser menos útil que la nueva, y por consiguiente los animales olvidan, olvidando en menor medida aquello que aprendieron recientemente. Cuando un nicho se vuelve más complejo, también se vuelve más complejo el procesamiento de su historia para su posterior desenvolvimiento. Para poder procesar su pasado, primero el animal debe tener memoria para poder tomar decisiones en el futuro en base al mismo [35]. Los aspectos que se estudian en este trabajo, tienen que ver con estos mismos que son independientes de un nicho en particular.

2.2. Reflejo vs. Condicionamiento

Las acciones reflejas son las respuestas que se dan en presencia de ciertos estímulos particulares y corresponden a comportamientos innatos. A diferencia del condicionamiento, las respuestas que se efectúan no requieren aprendizaje sino que existe una conexión innata entre el estímulo y la respuesta [24]. El médico inglés Charles Sherrington se dedicó a estudiar las características de los reflejos. Sherrington condujo la mayoría de sus estudios en animales espinales, que son animales a los cuales se les han seccionado la médula espinal a nivel cervical y por lo tanto el cerebro no puede recibir información sensora del cuerpo ni controlar los músculos. Cualquier acción refleja está controlada por neuronas en la médula espinal y por el cuerpo mismo. Los principios que Sherrington postuló sobre la acción refleja son válidos aún para animales sin lesiones espinales.

2.3. Condicionamiento Clásico y Operante

Por otra parte, a principios del siglo XX surge la psicología conductista como la corriente de la psicología que defiende el empleo de procedimientos estrictamente experimentales para estudiar el comportamiento observable, considerando al entorno como un conjunto de estímulos y respuestas. Las mayores influencias de la misma fueron los trabajos de condicionamiento clásico de Ivan Pavlov, los trabajos de Thorndike, Watson quienes dirigieron a la psicología hacia la restricción sobre métodos experimentales, y el trabajo de Skinner sobre condicionamiento operante.

El condicionamiento es un tipo de aprendizaje en el cual un organismo realiza una asociación de eventos. Un estímulo incondicionado (US) es un estímulo que genera una

respuesta incondicionada (UR). La UR es una respuesta reflexiva y es generada sin condicionamientos por el US, sin importar la historia del aprendizaje de un organismo. Por ejemplo, como a los efectos de los estudios de Pavlov que veremos a continuación, el alimento (que es un US) situado en la boca causa la salivación del individuo (la salivación es la UR). El refuerzo se define como una operación que incrementa la probabilidad de cierta clase de comportamientos. Es un término amplio y utilizado por diversas disciplinas. En particular, el refuerzo definido por Pavlov es definido como la operación que incrementa la probabilidad de ocurrencia de la CR cuando el CS es presentado. Según la terminología de Pavlov, la presentación de un US seguido del CS constituye un refuerzo del reflejo condicionado de la salivación (la CR). Estos elementos son los conceptos básicos para la teoría de condicionamiento.

Tanto el condicionamiento clásico como el condicionamiento operante, que son analizados a continuación, desde el enfoque psicológico se han convertido en los principales paradigmas de aprendizaje que permiten a los animales obtener características relevantes de su entorno de manera que los mismos puedan lograr la capacidad de ejecutar respuestas para obtener beneficios y evitar castigos. Los experimentos de condicionamiento clásico podrían ser definidos como aquellos en los cuales se genera una contingencia entre un estímulo y un resultado, mientras que los experimentos de condicionamiento operante podrían ser definidos como aquellos en los cuales se genera una contingencia entre una respuesta y un resultado.

2.3.1. Condicionamiento clásico

El precursor del condicionamiento clásico fue el fisiólogo ruso Ivan Pavlov, considerado uno de los padres de la psicología. Su principal interés era el estudio de los factores glandulares y nerviosos en el proceso digestivo. Como fisiólogo se interesó en el comportamiento como una herramienta para poder explicar el funcionamiento del cerebro.

El alimento, al ser colocado en la boca, produce saliva. Este fenómeno fisiológico permite que el alimento sea alterado químicamente para que, tras ser diluido, pueda producirse el proceso digestivo. Lo que Pavlov observa es que dicha secreción puede ser evocada cuando el individuo detecta, de alguna forma, la posible presencia de alimento.

Utilizando como sujeto de estudio a un perro en un aparato que lo inmoviliza y permite tener registro de la salivación del mismo, sus experimentos básicamente consisten en presentarle al animal un estímulo durante un breve período de tiempo (ejemplo, el sonido de una campana) antes de entregarle el alimento, el cual lo induce a salivar. Esta secuencia se repite un número de veces hasta que el perro saliva frente a la presencia

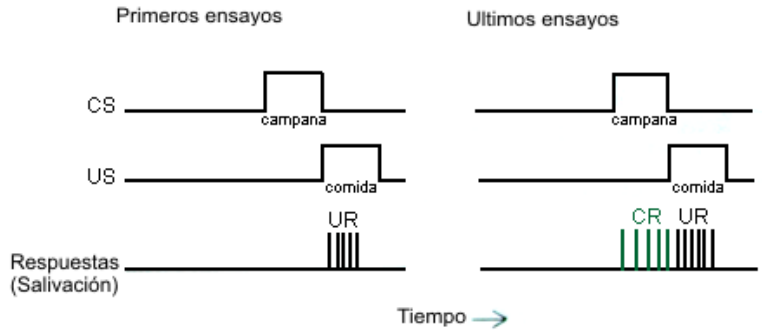

Figura 2.1: Eventos de un ensayo de condicionamiento clásico antes de establecerse la respuesta condicionada (izquierda) y luego de establecerse la misma (derecha) [Fuente: Mazur 1994]

del estímulo, previamente a que el alimento sea entregado. Este tipo de aprendizaje, recibió el nombre de condicionamiento clásico, donde los conceptos involucrados son el estímulo inicial (estímulo condicionado, CS), el alimento (estímulo incondicionado, US), la respuesta refleja de salivar (respuesta incondicionada, UR), y luego de aprendida la secuencia, la respuesta que anticipa al alimento (respuesta condicionada, CR). En la figura 2.1 puede verse un diagrama de eventos producidos durante un ensayo del experimento anterior.

Luego de sucesivas apariciones de secuencias CS-US, el sujeto comienza a exhibir gradualmente la CR y un incremento sostenido en su fuerza durante ensayos sucesivos hasta alcanzar un valor asintótico. Este proceso se denomina adquisición, y en el mismo el animal aprende que la presencia del CS es seguida por el US. Una vez aprendida la secuencia CS-US donde el perro saliva frente a la presencia del estímulo, si se deja de entregar alimento durante sucesivos ensayos el perro deja de salivar. Este fenómeno recibe el nombre de extinción, donde el animal aprende que el CS no estará seguido por el US. La adquisición y la extinción, constituyen las fases fundamentales de todos los experimentos de condicionamiento tanto clásico como operante.

Otros conceptos importantes, también descubiertos por Pavlov, son el de discriminación condicionante y generalización. La generalización se observa cuando un animal es entrenado para responder a un cierto estímulo, y luego se enseña un nuevo estímulo de características similares al primero, el animal transferirá parte del aprendizaje obtenido del primer estímulo al nuevo. Mientras más similares sean los estímulos, más parecida será la respuesta del animal. La discriminación consiste en presentar estímulos con ciertas características distintas (por ejemplo luz azul y luz verde), una de las cuales será

reforzada (recibirá un US) y otra no. El animal termina ejecutando una CR sólo frente al estímulo por el cual recibe una recompensa, es decir, termina aprendiendo a encontrar la diferencia entre los estímulos condicionados, que están relacionados con la recepción o no del estímulo incondicionado. Existen además otros fenómenos que se presentan en experimentos de condicionamiento, como el fenómeno de recuperación espontánea, la desinhibición, la readquisición rápida, entre otros [24].

2.3.2. Condicionamiento Operante

Si al perro del experimento de Pavlov lo entrenamos en la asociación CS-US, y luego solo le presentamos el CS sin entregarle el US, el perro intentará buscar la comida. De hecho, si colocamos la comida en un lugar de la habitación y le permitimos al perro moverse, buscará de manera azarosa donde está su alimento y lo comerá. Si repetimos el experimento otro día, veremos que el animal ya no buscará por cualquier lado, sino que irá donde antes había encontrado la comida. Los dos procesos involucrados son el de variación, donde existe un comportamiento no sistemático que eventualmente logra conseguir su objetivo, y el de selección, donde se vuelve a efectuar aquel comportamiento que le ha sido efectivo para obtener su refuerzo.

Muchas de las situaciones a las que se puede enfrentar un animal pueden evaluarse en función de la certeza que tiene con respecto a la eficiencia de su accionar. Por un lado puede enfrentarse a situaciones donde el resultado es altamente predecible, como tocar una superficie caliente y sacar la mano de la misma en una acción reflejo. Por otro lado existen circunstancias donde la acción adecuada puede ser bastante incierta (como por ejemplo, la búsqueda de alimento). Lo que la evolución ha producido, es darle la posibilidad al animal de realizar una variedad de acciones y proveerlo de capacidad para aprender basándose en la experiencia obtenida. Puede así intentar distintas estrategias para obtener alimento, evaluando su efectividad y repitiendo más veces aquellos métodos que le produzcan un mayor beneficio.

Básicamente, hay dos tipos diferentes de US, los placenteros y los no placenteros. Thorndike define a los estímulos placenteros como aquellos que el organismo busca obtener y preservar, y los no placenteros como aquellos que el organismo busca evitar y desentenderse. Si se presenta un estímulo placentero cuando el animal realiza el comportamiento deseado, el aprendizaje se denomina apetitivo. Por otro lado, si un estímulo no placentero es presentado y luego removido cuando el animal realiza el comportamiento deseado, el aprendizaje se denomina aversivo. Con respecto al análisis experimental de estas conductas en un laboratorio, un estímulo apetitivo usual en estos experimentos suele ser el alimento, mientras que un estímulo aversivo habitual es un shock eléctrico.

A continuación se detallan ambos tipos de aprendizaje.

Aprendizaje aversivo

En el caso del aprendizaje aversivo se distinguen dos respuestas principales, el escape y la evitación. En el escape, el animal recibe un estímulo y a continuación un shock eléctrico, ejecuta la respuesta de escape una vez aplicado el shock, y de esta manera cesa el shock. Mientras que en la evitación, el animal logra evitar el shock, por responder anticipadamente a la aplicación del mismo.

Solomon y Wynne realizaron un experimento mostrando varias de las propiedades del comportamiento aversivo. El ambiente del experimento consiste de una cámara con dos compartimientos rectangulares, separados por un tabique de varios centímetros de altura. El piso es metálico, y puede ser electrificado. El animal utilizado por Solomon y Wynne fue un perro, el cual puede saltar de un compartimiento al otro. Como estímulo condicionado hay dos luces que pueden iluminar separadamente cada uno de los compartimentos. En cada sesión el perro es entrenado en diez ensayos, en los cuales puede escapar o evitar el shock saltando la barrera formada por el tabique. El intento comienza apagando la luz del compartimiento donde se encuentra el perro, y dejando encendida la del otro compartimiento. Si el perro se queda en la parte oscura, luego de 10 segundos, recibe el shock hasta que no saltase al otro lado, escapándose del mismo. Pero también puede evitar directamente el refuerzo negativo, si salta antes de los 10 segundos. Se midió la latencia (el tiempo entre la aparición del estímulo y la acción efectuada) en función del número de intento en los primeros. La latencia era habitualmente mayor a los 10 segundos, siendo entonces una respuesta de escape, pero aproximadamente para el quinto intento, la latencia era menor a 10 segundos, realizando una respuesta de evitación. Luego de varias decenas de intentos, la latencia se veía reducida a entre 2 y 3 segundos, viendo también que muchos perros no experimentaban ningún shock luego de su primer respuesta de evitación. Esto generó la llamada paradoja de evitación, que cuestiona cómo es posible que la no ocurrencia de un evento (el shock) pueda servir como refuerzo para la respuesta de evitación. Esta cuestión no se plantea en el escape, dado que existe un cambio en el refuerzo por efectuar la respuesta del mismo.

Algunos investigadores sostenían que no tenía sentido decir que un shock que no es experimentado sirva como refuerzo para un comportamiento. Este hecho motivó el desarrollo de una teoría de evitación llamada Teoría de dos factores o de dos procesos. Estos efectos se han estudiado mediante modelos computacionales [31] [20] [15] que reproducen datos experimentales mediante estímulos aversivos.

Aprendizaje apetitivo

El psicólogo conductista Edward Thorndike fue el primer investigador en analizar sistemáticamente cómo los comportamientos no reflexivos (aquellos que no son innatos) pueden ser modificados como resultado de la experiencia. Los experimentos consisten en colocar un animal hambriento en un compartimiento. Si el animal realizaba la respuesta correcta, la puerta puede ser abierta y de esta manera el animal puede salir y obtener comida (el refuerzo apetitivo). Al comienzo de los experimentos, los animales usualmente exploran el compartimiento de una manera aparentemente azarosa y eventualmente efectúa la respuesta adecuada para salir. Midiendo el tiempo empleado desde que se colocaba al animal en el compartimiento hasta que salía al exterior, se observó que dicha latencia iba disminuyendo durante sucesivos ensayos. Se atribuyó dicha mejora en la performance a un fortalecimiento progresivo de la conexión del par estímulo-respuesta afirmando que a mayor satisfacción o disconformidad, mayor es el fortalecimiento o debilitamiento de dicho vínculo, respectivamente.

Thorndike entendió que un refuerzo positivo es aquel donde el animal no evita al US, y habitualmente intenta alcanzar y preservar. Mientras que un refuerzo negativo es aquel que el animal intenta normalmente de evitar y abandonar.

Estableció así la Ley del Efecto. Según esta ley, las respuestas que sean seguidas de consecuencias apetitivas, serán asociadas al estímulo y tendrán mayor probabilidad de ocurrencia cuando el estímulo vuelva a aparecer. Por el contrario, si la respuesta al estímulo va seguida de una consecuencia aversiva, la asociación será más débil, con lo que la probabilidad de ocurrencia será menor.

2.3.3. Protocolos de condicionamiento

Durante un ensayo de condicionamiento, tanto clásico como operante, la relación temporal entre el CS y el US puede variar notablemente dando lugar a diversas relaciones temporales entre dichos estímulos. El tiempo entre la aparición del CS y la aparición del US se denomina intervalo intra-estímulo (ISI). En la figura 2.2 se presentan algunas variaciones comunes de relaciones temporales.

Tal como se indica en la figura 2.2, en el condicionamiento simultáneo el CS y el US comienzan al mismo tiempo (ISI=0). En el caso del condicionamiento demorado (delay conditioning), la aparición del CS precede a la aparición del US. Se pueden presentar la posibilidad de que la duración del CS sea igual al ISI o que sea igual a la suma del ISI y la duración del US. En cambio, en el condicionamiento de traza (trace conditioning),

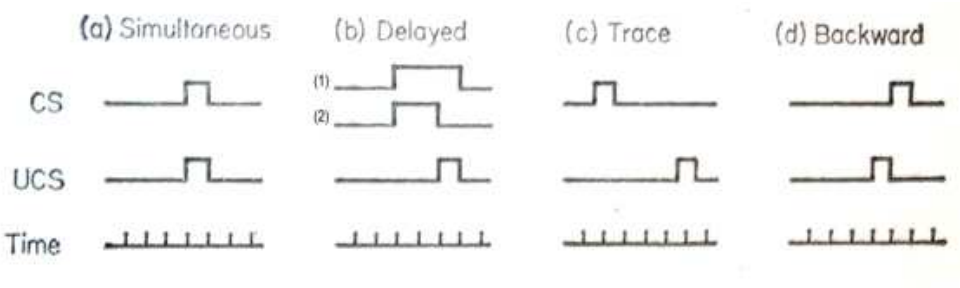

Figura 2.2: a) Condicionamiento simultáneo, b) Condicionamiento demorado (1) o (2), c) Condicionamiento de traza, d) Condicionamiento atrasado [Fuente: Mackintosh 1974]

la desaparición del CS precede a la aparición del US. En el caso del condicionamiento atrasado (backward conditioning) la desaparición del US es quien precede a la aparición del CS. En este último caso no se produce condicionamiento.

La magnitud del ISI en el desarrollo de un ensayo tiene una implicancia determinante en la capacidad del condicionamiento. Se dice que un ISI es óptimo si maximiza la capacidad de condicionamiento. Sin embargo, existen grandes discrepancias entre un mismo esquema para distintos tipos de refuerzo (una CR de parpadeo puede condicionarse con a lo sumo pocos segundos, mientras que una CR de aprendizaje por aversión gustativa podría llevar horas). Más allá de la naturaleza del refuerzo, la razón de dichas discrepancias y los valores óptimos de ISI sigue siendo hoy en día poco clara. [22]

2.3.4. La reglas de aprendizaje de Rescorla-Wagner

El principio de aprendizaje por asociación establece que mientras mayor sea la frecuencia del apareamiento de dos estímulos, mayor será la asociación entre los mismos. Este antiguo principio ha sido uno de los pilares de las teorías formales de aprendizaje. Sin embargo, en el año 1969, Kamin realizó un experimento donde una asociación inicial de un CS con un US bloqueaba a toda asociación posterior con otro CS [1]. Este efecto se conoce como blocking [22].

El fenómeno del blocking implicó la necesidad de generar nuevas teorías de aprendizaje [24]. En el año 1972, Rescorla y Wagner presentaron un modelo de aprendizaje

[1]En la primera fase se asocia al CS_1 con un US. Luego se asocia CS_1 y CS_2 con el mismo US. Finalmente para la contingencia de CS_2 con el US no puede establecerse una asociación.

capaz de explicar un gran número de resultados de la psicología experimental y al mismo tiempo permitió explicar el efecto de blocking. Si bien la regla de aprendizaje de Rescorla-Wagner (RW) carece de un sustento neurofisiológico, motivo por el cual es fuertemente cuestionada, se adapta de manera satisfactoria a experimentos fisiológicos realizados en neuronas dopaminérgicas que ponen en compromiso a las bases de las teorías formales del aprendizaje [40].

Concretamente, la regla RW asume que para un animal resulta útil poder predecir eventos futuros que sean importantes, tanto aquellos favorables como desfavorables. El modelo establece que por cada aparición de un CS particular, puede ocurrir alguna de las siguientes tres situaciones:

- Si el US recibido es mayor que el esperado, el CS se vuelve más excitatorio.
- Si el US recibido es menor que el esperado, el CS se vuelva más inhibitorio.
- Si el US recibido es igual al que se esperaba, no se producen cambios relativos al CS.

Una vez que se ha completado la etapa de aprendizaje, en la cual el condicionamiento alcanza un nivel asintótico, la fuerza de asociación estará directamente relacionada con la intensidad del US. Llamemos A_j al valor asintótico de la fuerza asociativa para un US_j y V_i a la fuerza asociativa entre el estímulo i y el US_j en un momento determinado. Mientras A_j representa la magnitud del evento observado, $V = \sum_{\forall i} V_i$ representa lo que animal espera recibir de éste. Si $A_j > V$ entonces cada CS presente en el ensayo tendrá un incremento en su fuerza asociativa, si $A_j < V$ tendrá un decremento y si $A_j = V$ no habrá ninguna modificación. El modelo establece que el aprendizaje se da en proporción a la diferencia $A_j - V$, siendo el factor de proporcionalidad S_i, que es la saliencia del estímulo CS_i. Un estímulo puede ser más saliente que otro por tener mayor intensidad o por resaltar más del contexto (por ejemplo, una luz encendida en una habitación apagada tendrá más saliencia que en una iluminada).

Finalmente, la regla RW del cambio en la asociatividad entre un CS_i y un US_j es:

$$\Delta V_{CS_i} = \delta \, CS_i \, (A_j - \sum_{\forall j} V_{CS_j}) \tag{2.3.1}$$

2.3.5. Criticas al modelo de Rescorla y Wagner

Si bien el modelo permite a su vez explicar de manera muy sencilla y con asunciones básicas un gran número de fenómenos, no debería de ser sorprendente que falle en explicar

todos los resultados experimentales [35]. Quizás el peor defecto que tiene el modelo RW sea su falta de capacidad para explicar la inhibición latente, que es el retraso en el condicionamiento producido por la presentación de un CS sin recibir US durante varios ensayos, antes de presentar dicho CS seguido del US. Según el modelo, la preexposición al CS no puede cambiar el valor de la asociatividad de cero, dado que si $V = 0$ (inicialmente lo es) y no hay US, no puede haber condicionamiento.

Capítulo 3

Conductas de cooperación

En este capítulo se presenta una introducción a mecanismos que permiten lograr la evolución de la cooperación. Asimismo se introduce un problema de suma no nula conocido como el dilema del prisionero, que ha resultado ser uno de los problemas de la teoría de juegos más utilizado por diversas disciplinas para el análisis de la evolución de la cooperación.

3.1. Biología evolutiva

La selección natural puede ser concebida como una competencia de supervivencia en la cuál solo aquellos organismos que estén mejor adaptados a las condiciones existentes y a los cambios del entorno venideros van a poder sobrevivir y, fundamentalmente, reproducirse. La evolución, que esencialmente es el proceso continuo de transformación de las especies a través de cambios producidos en sucesivas generaciones, está basada en dicha competencia y debería, por lo tanto, favorecer solo aquellos comportamientos egoístas. Cada gen, cada célula, cada organismo debería ser diseñado para promover su propio éxito evolutivo a expensas de sus competidores.

Como consecuencia de esto, la pregunta de cómo la selección natural puede conducir a comportamientos cooperativos entre individuos ha resultado ser una de las más importantes paradojas que ha fascinado a los biólogos evolutivos durante décadas.

Los dos principios fundamentales de la evolución, si es posible la reproducción, son la mutación, que son alteraciones en la replicación genética, y la selección natural. Ambos principios forman la base del cambio evolutivo. Sin embargo, la evolución es constructiva debido a la cooperación. Nuevos niveles de organización evolucionan cuando las unidades que compiten a bajo nivel comienzan a cooperar. La cooperación permite la especialización y en consecuencia promueve la diversidad biológica, y por eso tal vez sea uno de los aspectos más notables de la evolución en un mundo que resulta competitivo. Según Nowak [26], tal vez debería ser agregada la cooperación natural como el tercer principio fundamental de la evolución, al lado de la mutación y la selección.

3.2. Cooperadores y desertores

Un cooperador es alguien que paga un costo c, para que otro individuo reciba un beneficio b. Un desertor en cambio, no tiene costo ni otorga beneficios aunque los reciba. Los costos y beneficios están medidos en términos de la capacidad de un individuo para sobrevivir y reproducirse. Esta medida se denomina fitness y es una medida relativa, dado que el éxito evolutivo de un individuo no está determinado por su fitness absoluto sino por el relativo con respecto al fitness de los demás individuos.

En cualquier población de cooperadores y desertores, los desertores tienen un mayor fitness promedio que los cooperadores. La selección actúa incrementando la cantidad relativa de desertores y luego de un tiempo, los cooperadores desaparecen de la población (ver figura 3.1). Sin embargo, una población formada únicamente por individuos cooperadores tiene el más alto fitness promedio, mientras que una población formada solo por desertores tiene el menor. En consecuencia, la selección natural constantemente reduce el promedio de fitness de la población.

A su vez, el fitness de los individuos depende de la cantidad relativa de cooperadores en la población. Entonces la selección natural en poblaciones bien mezcladas, donde la interacción de cada uno es con todos los demás, necesita de algún mecanismo que sirva de ayuda para establecer comportamientos cooperativos entre individuos de la población. A continuación se introducen algunos mecanismos que hacen que la evolución de la cooperación sea posible en un mundo donde prevalecen individuos egoístas.

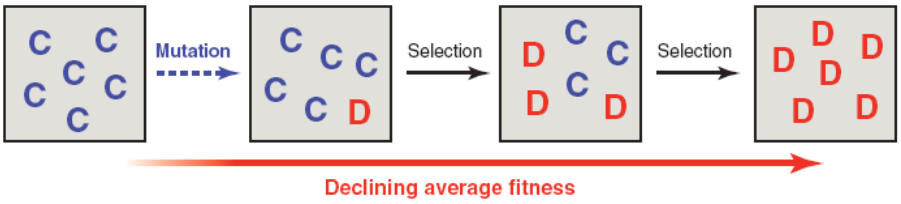

Figura 3.1: Bajo la ausencia de mecanismos para la evolución de la cooperación, la selección natural favorece a los desertores y disminuye el fitness promedio de la población. [Fuente: Nowak 2006]

3.3. Selección de parentesco

Según Hamilton [16], la selección natural puede favorecer a la cooperación si el donante y el receptor de un acto altruista están genéticamente relacionados. Más precisamente, la regla de Hamilton establece que el coeficiente de parentesco r debe de exceder el cociente costo-beneficio del acto altruista:

$$r > c/b$$

El coeficiente de parentesco r se define como probabilidad que un gen presente en un individuo sea una copia idéntica por descendencia de un gen presente en otro individuo [1]. Esta teoría de Hamilton se conoce como selección de parentesco (kin selection) o fitness inclusivo (inclusive fitness). Cuando se evalúa el fitness de un comportamiento inducido por un cierto gen, es importante incluir el efecto del comportamiento en el pariente que tal vez lleve ese mismo gen.

3.4. Reciprocidad directa

También puede observarse cooperación entre individuos no relacionados o incluso entre miembros de diferentes especies. Esas consideraciones impulsaron a Trivers [39] a proponer otro mecanismo para la evolución de la cooperación que es la reciprocidad directa.

[1] La probabilidad de que dos hermanos compartan el mismo gen por descendencia es $\frac{1}{2}$. La probabilidad entre abuelos y nietos será $1/4$ y entre primos es $1/8$.

Ahora se asume que hay repetidos encuentros entre dos individuos. En cada encuentro, cada jugador elige entre cooperar y desertar. Si un individuo coopera, luego el otro tal vez coopere o no. En consecuencia, puede que pague o no el hecho de cooperar. Este problema, conocido como el Dilema del Prisionero, es una pieza fundamental de la teoría de juegos para el estudio de la evolución de la cooperación y será detallado en una sección posterior.

3.5. Reciprocidad indirecta

La reciprocidad directa se basa en repetidos encuentros entre los mismos dos individuos. Ambos individuos deben poder proveer ayuda, lo cuál resulta menos costoso para el donante de lo que es de beneficioso para el receptor. Sin embargo, la interacción entre individuos puede ser asimétrica y efímera. Una persona está en una posición de ayudar a otros, pero no hay posibilidad para una reciprocidad directa. Ayudar a alguien establece una buena reputación [26], lo cual será recompensado por otros. Nowak asevera que, cuando decidimos cómo actuar, tenemos en cuenta las posibles consecuencias para nuestra reputación.

En el marco standard de reciprocidad indirecta, hay una elección al azar de pares de individuos que tienen un encuentro, pero que no necesariamente vuelvan a encontrarse. Un individuo actúa como el donante y el otro como el receptor. El donante puede decidir entre cooperar o no hacerlo. La interacción es observada sobre un subconjunto de la población quién puede informar al resto. Según Nowak [26], la reputación permite la evolución de la cooperación por reciprocidad indirecta. La selección natural favorece la estrategia que toma la decisión de ayudar basándose en la reputación del receptor. Aunque en animales se pueden encontrar simples formas de reciprocidad indirecta [4], solo los humanos parecen involucrarse con la complejidad del juego. La reciprocidad indirecta tiene demandas sustancialmente cognitivas. No solo se necesita recordar nuestras propias interacciones, sino que también debemos tener un monitoreo acerca de las interacciones del grupo.

Si bien los cálculos de reciprocidad indirecta resultan complejos, surge una regla básica que establece que la reciprocidad indirecta puede promover a la cooperación si la probabilidad q de conocer la reputación de alguien excede la razón costo-beneficio del acto altruista:

$$q > c/b$$

3.6. Redes de Reciprocidad

El argumento de la selección natural de la deserción (Fig 3.1) está basado en una población bien mezclada, donde todos interactúan de la misma forma con todos los demás. Sin embargo, las poblaciones reales no están bien mezcladas. Las redes sociales implican que algunos individuos interactúan más a menudo que otros. Una aproximación para estudiar este efecto es la Teoría de grafos evolucionarios [21], que permiten estudiar cómo las estructuras espaciales afectan a la dinámica de la evolución. Los individuos de una población ocupan los vértices de un grafo, los vértices determinan con quién interactúa. Los juegos sobre grafos son fáciles de estudiar por medio de simulaciones computacionales, pero son difíciles de analizar matemáticamente debido a la enorme cantidad de posibilidades de configuraciones a analizar. No obstante, existe una regla simple que determina cuándo una red de reciprocidad puede conducir a la cooperación [28], que establece que para cada individuo el cociente beneficio sobre costo debe exceder el promedio de la cantidad de vecinos k:

$$b/c > k$$

3.7. Selección de grupo

La selección actúa no solo sobre individuos sino también sobre grupos. Un grupo de cooperadores puede ser más exitoso que un grupo de desertores [38]. En un modelo simple de selección, una población es subdividida en grupos. Los cooperadores ayudan a otros en su propio grupo. Los desertores no ayudan. Los individuos se reproducen de manera proporcional a sus beneficios, y los descendientes se agregan al mismo grupo. Si un grupo alcanza cierto tamaño, se puede dividir en otros dos grupos. En este caso, otro grupo se extingue para limitar el tamaño total de la población. Aunque lo que se reproducen son los individuos, la selección surge a dos niveles distintos. Hay una competencia entre grupos porque algunos de ellos crecen más rápidamente y se subdividen más a menudo. En particular, los grupos formados puramente por cooperadores crecen más rápidamente que aquellos formados puramente por desertores, mientras que en grupos formados por cooperadores y desertores, los desertores se reducen más rápido que los cooperadores. Por lo tanto, la selección a bajo nivel (entre individuos) favorece a los desertores, mientras que la selección en altos niveles (entre grupos) favorece a los cooperadores. Asimismo, surge también una regla que indica que la selección de grupo permite la evolución de la cooperación si se cumple que: siendo n el tamaño máximo de un grupo y m el número de grupos, entonces

$$b/c > 1 + (n/m)$$

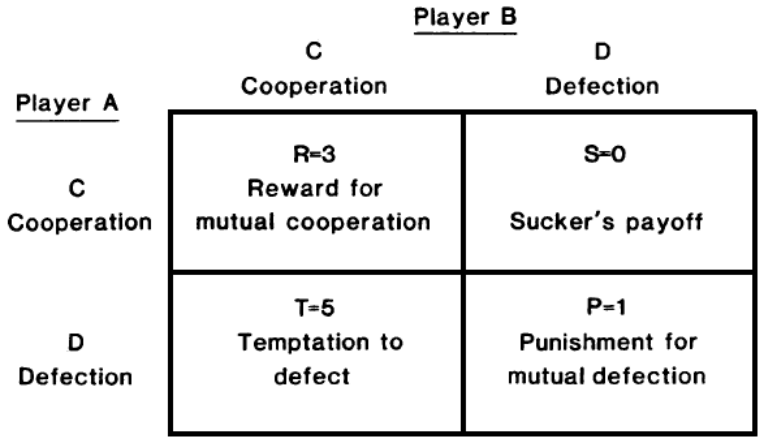

Figura 3.2: Matriz de pagos del dilema del prisionero. La paga del jugador A está asociada con valores a modo ilustrativo. El juego se define con las condiciones $T > R > P > S$ y $R > (S + T)/2$. [Fuente: Axelrod 1981]

3.8. El dilema del prisionero

Gran parte de los trabajos que se han realizado acerca de la evolución de la cooperación utilizan modelos de la teoría de juegos. Sin lugar a dudas, el trabajo que más repercusión ha tenido entre los modelos que examinan la reciprocidad y la evolución de la cooperación es el dilema del prisionero. El mismo, cobró una inusitada importancia a partir del trabajo de Axelrod y Hamilton [2].

En el juego, dos individuos pueden cooperar (C) o desertar (D). El pago a un jugador es en términos de fitness. En función de su propia elección y de la elección del contrincante, el jugador recibirá un puntaje que estará basado en una matriz de pago como se indica en la Figura 3.2.

No importa cuál es la decisión del otro, la estrategia egoísta de elegir la deserción conduce a una mayor paga que la cooperación. Con dos individuos destinados a encontrarse una única vez, la única estrategia que puede ser llamada una solución es la de desertar siempre, sin importar qué elección tome el contrincante. Esta estrategia se denomina deserción incondicional (ALLD).

Sin embargo, si ambos jugadores desertan, ambos obtienen un beneficio menor al de si ambos hubiesen cooperado. Aquí es donde se encuentra encerrada la paradoja.

Pero además de ser la solución en teoría de juegos, la deserción es también la solución en la evolución biológica. Es el resultado de una inevitable tendencia en la evolución de la mutación a la selección natural: si los pagos están en término de fitness, y la cantidad de interacciones entre pares de individuos es aleatoria y no se repiten, entonces cualquier población con una mezcla de estrategias evoluciona a un estado donde todos los individuos resultan de ser desertores. Según Axelrod [2], en muchos sistemas biológicos, los mismos individuos pueden encontrarse en una partida más de una vez. Luego la situación se transforma en lo que se denomina el dilema del prisionero iterado (DPI), que da lugar a un conjunto mucho más rico de posibilidades estratégicas.

Por medio de simulaciones computacionales, Axelrod y Hamilton evaluaron la performance de un conjunto de diversas estrategias presentadas en un torneo, en la búsqueda de aquellas estrategias que resulten evolutivamente estables. Esto significa que una vez que la estrategia es común en la población, no puede ser invadida o reemplazada por otra. El modelo asume que las elecciones se realizan de manera simultánea y con intervalos de tiempo discretos.

Si existe un número fijo de interacciones entre pares de individuos, la estrategia ALLD sigue resultando evolucionariamente estable y es la única estrategia que lo es. La razón de esto es debido a que la deserción en la última interacción es la opción óptima para ambas partes y en consecuencia también lo es para la interacción anterior, y así sucesivamente hasta la primera interacción.

Tal como marca Axelrod, un organismo no necesita un cerebro para emplear una estrategia. Las bacterias, por ejemplo, tienen una capacidad básica para jugar juegos en los cuales estas resultan altamente responsivas para seleccionar aspectos de su entorno, especialmente de su entorno químico. De esta forma pueden responder de manera diferenciada a aquello que están haciendo los otros organismos a su alrededor. Asimismo, estas estrategias son heredadas, y el comportamiento de una bacteria puede afectar el fitness de otros organismos alrededor de él, así como también el comportamiento de otros organismos puede afectar el fitness de una bacteria.

Mientras las estrategias puedan fácilmente incluir una sensibilidad diferencial a cambios recientes en el entorno o cambios acumulados en el tiempo, por otra parte el rango de sensibilidad es limitado. La bacteria no puede “recordar” o “interpretar” una compleja secuencia de eventos pasados, y probablemente no puede discernir el origen de cambios adversos o beneficiosos. Algunas bacterias, por ejemplo, producen sus propios antibióticos que resultan nocivos para los demás.

Al mismo tiempo, cuando uno sube peldaños en la escalera evolucionaria con respecto

a la complejidad neuronal, las partidas en los juegos se hacen más ricas en posibilidades. La inteligencia en primates, incluídos los humanos, permite un número considerable de mejoras: una memoria con mayores capacidades, más complejidad en el procesamiento de la información para determinar la próxima acción en función de la interacción pasada, y una mayor habilidad para distinguir entre diferentes individuos.

En el torneo organizado por Axelrod y Hamilton, algunas estrategias propuestas resultaban bastante intrincadas (incluso hasta un caso que usaba procesos markovianos e inferencia bayesiana). Sin embargo, la estrategia ganadora del torneo paradójicamente resultó ser la estrategia más simple. La misma fue planteada por Anatol Rapoport mediante un algoritmo que consistía en un par de lineas de código: tit for tat (TFT). Esta estrategia simplemente consiste en comenzar cooperando en la primera movida y luego hacer exactamente lo que hizo el contrincante en la movida inmediatamente anterior. Así, TFT resulta ser una estrategia robusta de cooperación basada en reciprocidad.

Axelrod y Hamilton [2] analizaron cuando TFT y ALLD son evolutivamente estables, es decir cuando son resistentes a una invasión de mutantes si ellos dominan la población. Demostraron las condiciones por las cuales TFT es resistente, mostrando que el éxito o fracaso de la cooperación depende de la probabilidad de encontrarse en un futuro con el mismo jugador. Sin embargo no demostraron que TFT sea evolucionalmente estable sino colectivamente estable, lo que significa que la performance de dicha estrategia es tan buena o mejor que la de cualquier mutante, ya que una estrategia de cooperar siempre (ALLC) puede introducirse en una población de TFT dado que obtendría la misma performance que cualquiera de éstos.

A partir de los trabajos de Axelrod, han surgido una gran cantidad de investigaciones en este campo, introduciendo diversas variantes estratégicas. Una de las variantes más interesantes fue la posibilidad de permitir a los contrincantes realizar jugadas "erróneas" (trembling hands, o fuzzy minds) en el desarrollo de sus estrategias. Estas perturbaciones en las estrategias resultan naturales desde un punto de vista biológico, dado que los animales cometen ciertos errores naturalmente haciendo que las conclusiones sobre las estrategias cambien notablemente. Por ejemplo, dos jugadores enfrentándose con la misma estrategia TFT, presentan una gran vulnerabilidad dado que si uno equivoca un movimiento y pasa a no cooperar, el otro jugador no cooperará en la jugada siguiente, produciendo una alternancia entre cooperar y desertar, disminuyendo así de manera significativa la performance de TFT debido a su falta de posibilidad de corregir errores. De ahí surgen diversas estrategias [25] para corregir este defecto en la recuperación de TFT.

En el modelo original de Axelrod, la decisión se toma de manera simultánea, ambos

jugadores deben decidir qué hacer al mismo tiempo. Sin embargo, las acciones de los animales usualmente están diferidas en el tiempo, existe un rol activo o de dador y uno pasivo o de receptor, y estos roles se intercambian en el transcurso de las interacciones. Además de esto, diversos experimentos de cooperación con animales han mostrado que es muy difícil obtener una reciprocidad sostenida [7]. En particular, se ha mostrado experimentalmente que los animales tienen una fuerte tendencia a la deserción en el DPI [7] [8] [10] [13]. Stephens et al [36] proponen que la cooperación no persiste cuando los animales juegan al DPI debido a la presencia de un efecto de fuertes descuentos temporales. A su vez, esta hipótesis está sustentada por estudios psicológicos de auto-control que establecen que los animales muestran una gran preferencia a recibir pequeñas recompensas inmediatas en lugar de grandes recompensas retardadas [1] [23] [29]. En los capítulos posteriores de este trabajo se analizan dichos fenómenos.

Capítulo 4

Redes Neuronales

En el siguiente capítulo se presenta una breve introducción a las redes neuronales artificiales que se utilizan para formalizar este trabajo. Asimismo, se detalla el mecanismo de aprendizaje de Hebb que permite modelar al aprendizaje operante de una manera más realista desde el punto de vista biológico.

4.1. Inspiración en la neurociencia

Tal como apuntan Hertz, Krogh y Palmer [19], el cerebro posee una serie de características altamente deseables en cualquier sistema artificial.

- Es mucho más rápido para realizar muchas tareas que la más veloz de las computadoras (por ejemplo, para el reconocimiento en imágenes).
- Es sumamente robusto y tolerante a fallas (diariamente mueren células sin afectar la performance de manera significativa).
- El procesamiento es altamente paralelo.
- Es flexible (puede adaptarse a un nuevo entorno "aprendiendo" sin necesidad de ser reprogramado).
- Puede manejar información difusa, probabilística, con ruido e inconsistente.

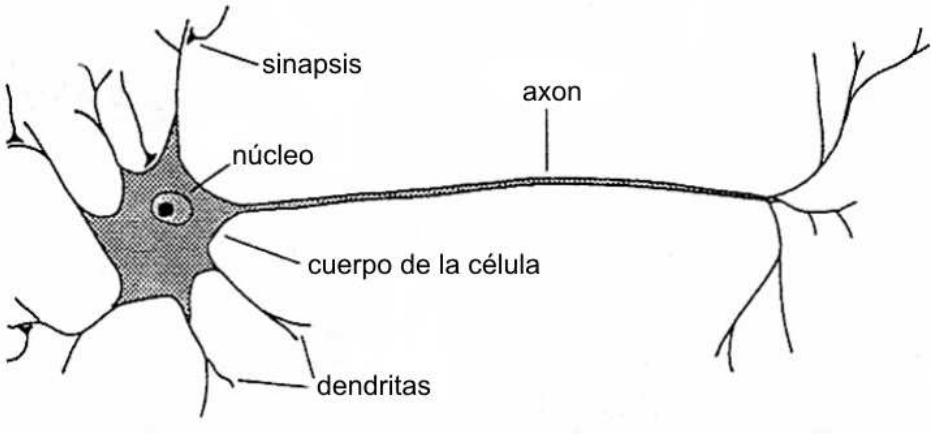

Figura 4.1: Diagrama esquemático de una neurona. [Fuente Hertz 1991]

- Es pequeño, compacto y disipa muy poca potencia.

Según Hertz, lo único en lo cual la computadora parecería superar al cerebro es en el cálculo aritmético. La inspiración original para las redes neuronales artificiales proviene del estudio del sistema nervioso central y de las neuronas, que constituyen las unidades elementales de procesamiento. Si bien estos modelos son simplificaciones extremas desde un punto de vista neurofisiológico, resultan ser una poderosa herramienta de cálculo y al mismo tiempo representan un nuevo paradigma metodológico en el campo de las ciencias cognitivas.

El cerebro humano está compuesto por unas 10^{12} neuronas de diferentes tipos ubicadas en el encéfalo, la médula espinal y los ganglios nerviosos y se encuentran en contacto con todo el cuerpo. Las dendritas transmiten los potenciales de acción desde las neuronas adyacentes hacia el cuerpo celular o soma, donde se encuentra el núcleo de la neurona. Como única salida del soma se extiende una fibra prolongada denominada axon, que a su vez puede ramificarse al final de la misma. Al final de estas ramificaciones se encuentran las uniones sinápticas o sinapsis hacia otras neuronas. Los terminales receptores de estas juntas pueden encontrarse tanto en las dendritas como en el cuerpo de la célula mismo (una sola neurona puede tener entre 1.000 y 10.000 sinapsis conectadas con hasta 70.000 neuronas).

La transmisión de una señal de una célula a otra mediante la sinapsis es un complejo proceso químico en el cual intervienen sustancias denominadas neurotransmisoras que son liberadas desde ambos lados de la unión neuronal. El efecto es el de subir o bajar el potencial eléctrico dentro del cuerpo de la neurona receptora. Si dicho potencial supera

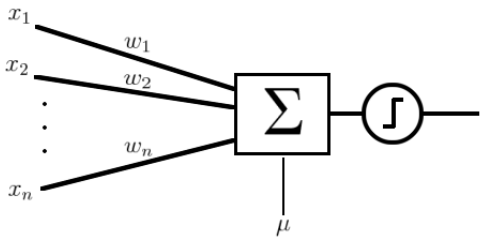

Figura 4.2: Diagrama esquemático de una neurona McCulloch-Pitts.

un umbral, un pulso o potencial de acción de una intensidad y duración fija es enviado hacia al axon de la misma. En ese caso estamos en presencia de un disparo de la neurona. Dicho pulso se dirige por las ramificaciones hacia las uniones con otras células. Luego del disparo, la célula debe esperar durante un período denominado refractario antes de poder volver a disparar [19].

4.2. Redes neuronales artificiales

Para simular las propiedades observadas en los sistemas neuronales biológicos, las redes neuronales artificiales son formalismos matemáticos que utilizan simples unidades de procesamiento o neuronas, las cuales al combinarse en diversas arquitecturas generan comportamientos complejos (propiedades emergentes).

Los precursores del modelado neuronal son McCulloch y Pitts (Fig. 4.2), que en 1943 propusieron un modelo muy simple de neurona en base a estudios neurofisiológicos. El mismo computa la suma ponderada de sus entradas provenientes de otras unidades y da como salida un 1 o un 0 dependiendo si la suma supera o no un determinado umbral, lo que significaría que la neurona está disparando o no, respectivamente.

En este modelo, el estado de disparo de la neurona i es expresado matemáticamente como:

$$x_i(t+1) = \sigma(\sum w_{ij} x_j(t) - i) \tag{4.2.1}$$

El peso w_{ij} se denomina peso sináptico y representa la fuerza de la conexión sináptica entre la neurona i y la neurona j. Si el mismo es positivo, se corresponde con una sinapsis excitatoria, y en el caso de ser negativo la sinapsis será inhibitoria. El valor n_k indica la activación de la neurona k. El valor μ_i es el valor del umbral de disparo para

la neurona i (si el umbral es superado la neurona dispara, y en el caso contrario no). Finalmente, la función σ se denomina función de activación y es la función característica de $\mathbb{R}_{\geq 0}$.

Pese a su extremada sencillez, McCulloch y Pitts probaron que la utilización de un conjunto particular de neuronas con sus pesos sinápticos elegidos correctamente, podía realizar cualquier función computable (esto no significa que la computación la realice de manera rápida ni conveniente).

Otro trabajo que ha resultado ser fundacional fue aquel realizado por el psicólogo Donald Hebb en 1949. En el mismo formula una hipótesis acerca de la forma en la cuál se producen cambios en los pesos sinápticos de las neuronas en respuesta a las sucesivas experiencias. Esta hipótesis se convirtió luego en una de las leyes básicas del aprendizaje. Lo que postula su regla es que la efectividad de una sinapsis entre dos neuronas se ve incrementada por la repetida activación de una neurona por otra, a través de dicha sinapsis. Más abajo se presenta en detalle esta regla de aprendizaje que será usada posteriormente en este trabajo.

4.2.1. Topología de redes

Como dijimos anteriormente, conectar las unidades simples de procesamiento de una manera adecuada nos permite generar comportamientos complejos. Imitando a un sistema neuronal biológico, las redes consisten en un conjunto de neuronas de entrada sensoras, que pueden conectarse con una red de neuronas intermedias u ocultas, que finalmente se conectan a las neuronas de salida. El tipo de patrón de conexiones que presenta una red neuronal, define la topología de la misma. En la Figura 4.3 se observan los dos tipos de topologías diferenciados.

El tipo más sencillo de topología de red es la feed forward, formada por sucesivas capas desde la entrada (pudiendo pasar por capas ocultas) hasta la salida. En la misma las señales circulan en una sola dirección desde las neuronas de entrada hacia las de salida, sin existir ciclos ni conexiones entre neuronas de una misma capa. A su vez, las redes feedforward se dividen en redes monocapa (ejemplo: perceptrón simple) y redes multicapa (ejemplo: perceptrón multicapa, redes de Kohonen).

Por otro lado, las redes recurrentes permiten la formación de ciclos en las conexiones neuronales de la red, y de esa forma permiten propagar señales desde capas posteriores a capas anteriores (ejemplo: Hopfield, Máquinas de Boltzman, Aprendizaje por refuerzo). [19]

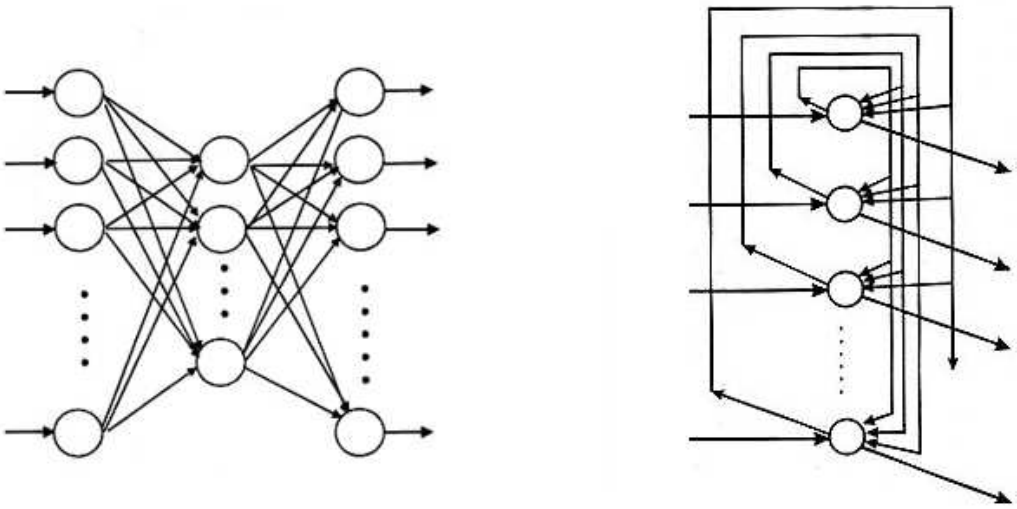

Figura 4.3: En la izquierda se ve una red feedforward - perceptrón multicapa. La derecha es un ejemplo de red recurrente - Hopfield.

4.2.2. Paradigmas de aprendizaje

Una de las características fundamentales a considerar de una red neuronal, es cuál es su paradigma de aprendizaje. Esencialmente existen tres paradigmas: el aprendizaje supervisado, el no supervisado y el aprendizaje por refuerzo. Usualmente, cualquier tipo de arquitectura de red dada puede utilizar cualquiera de dichos paradigmas de aprendizaje.

El aprendizaje supervisado o aprendizaje con maestro consiste en suponer la existencia de un maestro que tiene el conocimiento del entorno, y por lo tanto puede decir qué tan cerca o qué tan lejos se está del comportamiento deseado. Para esto existe un mapeo de entradas y salidas, y es el maestro quien sabe qué salida debe ser dada por la red neuronal frente a una entrada determinada. La diferencia entre lo deseado y lo que la red ha efectuado, es un costo de error (comúnmente el error cuadrático medio), que es el que controla el aprendizaje de la red. En este paradigma no sólo se sabe qué esta bien, sino también se tiene una medición de qué tan cerca o lejos se está de obtener el resultado correcto.

En el aprendizaje no supervisado, la red debe encontrar características, patrones o correlaciones en los datos de entrada, y codificarlos para la salida. Para que la red pueda hacer algo útil, los datos de entrada deben tener cierta redundancia, ya que sin éstos no

se podrían encontrar características ni patrones. Una red no supervisada nos puede dar como representación de sus salidas varias posibilidades [19]:

- **Familiaridad**: Una salida continua nos puede decir qué tan similar es una nueva entrada con respecto al promedio de las señales vistas en el pasado.
- **Análisis de Componentes Principales**: Si se extiende el caso anterior a varias salidas se puede construir un conjunto de ejes de los cuales se pueda medir similitud en distintos aspectos. Lo que se quiere lograr es reducir la dimensionalidad de entrada, para obtener un vector que contenga las características salientes del vector de entrada.
- **Agrupación (Clustering)**: Un conjunto de salidas binarias, de las cuales una sola está activa a la vez, nos puede informar a cual de varias categorías pertenece un determinado dato de entrada.
- **Obtener Prototipos**: Al igual que en el caso anterior, la red genera una clasificación del espacio de entrada, pero lo que devuelve es un prototipo de la clase.
- **Codificación**: La salida puede ser una versión codificada de menor cantidad de bits que la entrada, asumiendo que existe una red que pueda realizar la operación de decodificación.
- **Mapas de características**: Si las salidas tienen una conformación geométrica determinada, con sólo una neurona activa por vez, puede mapear los datos de entrada, en alguna posición de la estructura de salida. El objetivo es crear mapas topográficos de la entrada, con lo que se esperaría que emerja una organización global de la salida. Estos mapas se encuentran en varias partes del cerebro humano, como ser la corteza visual y la auditiva.

En el aprendizaje reforzado, el conocimiento que se adquiere del mapeo de entrada-salida se da a través de una interacción con el entorno de tal manera de maximizar la señal de refuerzo [37]. Una característica de este tipo de aprendizaje es el compromiso entre exploración y explotación. Para tratar de obtener mayores refuerzos, se deben preferir las acciones que realizó en el pasado y resultaron efectivas. Sin embargo, para descubrir tales acciones que no ha realizado antes, debe tomar acciones que no necesariamente correspondan con aquellas que hayan sido las aprendidas hasta el momento. Esto quiere decir que debe explotar lo que ya sabe para obtener beneficios, pero al mismo tiempo debe explorar para poder tomar mejores decisiones en el futuro. Otra característica importante es que el aprendizaje reforzado está fuertemente emparentado con la programación dinámica planteada por Bellman y el condicionamiento operante de los animales desde el punto de vista de la psicología experimental.

4.2.3. Algoritmos de Aprendizaje

El aprendizaje de las redes neuronales se logra mediante la flexibilidad de las mismas anteriormente enunciada. Esto significa que lo hace mediante su capacidad de adaptarse a un entorno para cumplir un objetivo determinado, que puede ser maximizar beneficios o minimizar un criterio de costos. El proceso de aprendizaje de la red es un proceso por el cual los parámetros de la red se adaptan al entorno cuando el mismo entorno estimula a la red. Existen diversos algoritmos para entrenar una red neuronal como el aprendizaje de Boltzman, competitivo, por refuerzo [19], pero vamos a centrarnos solo en dos que posteriormente formaran parte de este trabajo. Ellos son el Aprendizaje por gradiente descendente y el aprendizaje Hebbiano.

Aprendizaje por gradiente descendente (la regla delta)

Este tipo de aprendizaje plantea la búsqueda de una regla que permita encontrar, desde una condición inicial arbitraria y mediante aproximaciones sucesivas, un conjunto de pesos sinápticos plausibles de manera que se minimice una función de costo. La misma puede definirse como:

$$E(w) = 1/2 \sum_i \mu(\zeta_i^\mu - \sum_k w_{ik} \cdot \zeta_k^\mu)^2 \tag{4.2.2}$$

Dada la función de error $E(w)$, puede buscarse el mínimo de la misma en el espacio de la variable w. Para esto, el algoritmo del gradiente descendente sugiere cambiar cada w_{ik} por una cantidad Δw_{ik} proporcional al gradiente de E [19]. Finalmente, la expresión resultante se expresa como:

$$\Delta w_{ik} = \eta \cdot (\zeta_i^\mu - o_i^\mu) \cdot \zeta_k^\mu \tag{4.2.3}$$

Esta regla, comúnmente denominada delta, resulta ser formalmente equivalente a la regla RW de condicionamiento clásico que vimos anteriormente.

Aprendizaje Hebbiano

Tal como indica Haykin [17], la regla de Hebb enunciada anteriormente puede sintetizarse como:

1. Si dos neuronas en cada extremo de una sinapsis se activan de manera sincrónica (es decir, simultáneamente), la fuerza asociativa de dicha sinapsis deberá incrementarse.

2. Si dos neuronas en cada extremo de una sinapsis se activan de manera asincrónica, la fuerza asociativa de dicha sinapsis deberá verse disminuída.

Las principales características de este tipo de sinapsis son que cuenta con un mecanismo de dependencia temporal (es decir, depende del momento exacto de la ocurrencia de las señales presinápticas y postsinápticas), es un fenómeno local (significa que las señales deben ser contiguas espacio-temporalmente), y posee también mecanismos de correlación (dado que el cambio se debe a una ocurrencia de un conjunto de eventos). Para dejar en claro el tipo de sinapsis, podemos clasificar las sinapsis en Hebbianas cuando señales positivamente correlacionadas incrementan la efectividad de la sinapsis, y señales negativamente correlacionadas lo disminuyen, y anti-Hebbianas cuando se da a la inversa. La formalización matemática de la regla es:

$$\Delta w_{ij} = \gamma \psi_i \psi_j \tag{4.2.4}$$

donde Δw_{ij} representa al cambio del peso sináptico dada la actividad ψ_i en la neurona presináptica i, la actividad ψ_j de la neurona postsináptica j y γ es el parámetro de aprendizaje .

Si bien el estudio de las modificaciones sinápticas en el sistema nervioso central resulta complejo, existen evidencias neurofisiológicas relevantes que sustentan a esta regla como los resultados experimentales sobre la potenciación de largo plazo en el hipocampo [17].

Capítulo 5

Un modelo de cooperación evaluado con el dilema del prisionero iterado

En este capítulo se presenta un modelo computacional de aprendizaje operante, basado en evidencias neurofisiológicas y conductuales, capaz de predecir resultados obtenidos en diversos estudios de psicología experimental. Se explicarán dichas bases, la arquitectura de la red neuronal que lo modela y el modelo matemático subyacente. Finalmente, se explicará el funcionamiento del modelo.

5.1. Antecedentes

Como se ha mencionado anteriormente, el DP ha sido una herramienta abordada por diversas disciplinas para el estudio de cooperación. Existen modelos que utilizan este paradigma de cooperación para explicar el comportamiento económico de poblaciones en contraposición a los modelos económicos clásicos que consideran a los individuos como entidades racionales y egoístas [5].

Gutnisky y Zanutto muestran que un modelo de aprendizaje operante se comporta como una estrategia robusta al ser evaluado en el dilema del prisionero [14]. Además de esto, ante el problema de recuperación en TFT cuando se introducen perturbaciones

de las estrategias enunciadas anteriormente, el modelo de aprendizaje operante de Gutnisky y Zanutto tiene una performance ampliamente superior que TFT en estos casos. Sin embargo, en la construcción del modelo no se tuvieron en cuenta ciertos efectos que desean estudiarse y para los cuales dicho modelo resulta limitado.

En el presente trabajo se estudian los efectos del retraso en la entrega del refuerzos y su acumulación en la cooperación entre animales en el marco del DP.

5.2. Modelo propuesto

El modelo que se propone en este trabajo toma como punto de partida al modelo de condicionamiento operante de Lew et al [20] para estímulos apetitivos y aversivos. Utilizando las redes neuronales como una herramienta para la formalización de hipótesis neurofisiológicas y conductuales, este modelo predice resultados experimentales relevantes de condicionamiento operante como el desarrollo de preferencias, la Ley de matching, la extincion de respuestas en refuerzo parcial, el contraste negativo sucesivo, recuperación espontánea y los experimentos de escape y evitación simulados en [31].

El modelo que es presentado a continuación es una simplificación en algunos aspectos del modelo de Lew et al [20]. Por empezar, ha sido realizado con el fin de estudiar el aprendizaje por mecanismos de condicionamiento operante usando solo estímulos apetitivos. Por este motivo, las realimentaciones entre las respuestas y el área dopaminérgica, y las alimentaciones de US hacia las respuestas, ambas definidas para estudiar el aprendizaje aversivo en el modelo de Lew et al [20], han sido removidas. Otra modificación realizada es el mecanismo de elección de la respuesta ejecutada, que se basa en el cálculo de probabilidades de ejecución de cada respuesta, un mecanismo utilizado en modelos biofísicos de la Ley de matching [34].

Además, también es capaz de permitir estudiar el efecto que tiene el hecho de acumular alimento de manera perceptible pero no accesible por parte del individuo. Existe evidencia experimental que indica que este efecto es un mecanismo válido por el cual es factible lograr niveles de cooperación de manera sostenida en el marco del DP [36]. El objetivo entonces es poder explicar dicho efecto utilizando un mecanismo de aprendizaje operante.

5.2.1. La predicción del US

Desde bases neurobiológicas y conductuales, se asume que los animales tienen la capacidad para computar una predicción de US. Muchos experimentos de comportamiento sugieren que el aprendizaje está controlado por la expectación de los futuros eventos. Como vimos anteriormente, Rescorla y Wagner han propuesto que los animales aprenden comparando lo que esperan recibir en una situación dada y lo que efectivamente reciben. Por otro lado, existen sustratos neuronales involucrados en la predicción y el refuerzo, tales como los que están involucrados en las neuronas dopaminérgicas del área ventral tegmental (VTA) y la de la sustancia nigra pars compacta (SNc) [33].

Si bien la regla RW resulta simple y responde de manera análoga a estudios fisiológicos realizados en neuronas dopaminérgicas, tiene un defecto importante que hasta aquí no ha sido mencionado: no es un modelo de tiempo real sino que el cambio de los pesos se realiza a nivel inter ensayo. Esto en sí, significa que considera al aprendizaje como un proceso discreto. Esta falencia es reparada por la regla de RW en tiempo real [30], dado que propone que la actualización de las asociaciones se realice en cada instante de tiempo en lugar de ser actualizadas a nivel de ensayo.

Otros modelos computacionales como el método de Diferencias Temporales (TD) [37] predicen con bastante precisión el disparo dopaminérgico de VTA/SNc [33]. Sin embargo, a nivel neurofisiológico existen evidencias de que el tiempo de dopamina tiene un efecto prolongado en la corteza prefrontal y estriado (entre 1 y 1.5 seg) superando ampliamente al tiempo del disparo o potencial de acción [11].

Entonces, por ser un mecanismo más simple y que permite explicar los fenómenos a analizar y dado que el tiempo de cada ensayo será discreto en el orden de los segundos entre paso y paso, la RW de tiempo real resulta un mecanismo adecuado para la actualización de los pesos de las asociaciones.

Con respecto a la correlación de la predicción del refuerzo y el refuerzo, el efecto de omisión de este último ante la presencia de un estímulo que lo predice genera un error de predicción que deprime la actividad neuronal dopaminérgica por debajo del nivel basal [33]. Este resultado fue analizado mediante estudios de neuroimágenes en la corteza orbitofrontal en monos. El resultado es que ante la omisión del refuerzo se produce un efecto conocido como error de predicción negativa, mediante el cual se genera una señal negativa donde antes era positiva [27].

5.2.2. Memoria de corto plazo

La memoria de corto plazo, muchas veces llamada primaria o de trabajo, es un concepto abordado por la psicología experimental. Su principal característica es que es una memoria limitada en capacidad y duración. Desde un punto de vista biológico, la traza se define como la actividad neuronal persistente que genera dicha memoria frente a la ausencia de estímulos.

El término memoria de trabajo es usado comúnmente en lugar de memoria de corto plazo. Sin embargo, existe una diferencia conceptual dado que la memoria de trabajo solo es aquella que es usada para guiar la tarea que un individuo está realizando, mientras que la memoria de corto plazo no tiene esa restricción y en consecuencia resulta ser un término más general.

Existe evidencia indicando que en las cortezas orbito frontal [3] e infero-temporal [9] se generan memorias de corto plazo de los estímulos. Las neuronas de dichas áreas continúan en actividad cientos de milisegundos después de la desaparición de un estímulo, y la memoria persiste incluso varios segundos después de dicha desaparición.

5.2.3. La red neuronal

Esquemáticamente, la red neuronal propuesta se presenta en la figura 5.1.

El modelo basado en las evidencias neurofisiológicas y conductuales hasta aquí enunciadas, recibe como entrada todos los estímulos condicionados y el estímulo incondicionado. Tiene una neurona encargada del cálculo de la predicción, y para cada respuesta posible existe una neurona de respuesta.

Las vías que se mencionan a continuación y que son propuestas para participar en el mecanismo operante tienen sus bases en estudios de anatomía en vertebrados tales como ratas y monos. En los monos, la corteza prefrontal es una región de convergencia de cinco vías corticocorticales originadas en las áreas: sensorial somática primaria, auditiva, visual, olfativa y gustativa. Estas vías son relativamente independientes unas de otras hasta que alcanzan la corteza prefrontal que las asocia. En los primates, la corteza prefrontal es el origen de una cascada de conexiones que fluyen hacia la corteza promotora, y de ahí a la corteza motora primaria [3].

Por otra parte, las neuronas de VTA/SN_c se encuentran en el sistema dopaminérgico

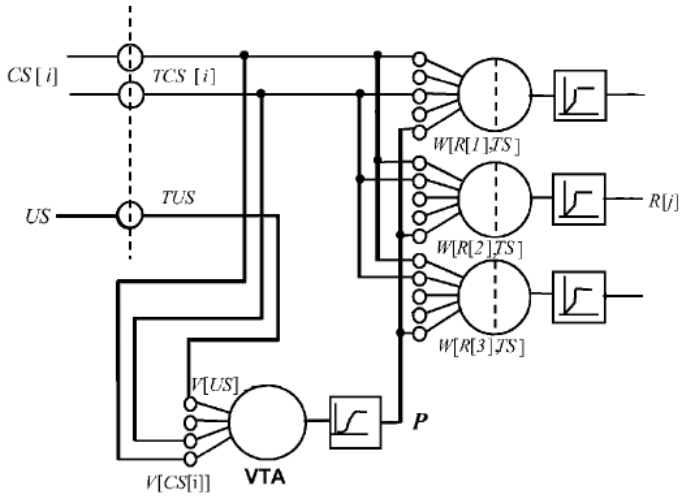

Figura 5.1: Modelo de red neuronal, se utiliza una neurona artificial que computa la prediccion (P) del US y una para cada respuesta. Las entradas a la neurona que computa P representan las trazas (memorias de corto plazo) de los estimulos condicionados (TCS[i]), del estimulo incondicionado (TUS). Los pesos sinapticos V[CS[i]], y V[US] representan las asociaciones entre las entradas y P. Las neuronas de respuesta tienen como entradas a P y las trazas TCS[i] de los estimulos condicionados (CS[i]) y la traza TUS del estimulo incondicionado (US). Los pesos sinapticos W[R[j],TS], son las asociaciones entre P, los estímulos y las neuronas de respuestas R[j].

y se conectan con la corteza prefrontal y ganglios basales mediante el sistema dopaminérgico mesocortical y nigro striatal.

En el modelo, existe una neurona que computa la predicción que representa a un cluster de neuronas de VTA/SN_c. Basándonos en el trabajo de Schultz et al [33], cuando la asociación entre un CS y US es aprendida, la neurona dispara ante la ocurrencia del CS, prediciendo al US. Asimismo, reportan los errores en la predicción por omisión de refuerzos, tal como fue explicado anteriormente. La neurona de predicción recibe como entrada a las trazas de los estímulos.

Las neuronas de respuesta del modelo, representan a un cluster de neuronas que se activan en forma de cascada en áreas frontales, promotoras y motoras, generando la respuesta conductual del animal. Las mismas reciben las trazas de los estímulos y la salida de la neurona de predicción.

Los pesos sinápticos de todas las entradas salvo del US (que permanece fijo) son plásticos, es decir, pueden modificarse. Schultz en [32] sugiere que la acción de la predicción *"puede ser formalizada aplicando la regla RW para actualizar los pesos sinápticos que la modulan"*. El modelo entonces va a utilizar este mecanismo para la actualización de dichos pesos sinápticos. Por otra parte, ese trabajo de Schultz también sugiere que *"la liberación de dopamina para el control de las sinapsis prefrontales podría modificar los pesos sinápticos de las neuronas de la corteza prefrontal de acuerdo a la regla Hebbiana"*, dependiendo del nivel de predicción. En consecuencia, el modelo va a actualizar el peso sináptico de las neuronas de respuesta siguiendo este mecanismo.

La selección de la respuesta que se ejecuta, por parte de la red, se realiza de manera estocástica. Para ello se utilizó en la estrategia del modelo para la Ley de matching propuesto por Wang [34]. La probabilidad de disparo de una neurona se calcula en función de la intensidad de la salida de la neurona. Aquella neurona que tenga una salida mayor, tendrá mayor probabilidad de ser la ganadora. El modelo no permite el disparo simultáneo de más de una neurona de respuesta, sino que existe una competencia durante cada ensayo.

Mientras que una teoría no formal carece de la posibilidad de describir qué es lo que pudiese ocurrir en un experimento suficientemente complejo, una teoría formal puede mostrar no solo qué responderá sino cómo. Es de interés para la constatación de las hipótesis del presente modelo poder evaluar su desenvolvimiento en simulaciones de experimentos de cooperación, y verificar los resultados que se predicen de manera experimental.

5.2.4. El modelo matemático

Con respecto a las memorias de corto plazo, las trazas de los estímulos son generadas ante la presencia de dichos estímulos y persisten según la siguiente dinámica que se indica a continuación. La traza del estímulo i (CS o US) se calcula como:

$$\tau S[i]_n = \begin{cases} (1-\varepsilon)\cdot \tau S[i]_{n-1} + \varepsilon \cdot S[i]_n, & \text{si } \ S[i]_n \neq 0 \\ \beta \cdot \tau S[i]_{n-1}, & \text{sino} \end{cases} \tag{5.2.1}$$

Donde $S[i]_n$ representa al valor de la saliencia del estímulo i en el instante n. Por otro lado, las trazas de las respuestas reciben como entrada a la salida de las neuronas de respuesta y se calculan según la expresión 5.2.2.

$$\tau R[i]_n = (1-\alpha_{DOWN})\cdot \tau R[i]_n + \alpha_{UP}\cdot R[i]_n \tag{5.2.2}$$

Donde $R[i]_n$ es la salida de la neurona de respuesta i en el instante n.

El valor de la amplitud de las trazas se encuentra acotado. Esto permite que las trazas de respuesta no crezcan indefinidamente, independientemente del valor de $R[i]$, y en consecuencia tampoco lo hagan sus pesos sinápticos.

La eficacia sináptica de la neurona que computa la predicción es calculada según la expresión 5.2.5.

$$X_n = V[US]_n \cdot \tau US_n + \sum_{i=1}^{Ncs} V[i]_n \cdot \tau CS[i]_n \tag{5.2.3}$$

donde $V[S]_n$ es el peso sináptico del estímulo S en el instante n, y τS el valor de la traza del estímulo S. Ncs es el número de estímulos no condicionados.

La salida de la neurona de predicción del estímulo incondicionado (P) en el instante n tiene la función de controlar el cambio sináptico de las neuronas de respuesta. En base a la eficacia sináptica descripta anteriormente, se calcula a P siguiendo la función una sigmoidea indicada en 5.2.5.

$$P_n = \frac{1}{1+e^{-\upsilon(X_n-\sigma)}} \tag{5.2.4}$$

Donde υ regula la rapidez del crecimiento de P y σ controla el retraso de su efecto. Por otro lado, las salidas de las neuronas de respuesta del modelo se calculan respondiendo a la dinámica descripta en 5.2.5.

$$R[i]_n = W[i][P]_n \cdot \gamma \cdot P_n + \sum_{i=1}^{Ncs} W[j][i]_n \cdot \tau CS[i]_n + noise(n) \tag{5.2.5}$$

La predicción es escalada por el factor γ, que regula la entrada de P en las neuronas de respuesta. Los $W[i][k]$ son los pesos sinápticos de la neurona de respuesta i a los CS y a la neurona de predicción. La función $noise(n)$ es ruido blanco en el instante n, y se obtiene muestreando una distribución uniforme $U(0, 1)$ escalada.

Como dijimos anteriormente, la selección de la respuesta a ejecutar sigue un proceso estocástico, donde la probabilidad de ejecutar la respuesta i estará dada por la expresión 5.2.6, donde Nr es el número de respuestas del modelo.

$$Pr(i) = \frac{|R[i]|}{\sum_{i=1}^{Nr} |R[j]|} \tag{5.2.6}$$

La forma de la elección de la respuesta planteada en [20] [14] [15] es cambia de ser un proceso determinista a uno estocástico. Esto permite ajustarse más al modelo biofísico del disparo de neuronas de respuestas motoras propuesto por Wang [34].

Con respecto a los pesos sinápticos de la neurona de predicción con el estímulo S se calculan siguiendo la regla de RW como se indica en la expresión 5.2.7

$$V[S]_n = V[S]_{n-1} + \eta(US_n).\tau S_n.(\Gamma(US_n, P) - P) \tag{5.2.7}$$

Donde $\eta(US_n)$ puede tomar dos valores que representan el tiempo de crecimiento (η_i) y de decrecimiento (η_d) de los pesos sinápticos, según esté presente o no el US en el instante n, respectivamente. Estos pesos sinápticos se encuentran acotados entre -1 y 1. El peso sináptico $V[US]$ se encuentra fijado en un determinado valor por las razones enunciadas anteriormente. La función Γ se define como:

$$\Gamma(US, P) = \begin{cases} -US, & \text{si } P > \lambda \text{ y } US = 0 \\ US, & \text{sino} \end{cases} \tag{5.2.8}$$

Esto modela al error de predicción para el caso de omisión de refuerzos, como se explicó anteriormente.

Los pesos sinápticos de la neurona de respuesta i se actualizan siguiendo las ecuaciones 5.2.9 y 5.2.10.

$$W[j][P]_n = W[j][P]_{n-1} \cdot \psi + \phi \cdot \gamma \cdot P_n \cdot \tau R[j]_n \cdot \Omega \tag{5.2.9}$$

$$W[j][CS_i]_n = W[j][CS_i]_{n-1} \cdot \psi + \phi \cdot \tau CS[i]_n \cdot \tau R[j]_n \cdot \Omega \tag{5.2.10}$$

El primer término de la ecuación es un momento de primer orden que previene el crecimiento no acotado de los pesos sinápticos. Siguiendo la regla de aprendizaje Hebbiano o anti-Hebbiano según el valor de la predicción del US, si P es menor que un umbral λ entonces $\Omega = \lambda$, en caso contrario $\Omega = -\lambda$.

Para simular la conducta exploratoria de animales, la probabilidad de generar respuestas aleatorias (Pb) decrece exponencialmente desde un valor inicial Pb_0. Esto permite simular el comportamiento de animales hambreados al comienzo de un experimento.

$$Pb_n = Pb_{n-1} \cdot \omega \tag{5.2.11}$$

5.2.5. Cómo funciona el modelo

Para entender el funcionamiento básico del modelo, a continuación se explica cómo se asocian la presentación de un estímulo, la selección de una respuesta y el refuerzo.

Se supone a un animal hambriento en una caja que contiene en su interior una luz y dos palancas P_1 y P_2. La palanca P_1 al ser presionada cuando la luz está encendida, activa un expendedora que le proporciona alimento al animal. Dependiendo de la cantidad de alimento proporcionado, al animal se lo alimenta durante una mayor o menor cantidad de tiempo. Bajo las mismas condiciones, al presionar la palanca P_2 no se proporciona alimento. Cualquiera de las palancas, al ser presionadas apagan la luz de manera que (por más que se siga presionando) no recibirá más alimento de ninguna forma hasta que la luz sea nuevamente encendida.

A continuación se analizará el funcionamiento del modelo con el siguiente experimento. Se enciende la luz durante 10 segundos (CS_i). La señal de la luz, proveniente de la corteza visual, genera una traza de memoria de corto plazo. Si se presiona la palanca P_1 (respuesta R_1) por azar, se recibirá alimento (US). Dado que el V_{US} es fijo, si el US es lo suficientemente significativo, hará que la neurona de predicción dispare por arriba de un umbral, incrementando el valor de los pesos sinápticos de las conexiones entre el CS_1 y la respuesta R_1 (el aprendizaje es Hebbiano), y entre el CS_1 y la neurona de predicción.

La salida de cada neurona de respuesta R_i se calcula en función de: los pesos sinápticos de todos los CS con la respuesta R_i y de la neurona de predicción con la respuesta R_i, el valor de las trazas de todos los CS y el valor de la predicción P. Como se expresó en la sección anterior, cada respuesta se ejecuta con una probabilidad definida en función de la salida de dicha respuesta. Entonces, en este caso, la próxima vez que aparezca CS_1 (es decir, se encienda la luz) la respuesta R_1 tendrá más chances de ser ejecutada, dada la modificación de los pesos sinápticos asociados al CS_1 y al incremento de P.

Si existe una diferencia entre el valor de predicción del US y el US efectivamente otorgado, el valor del peso sináptico se actualiza en función de dicha discrepancia, salvo que exista omisión de US cuando el mismo está siendo predicho. En este caso, el efecto del error de predicción negativo genera una pronunciada depresión en los pesos sinápticos de los CS con la neurona de predicción.

En los sucesivos ensayos, los pesos sinápticos asociados al CS_1 con la neurona de predicción, crecerán gradualmente hasta el punto que los mismos permitan predecir la venida del US. Esto significa que la sola presentación del CS hará que la neurona de predicción dispare, y el animal responda presionando la palanca P_1 al encenderse la luz.

Por otro lado, cuando el animal elige la respuesta incorrecta (R_2), sucede lo contrario. La neurona de predicción no dispara, el aprendizaje se hace anti-Hebbiano y el peso sináptico entre el CS_1 y la respuesta incorrecta se deprime, causando posteriormente una reducción en la salida de la respuesta R_2, y en consecuencia una disminución en la probabilidad de que se ejecute dicha respuesta en los ensayos subsiguientes.

Es importante notar que la asociación no se podría hacer sin la existencia de algún tipo de memoria, dado que la desaparición del CS es previa a la aparición del US. La superposición de los valores de las trazas de los todos estímulos del sistema juega un papel fundamental para las asociaciones de los eventos del ensayo.

Algo que no ha sido mencionado hasta ahora es el modelado de la presentación del alimento por medio del animal. Cuando el alimento es visto por el animal, el modelo recibe un estímulo particular (CS_{VE}) cuya saliencia es proporcional a la cantidad de US percibido. Este hecho ha sido suficiente para poder explicar el efecto de la acumulación que se estudia posteriormente en este trabajo.

En la figura 5.2 se presenta un diagrama de eventos para este ensayo propuesto y esos mismos eventos son reproducidos en el modelo. El tiempo es discretizado en pasos en el orden de segundos reales. En cada paso se realiza la iteración de un proceso que actualiza las variables del modelo. Los términos pasos e iteraciones, de aquí en más, son

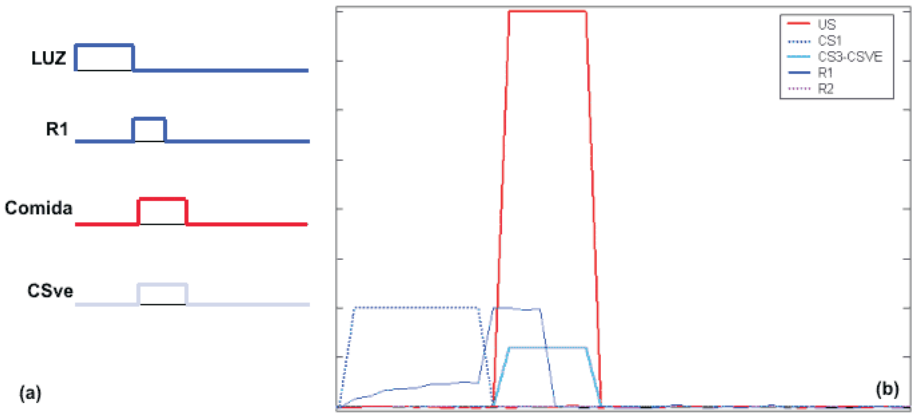

Figura 5.2: (a) Esquema de eventos en un ensayo. (b) El mismo ensayo pero en el modelo.

intercambiados pero se refieren al hecho de la actualización del estado del modelo en si.

Otra situación posible se representa esquemáticamente en la figura 5.3. Aquí se supone que el animal ve al alimento antes de poder consumirlo. Esto, por ejemplo, podría deberse a que el comedero no se encuentra cerca de la palanca donde se ejecutó la respuesta.

El efecto de la presentación del alimento es modelado como un estímulo más (CS_{VE}) con el fin de estudiar si éste podría ser un mecanismo posible para describir ciertos resultados experimentales que involucran la acumulación de alimento como una vía para lograr, bajo determinadas condiciones, el altruismo recíproco en animales.

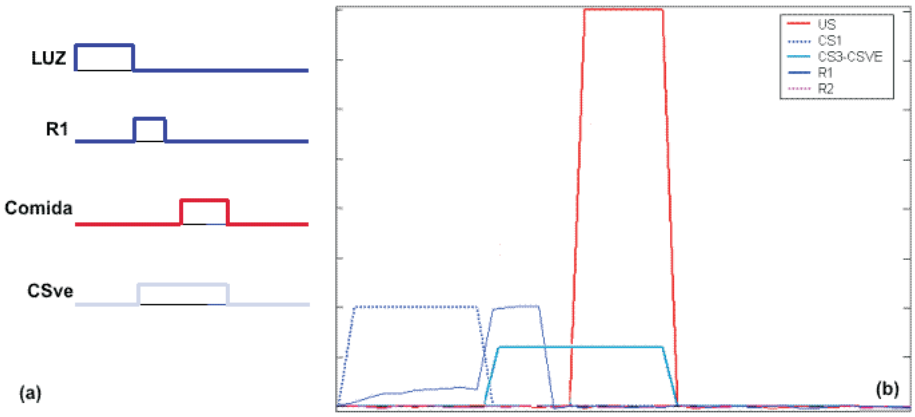

Figura 5.3: (a) Eventos de un ensayo donde la presentación del US es posterior a la ejecución de la respuesta. (b) Este mismo ensayo en el modelo.

Con respecto a los algoritmos que utiliza el modelo, si bien independientemente resultan sencillos de entender, en su conjunto conforman un sistema complejo que dificulta la tarea de realizar un seguimiento de la ejecución de una manera exhaustiva. Para simplificar este problema, y a los efectos de lograr facilitar la comprensión del mismo, a continuación se describen dos pseudo códigos que resumen el funcionamiento cómo es el comportamiento del modelo a lo largo de un ensayo. Las ecuaciones presentadas son las mismas que han sido definidas en la sección 5.4.2.

Un experimento está dado por una sucesión de ensayos separados por un tiempo inter ensayo. Durante este período se computan en cada paso el valor de los pesos sinápticos y las trazas, pero no se ejecutan respuesta y tampoco se reciben estímulos. Los valores de las trazas y de los pesos sinápticos son inicializados en cero. Ambos procedimientos descriptos a continuación se ejecutan durante cada paso de el ensayo de un experimento.

Procesamiento del modelo durante el paso de un ensayo

INPUT: modelo, CS, US, paso_actual

01. Inicializar el valor de todas las trazas en cero
02. Computar el valor de las trazas de estímulos y respuestas

$$\tau S[i]_n = \begin{cases} (1-\varepsilon)\cdot \tau S[i]_{n-1} + \varepsilon \cdot S[i]_n, & \text{si } \ S[i]_n \neq 0 \\ \\ \beta \cdot \tau S[i]_{n-1}, & \text{sino} \end{cases}$$

$$\tau R[i]_n = (1-\alpha_{DOWN})\cdot \tau R[i]_n + \alpha_{UP}\cdot R[i]_n$$

03. Actualizar el valor de la salida de la neurona de predicción

$$X_n = V[US]_n \cdot \tau US_n + \sum_{i=1}^{Ncs} V[i]_n \cdot \tau CS[i]_n$$

$$P_n = \frac{1}{1+e^{-\upsilon(X_n-\sigma)}}$$

04. Actualizar el valor de la salida de todas las neuronas de respuesta

$$noise(n) = \rho * U(0,1)$$

$$R[i]_n = W[i][P]_n \cdot \gamma \cdot P_n + \sum_{i=1}^{Ncs} W[j][i]_n \cdot \tau CS[i]_n + noise(n)$$

05. Actualizar la probabilidad de elegir una respuesta al azar

$$Pb_n = Pb_{n-1}\cdot \omega$$

06. Si no hay una respuesta ejecutándose
07. Si el paso_actual del ensayo es mayor que el limite inferior de pasos para activar una respuesta
08. Si al menos una respuesta supera el umbral de disparo
09. Elegir una respuesta ganadora con probabilidad

$$Pr(i) = \frac{|R[i]|}{\sum_{i=1}^{Nr}|R[j]|}$$

10. Las demás respuestas se fuerzan a cero + $noise(n)$.
11. Mantener esa respuesta activa durante el tiempo de ejecución de una respuesta.
12. FinSi
13. FinSi
14. Si el paso_actual del ensayo es mayor que el limite superior de pasos para activar
15. Elegir una respuesta al azar con probabilidad Pb
16. FinSi
17. Si se supero el tiempo de ejecución de una respuesta
18. Forzar la salida de la respuesta en ejecución a 0 + $noise(n)$
19. FinSi
20. FinSi
21. Se actualizan los pesos sinápticos asociados a la neurona de predicción

$$V[S]_{n+1} = V[S]_n + \eta(US_n).\tau S_n.(\Gamma(US_n) - P)$$

$$\Gamma(US, P) = \begin{cases} -US, & \text{si } P > \lambda \text{ y } US = 0 \\ US, & \text{sino} \end{cases}$$

22. Se actualizan los pesos sinápticos asociados a las neuronas de respuesta

$$W[j][P]_n = W[j][P]_{n-1} \cdot \psi + \phi \cdot \gamma \cdot P_n \cdot \tau R[j]_n \cdot \Omega$$

$$W[j][CS_i]_n = W[j][q]_{n-1} \cdot \psi + \phi \cdot \tau CS[i]_n \cdot \tau R[j]_n \cdot \Omega$$

Durante cada paso, al mismo tiempo se realiza el cálculo del US según el siguiente criterio.

Asignación de refuerzo durante un ensayo

INPUT: modelo, paso_actual

```
Si la respuesta que se activo es correcta y paso_actual es igual al retraso de la venida del refuerzo
    Asignar US durante la cantidad de pasos asociados al refuerzo
FinSi
```

Estos pseudo códigos no describen exactamente cómo se realiza el procesamiento del modelo según la implementación del mismo. Se dispusieron de esa forma con el fin de lograr una mejor comprensión del cálculo del modelo.

Los parámetros utilizados en las simulaciones se presentan en el Apéndice A.

Capítulo 6

Simulaciones

En este capítulo, inicialmente se presentan algunos experimentos básicos que han de ser explicados por un modelo formal de condicionamiento operante. A medida que se avanza en las diferentes secciones, los experimentos aumentan su complejidad. Finalmente se trata de explicar por medio de simulaciones en el modelo de condicionamiento operante propuesto, los efectos de la acumulación como mecanismo para lograr cooperación en el DPI.

6.1. Selección de una respuesta - extinción

El objetivo de este experimento es simular que se produzca el condicionamiento de una única respuesta que otorga recompensas, generando un incremento entre la asociación del estímulo y dicha respuesta. Este proceso es conocido como adquisición (ver sección 2.3.1). Para dar un paso más en el análisis en la manera en que se pueda verificar los efectos de la extinción (el otro fenómeno básico de condicionamiento, explicado en 2.3.1), se debe asociar que la presencia del CS no estará seguida del US.

6.1.1. Simulación

El modelo considerado de aquí en más tiene 4 CS: CS_0, CS_1, CS_2, CS_{VE} y 3 posibles respuestas: R_0, R_1, R_2. No existe una restricción potencial de la cantidad de estímulos y

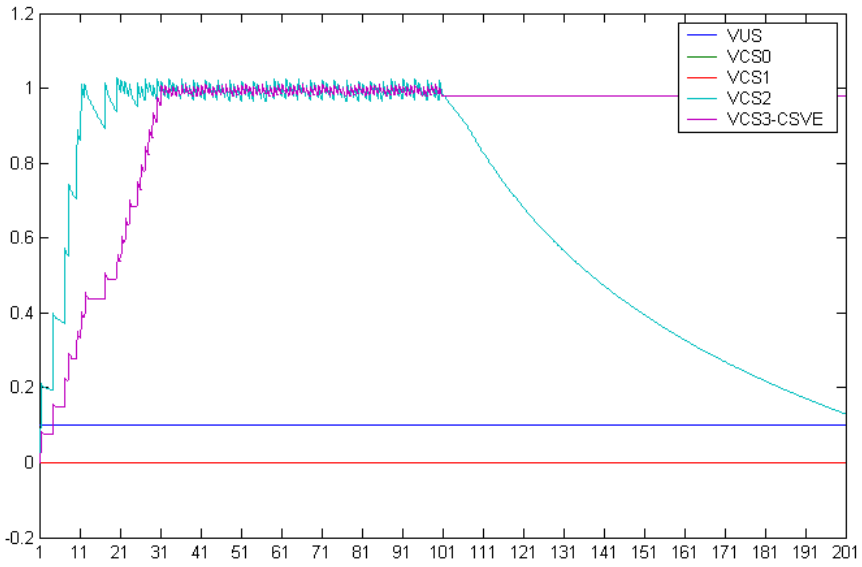

Figura 6.1: Pesos sinápticos de la neurona de VTA/SN_c.

respuestas del modelo, si no que esta cantidad resulta suficiente para afrontar el desafío de los experimentos a estudiar. Para simplificar la escritura en algunas situaciones, se describen los estados en un vector CS, para el caso de los CS de entrada, y R para el caso de las respuestas. Los parámetros que utiliza el modelo se encuentran en el apéndice A.

En este experimento, como entrada, se recibe CS=[0,0,1]. Se realizan 100 ensayos de adquisición y luego 100 ensayos de extinción, de manera consecutiva. La condición impuesta para los primeros 100 ensayos es que se recibe US de 6 pasos si y solo si ante la presencia de CS_2, el modelo ejecuta la respuesta R_2. La condición para los siguientes 100 ensayos para comprobar los efectos de la extinción es que, ante la misma entrada, sin importar la respuesta que ejecute, la recompensa será nula.

6.1.2. Resultado

En la figura 6.1 se muestran los pesos sinápticos de VTA/SN_c a lo largo del experimento. Se puede observar inicialmente el incremento sostenido de $V[CS_2]$ y de $V[CS_{VE}]$ hasta estabilizarse alrededor de 1, valor donde se encuentran acotadas. Si se observa la ecuación de actualización de V, claramente se observa que solo es posible un incremento en el valor de V ante la presencia de US (ver figura 6.3) y por las condiciones impuestas

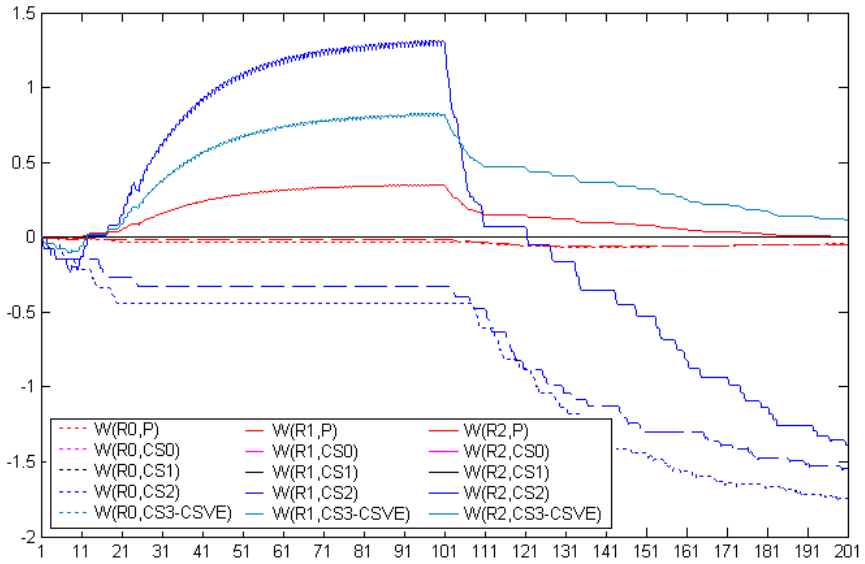

Figura 6.2: Pesos sinápticos de las neuronas de respuesta.

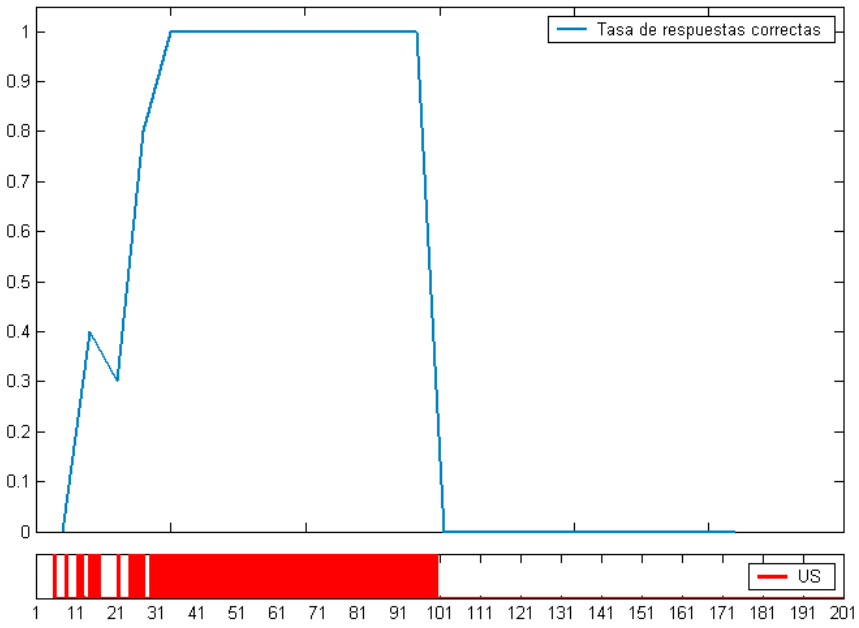

Figura 6.3: Tasa de respuestas correctas y US asignado.

por el experimento, solo es posible si el modelo responde R_2 ante la presencia del estímulo CS_2. Luego, al comenzar el proceso de extinción se puede ver una caída en el valor del peso sináptico de CS_2 mientras que el peso sináptico de CS_{VE} permanece estable alrededor de 1. Este efecto es el correcto según la regla de RW puesto que, en primer lugar, se produce un efecto inhibitorio en el peso de CS_2 dado que existe una asociación alta con el US y sin embargo no se recibe US. Entonces, ante el estímulo CS_2 se esperaba recibir US pero no se recibe, entonces la fuerza asociativa se deprime. En segundo lugar, el peso sináptico asociado a CS_{VE} no varía. Esto puede resultar confuso, pero efectivamente sigue respetándose la regla RW, puesto que al no presentarse nuevamente el CS_{VE}, la regla de RW no se evalúa y en consecuencia el peso sináptico no se modifica.

En la figura 6.2 se puede observar que durante la adquisición, los pesos sinápticos que se incrementan solo son aquellos pertencientes a R_2 (los demás, permanecen por valores menores o iguales que cero). En particular, el peso que cobra más fuerza es el de R_2 con CS_2, mientras que el peso de dicha respuesta con los demás CS no varian de 0. Pese que al comienzo del experimento, el aprendizaje es Hebbiano y anti-Hebbiano puesto que el modelo está explorando y respondiendo al azar, aproximadamente luego del ensayo 20, los pesos sinápticos de dicha neurona de respuesta comienzan a tener una asociación Hebbiana de manera sostenida, y en consecuencia, el valor del peso sináptico se incrementa. Por otro lado, durante el proceso de extinción se puede observar una depresión sostenido en los valores de los pesos sinápticos. Durante esta etapa, la asociación es anti-Hebbiana.

6.2. Reversión

El entrenamiento de reversión involucra revertir las contingencias del refuerzo. Luego del condicionamiento de una respuesta R_j por parte de un CS_i, se deja de recibir CS_i y se comienza a recibir CS_k $(k \neq j)$ y trata de lograr el condicionamiento de R_j por parte de CS_k.

6.2.1. Simulación

Se realizan 200 ensayos de adquisición y luego 200 ensayos de reversión, de manera consecutiva.

Inicialmente, como entrada recibe CS=[0,0,1]. La condición impuesta para los primeros 200 ensayos es que se recibe un US de 6 pasos si y solo si ante la presencia de CS_2, el

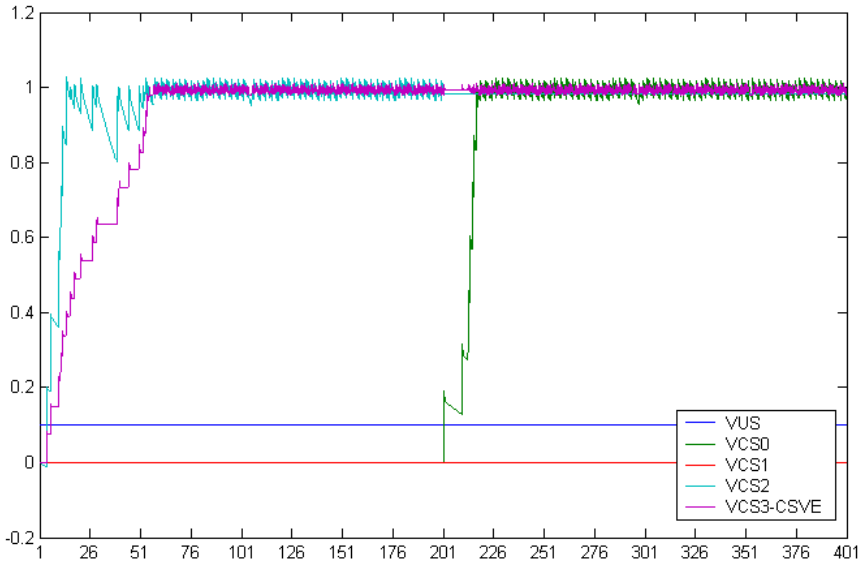

Figura 6.4: Pesos sinápticos de la neurona de VTA/SN_c.

modelo ejecuta la respuesta R_2.

Luego recibe como entrada, CS=[1,0,0], y la condición impuesta para los siguientes 200 ensayos pasa a ser que se recibe US de 6 pasos si y solo si ante la presencia de CS_0, el modelo ejecuta la respuesta R_2.

6.2.2. Resultado

Con respecto a la figura 6.4, que muestra los pesos sinápticos asociados a la predicción de CS_2 y de CS_{VE} incrementan su fuerza al igual que en el caso de la adquisición mostrado en la figura 6.1. Por otra parte, en la reversión crece el valor del peso sináptico de CS_0, mientras que el peso de CS_2 permanece constante. Al igual que fue explicado en la sección inmediata anterior, la actualización de estos pesos sinápticos se realiza según la regla RW y en consecuencia, se actualiza el valor de un peso sináptico de un CS solo si dicho CS forma parte de alguna contigencia CS-US. Como el CS_2 no vuelve a presentarse en el resto del experimento, su valor permanece constante. Esto no sucede con el peso de CS_{VE}, que sufre pequeñas variaciones cerca de su cota superior.

Por otra parte, en la figura 6.5 se exhiben los valores de los pesos sinápticos asociados

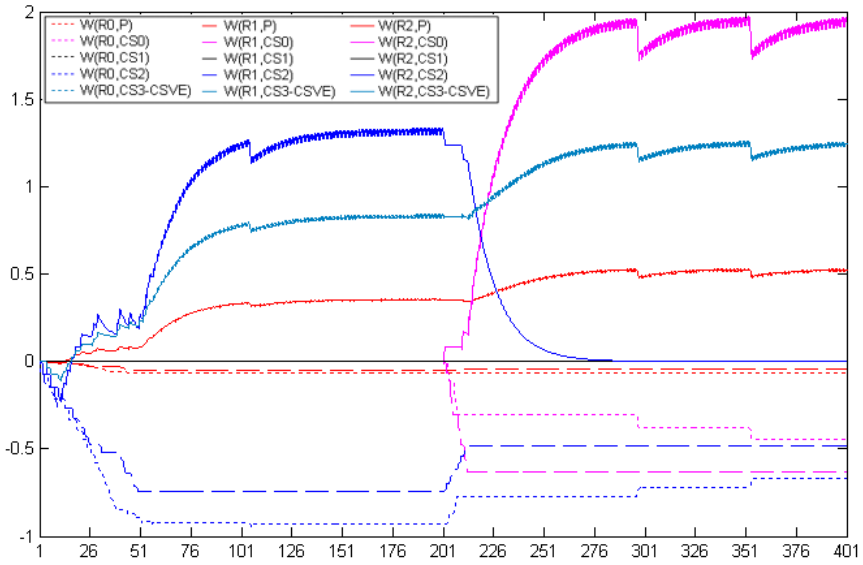

Figura 6.5: Pesos sinápticos de las neuronas de respuesta.

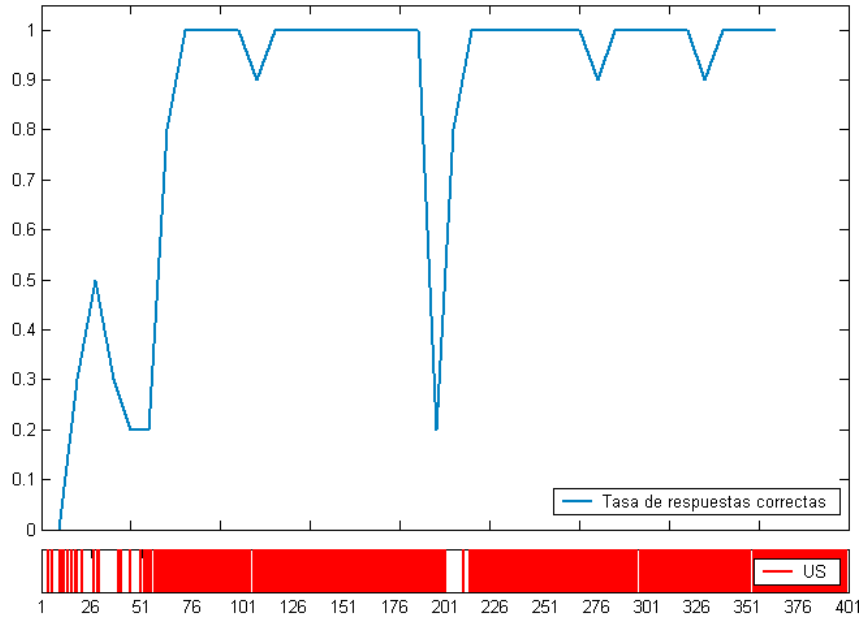

Figura 6.6: Tasa de respuestas correctas y US asignado.

a las neuronas de respuesta. Durante el proceso de reversión, se ve una caida sostenida en el peso que asocia R_2 con CS_2, mientras que al mismo tiempo se fortalece la asociación entre R_2 y CS_0. Sin embargo, se puede notar que los pesos de P y CS_2, casi no sufren la consecuencia de la reversión, y se incrementan de manera sostenida.

6.3. Discriminación

El concepto de discriminación está asociado a la habilidad de un animal para tratar situaciones similares de manera diferente. Esto se refiere a la variación en las situaciones (reflejada en los estímulos) que el animal es capaz de tratar de manera diferente [35].

Durante un experimento de discriminación, se muestran distintos CS. El individuo tiene la capacidad de discriminar de manera consistente entre la presencia de un CS y otro. Para verificarlo, se trata de comprobar si cada CS_i puede condicionar una respuesta distinta R_j.

6.3.1. Simulación

El modelo recibe como entrada CS=[1,0,0] y CS=[0,0,1] de manera alternada entre ensayo y ensayo. Se recibe un US de 6 pasos si y solo si ante la presencia de CS_0 se ejecuta la respuesta R_0 y ante la presencia de CS_2 se ejecuta la respuesta R_2. La cantidad total de ensayos es de 250.

6.3.2. Resultado

En la figura 6.7 se puede observar el incremento de los pesos sinápticos asociados a la neurona de predicción de CS_0, CS_2 y CS_{VE}. Por otra parte, la figura 6.8 muestra cómo es el incremento de los pesos sinápticos de las neuronas de respuestas y los CS correspondientes. Los valores más grandes de W los tienen R_2 con CS_2 y R_0 con CS_0, respectivamente. Ambos pesos tienden al mismo valor asintótico. Esto mismo ocurre entre R_2 y R_0 con P y entre R_2 y R_0 con CS_{VE}.

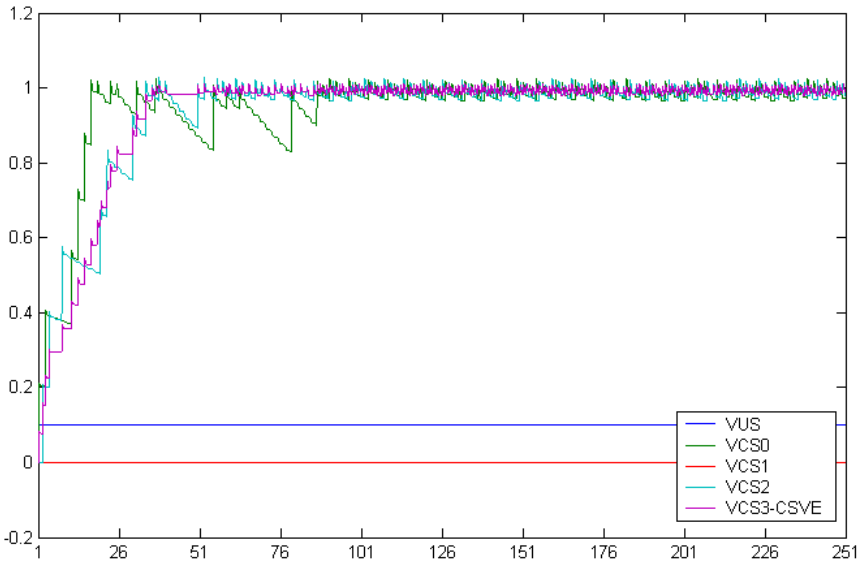

Figura 6.7: Pesos sinápticos de la neurona de VTA/SN_c.

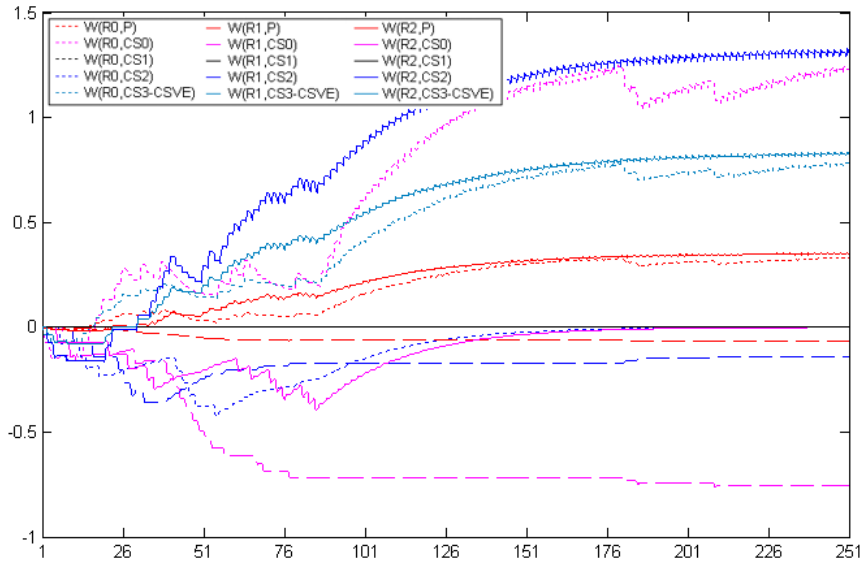

Figura 6.8: Pesos sinápticos de las neuronas de respuesta.

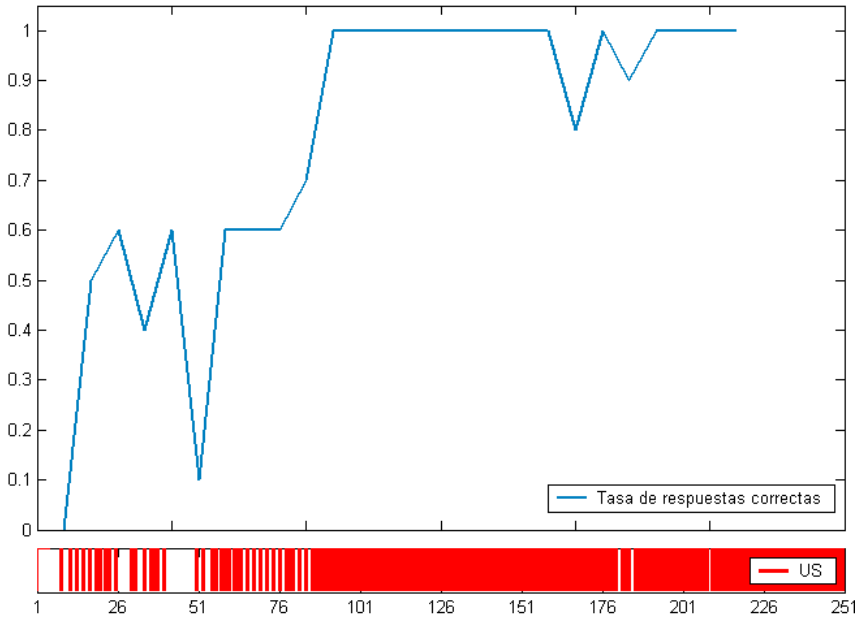

Figura 6.9: Tasa de respuestas correctas y US asignado.

6.4. Reversión de discriminación

Es una combinación de los dos experimentos anteriores. Una vez que se aprende a discriminar, se efectúa la reversión de respuestas.

6.4.1. Simulación

El modelo recibe como entrada CS=[1,0,0] y CS=[0,0,1] de manera alternada entre ensayo y ensayo. Se recibe un US de 6 pasos si y solo si ante la presencia de CS_0 se ejecuta la respuesta R_0 y ante la presencia de CS_2 se ejecuta la respuesta R2. Una vez que se realizan 250 ensayos, que como vimos anteriormente, la discriminación de estímulos la efectúa de una manera sostenida, la regla de pago cambia y ahora recibe un US de 6 pasos si y solo si ante la presencia de CS_0 se ejecuta la respuesta R_2 y ante la presencia de CS_0 se ejecuta la respuesta R_2. La cantidad total de ensayos de esta fase es de 250.

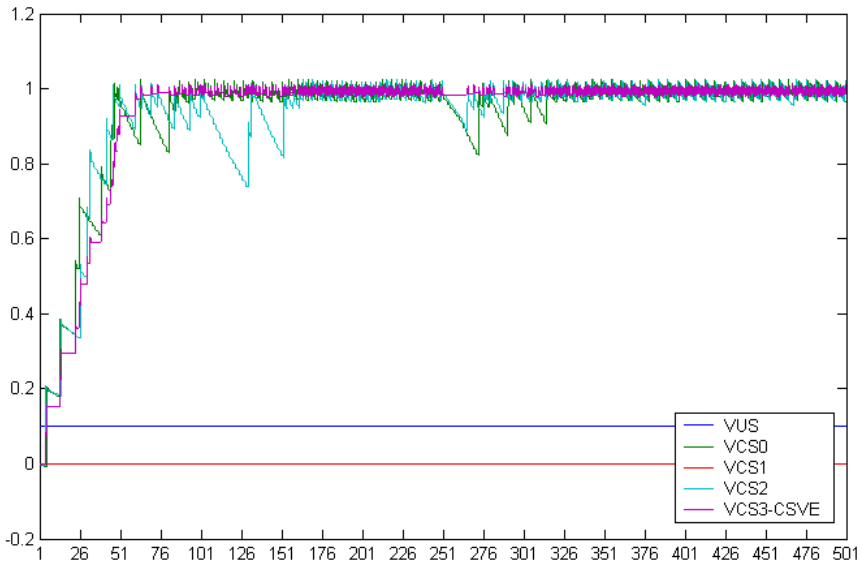

Figura 6.10: Pesos sinápticos de la neurona de VTA/SN_c.

6.4.2. Resultado

A medida que los experimentos van resultando cada vez más complejos, es más difícil de interpretar qué es lo que está sucediendo en un experimento mirando los gráficos que se generan como resultado. Sin embargo, el resultado de este experimento puede ser analizado sin inconvenientes. Durante los primeros 250 ensayos, se produce el mismo resultado que en el experimento anterior, luego durante el aprendizaje de la reversión, los pesos caen pero se fortalecen aquellos pesos que corresponden, de manera que se logra la reversión de la discriminación de manera satisfactoria.

6.5. Ley de Matching

En una gran cantidad de situaciones, los animales deben optar por una respuesta u otra, que les determinará más o menos posibilidades de obtener refuerzos. Es sorprendente encontrar que en gran cantidad de oportunidades, animales de diversas especies responden a cada evento en la misma proporción que los refuerzos obtenidos.

En el experimento de Herrnstein [18], se usan palomas que son expuestas a una elección entre dos posibles respuestas en un esquema de condicionamiento de intervalos

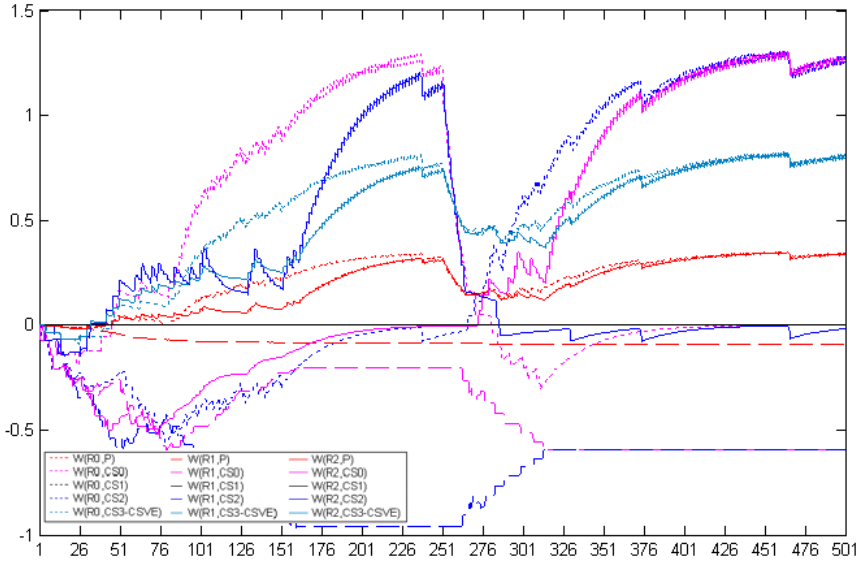

Figura 6.11: Pesos sinápticos de las neuronas de respuesta.

variables de refuerzos (denominado VI): al presionar una de las palancas se da refuerzo cada 135 segundos, mientras que presionar en la otra lo hace cada 270 segundos (siendo independientes una de otra). Lo que se obtiene como resultado es que la proporción con que se entregan los refuerzos es igual a la proporción con que las palomas presionan cada una de las palancas. En el caso que se mencionó, la primera palanca es apretada la tercera parte de las veces y la segunda palanca en la dos tercera parte.

En base a esta evidencia, Herrnstein formuló el principio conocido como Ley de matching (*Matching law*). Dicha ley establece que si R_1 y R_2 son las respuestas en cada una de las palancas, B_1 y B_2 son los retrasos asignados a cada palanca, la ley establece que:

$$\frac{R_1}{R_1 + R_2} = \frac{B_1}{B_1 + B_2} \tag{6.5.1}$$

Una de las razones por la cual los investigadores vieron como un importante principio a la ley de matching es que pudo ser comprobada en una gran cantidad de experimentos en diferentes especies de animales.

La Ley de matching está originalmente formulada para refuerzos de igual magnitud y la variable que se controla es el retraso en los refuerzos asociados a cada palanca. A

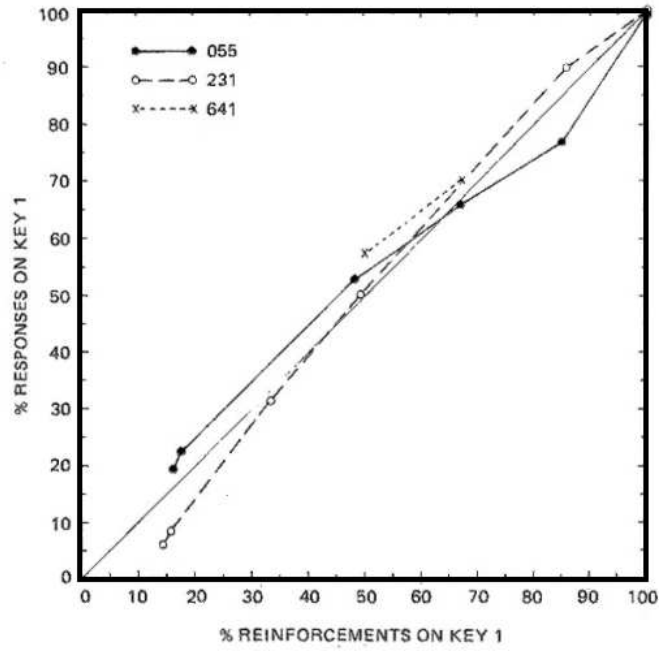

Figura 6.12: Ley de matching de Herrnstein. Los puntos representan sus resultados en palomas obtenidos por Herrnstein. La diagonal simboliza el matching perfecto.

partir de esta propiedad surgieron diversos estudios que modificaron a la Ley de matching original en cuanto a las variables involucradas. Alternativas como el tipo de refuerzo, la cantidad y la incorporación de castigos para las opciones a elegir, han mostrado la versatilidad de este principio.

En particular, Catania encontró resultados consistentes con la ecuación 6.5.2 en un estudio con palomas que tienen la opción entre 3 y 6 segundos de refuerzo [6]. De esa forma se establece la Ley de matching para refuerzos definiendo que, si R_1 y R_2 son las respuestas en cada una de las palancas, US_1 y US_2 son las magnitudes de los refuerzos asociados a cada palanca, entonces:

$$\frac{R_1}{R_1 + R_2} = \frac{US_1}{US_1 + US_2} \tag{6.5.2}$$

Esta variación es la que se va a estudiar en este apartado, con el fin de observar cómo se comporta el modelo ante la opción de elegir distintas cantidades de refuerzo.

6.5.1. Simulación

Se presentan distintas relaciones entre los refuerzos y para cada configuración se ejecutan 500 experimentos. Con respecto a la configuración de los refuerzos evaluados, el US_1 varía desde 3 hasta 17 y $US_2 = 17 - US_1$.

6.5.2. Resultado

En la figura 6.13 se puede apreciar en los resultados obtenidos para la ley de matching en refuerzos. Luego del aprendizaje, el modelo responde en cada opción según el refuerzo otorgado. El equilibrio en el nivel de respuestas ejecutadas se va logrando cuando los refuerzos de ambas opciones van tendiendo al mismo valor, entonces la cantidad de respuestas en cada palanca alcanza un nivel de equiprobabilidad. Esta equiprobabilidad representa la indiferencia del animal por sobre una de las respuestas.

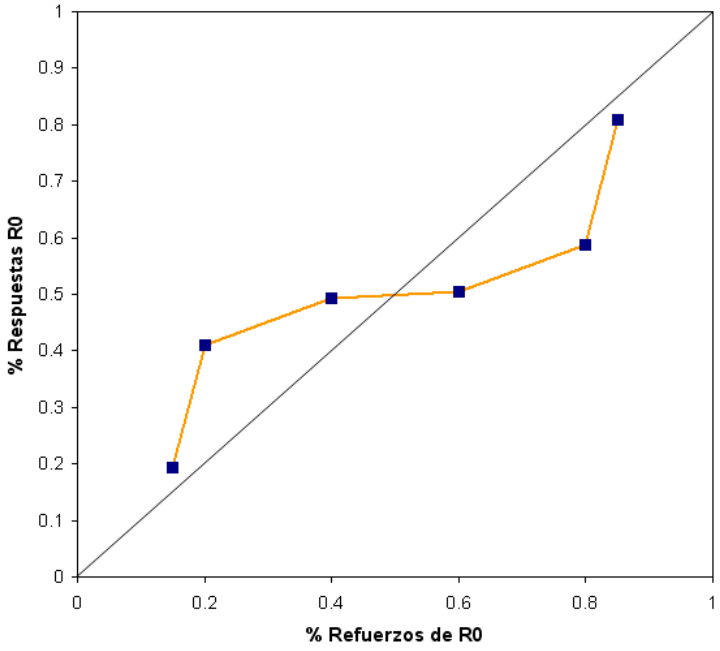

Figura 6.13: Ley de matching verificada en el modelo.

6.6. Experimentos sobre descuentos temporales y auto-control

En este punto se realiza un análisis sobre experimentos de elección de auto-control (self-control choice) y sobre la devaluación de refuerzos retrasados. La devaluación de refuerzos, como será analizado más adelante, cumple un rol preponderante a la hora lograr sostenidos niveles de cooperación en el DPI.

Los animales se enfrentan a decisiones que involucran conflictos de intereses entre un plazo corto o un plazo largo de tiempo. Tomar una decisión que permita obtener un beneficio inmediato pero pequeño en lugar de esperar durante un tiempo determinado por un beneficio mayor, es un fenómeno psicológico conocido como impulsividad. En un experimento típico de elección de auto-control, se debe elegir entre una respuesta que otorga un beneficio inmediato pero reducido y una respuesta que otorga un mayor beneficio, pero que se encuentre retrasado en el tiempo.

Las preferencias de un individuo entre una respuesta y otra pueden variar a lo largo del tiempo. Herrnstein y Mazur, argumentan que esta tendencia al cambio de preferencias es una de las evidencias en contra de la teoría de la optimización, puesto que si se siguiese una estrategia que optimiza su satisfacción a largo plazo, se elegiría una respuesta consistente entre una alternativa o la otra [24]. Sin embargo, hay diversos estudios experimentales que han demostrado que no resulta así. Los cambios de las preferencias en situaciones elección de auto-control son comprensibles si se considera cómo varía la efectividad de los refuerzos frente a los cambios en el retraso de los mismos. Esta teoría se conoce como teoría de Ainslie-Rachlin. La teoría asume, en primer lugar, que el valor de un refuerzo disminuye a medida que se incrementa el retardo entre ejecutar una respuesta y recibir dicho refuerzo. Por otro lado, también asume que un sujeto va a elegir cualquier refuerzo que tenga mayor valor al momento de realizar la elección.

Hay diversos estudios de psicología experimental de elección de auto-control. Incluso se realizan experimentos en humanos con el fin de estudiar los comportamientos económicos. Pero a diferencia de estos estudios, en el caso de experimentos con animales un retardo de segundos puede llevar a producir la diferencia entre un comportamiento impulsivo y uno de auto-control [24].

En un estudio realizado por Green, Fisher, Perlow y Sherman [12], se mostró en palomas los efectos de la impulsividad. Las palomas tienen 2 llaves como opciones. Una es elegir picar la llave que les otorga 2 segundos de alimento (retrasado 2 segundos),

y otra llave que les otorga 6 segundos de alimento (retrasado 6 segundos). Los ensayos se ejecutan cada 40 segundos independientemente de la opción tomada, y el resultado es que luego de sucesivos ensayos, las palomas se quedan con la preferencia de menor beneficio a largo plazo, eligiendo consistentemente la opción de los 2 segundos. Eligiendo esa opción, las aves pierden cerca de dos tercios de su potencial beneficio. Bajo otras condiciones, agregando un retraso de 18 segundos a los retrasos de ambas opciones, se logra revertir esta situación y las aves toman la opción de los 6 segundos de alimento durante el 80 % de los ensayos.

A continuación se realiza un análisis del retraso de refuerzos con el fin de verificar si se produce un descuento temporal, es decir, si el hecho de retrasarlos provoca una devaluación de los mismos tal como lo indica la evidencia experimental.

6.6.1. Simulación

La simulación consiste en analizar el comportamiento del modelo para diferentes esquemas de retrasos en los refuerzos. Se espera obtener, por medio de los resultados, una noción de cómo se comporta el modelo frente dichos retrasos, con el fin de comprobar si efectivamente se produce la devaluación de los refuerzos retrasados en el tiempo, tal como ocurre con los animales.

Para ello, se construye una grilla para cada par de retraso de la forma (*retraso del refuerzo inmediato*, *retraso de refuerzo retardado*). Se realiza la simulación de 1000 experimentos para cada par de retraso donde cada experimento consiste de 400 ensayos para cada par de retraso.

Como entrada se recibe CS=[1,0,0]. El US asociado a R_0 es de 4 pasos y el asociado a R_2 es de 12 pasos.

6.6.2. Resultado

En la figura 6.14 se presenta una grilla donde se muestra la probabilidad de elegir el refuerzo retardado para los distintos pares de retraso.

En el eje vertical se encuentran los retrasos del refuerzo menor pero inmediato, y en el horizontal se encuentran los retrasos de los refuerzos mayores retardados. Si se considera cada fila, a medida que se va aumentando el valor del retraso en la entrega del refuerzo, la probabilidad de elegir el refuerzo mayor decae. Pese a que el individuo pierde dos tercios de su potencial beneficio, muestra claramente los efectos de la impulsividad

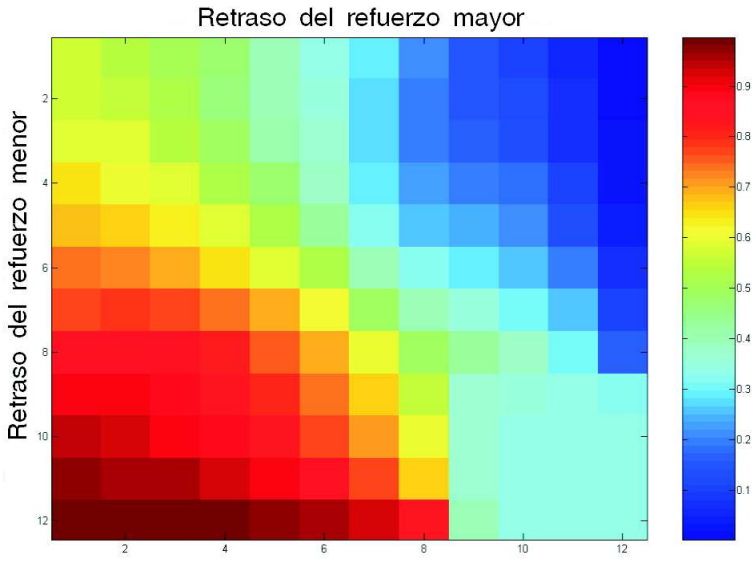

Figura 6.14: Probabilidad de elección de refuerzo retardado para pares (*retraso de refuerzo inmediato*, *retraso de refuerzo retardado*).

en la decisión tomada.

En la zona celeste (entre 10 y 12) no hay preferencia y la elección de la respuesta ejecutada se realiza al azar. Esto se debe a que el condicionamiento (como se explica en el capítulo 2) se realiza por medio de la asociación de eventos. Si los eventos se encuentran lo sucientemente distantes, existe una mayor dicultad para producir una asociación de los mismos. Como consecuencia, la posibilidad de lograr el condicionamiento se vuelve menor. Por este motivo, al no existir tales asociaciones en esa región, el modelo responde al azar.

6.7. El efecto del agrupamiento de ensayos y la acumulación

En este experimento de elección, se considera que es posible acumular los beneficios de cada opción durante una secuencia de ensayos, y que resultan accesibles solo luego de un cierto número de los mismos. La hipótesis del trabajo de Stephens [36] es que si la impulsividad ocurre debido a que las recompensas inmediatas están más valuadas que las retrasadas, entonces la acumulación de la recompensa debería reducir o eliminar la impulsividad debido a que le resta valor a las mismas.

Entonces, el objetivo de este experimento es observar la influencia de la acumulación de alimento para reducir los niveles de impulsividad de los individuos sometidos a un experimento de elección de auto-control.

Stepehns utiliza para sus estudios urracas azules (Cyanocitta cristata) en un aparato que denomina V, que se representa esquemáticamente en la figura 6.15. Este aparato está conformado por dos compartimientos en forma de V unidos por medio de una pared transparente. La expendedora de pellet (el pellet es una unidad de alimento) cuenta con un acumulador transparente y una tapa activada mediante un mecanismo eléctrico. En el caso de que la tapa permanezca abierta, los pellets caen directamente al comedero. Si la tapa permanece cerrada, los pellets quedan contenidos en el acumulador transparente, de manera que los mismos resultan visibles pero no accesibles. El aparato cuenta con tres perchas en cada compartimiento V, cada una de ellas debajo de una luz.

Inicialmente, cada pájaro es entrenado para situarse en la percha más próxima a cada luz mediante la técnica de entrenamiento de "shaping" o aproximación sucesiva [24]. El experimento comienza con el animal en la percha que se encuentra en el vértice de la V. Para inducir a que los pájaros vayan a cada posición, solo se prende la luz asociada a cada percha. Luego, una vez que el pájaro se encuentra en la posición inicial, se apagan todas las luces y se enciende una de las dos perchas delanteras. Dado que el pájaro está entrenado, éste viaja hacia la percha que está debajo de la luz encendida, y mediante un microswitch activa la expendedora de comida que deja caer cierta cantidad de pellets al comedero. La luz permanece encendida hasta que luego de un tiempo (en el cual el pájaro consume el alimento entregado) se apaga y se vuelve a encender la luz de la percha de la posición inicial para volver a repetir el procedimiento. Durante este proceso, el sujeto es entrenado a ir a ambas perchas delanteras.

Finalizado este entrenamiento inicial comienza en sí la parte principal del experimento, cuyos ensayos tienen la siguiente secuencia de eventos:

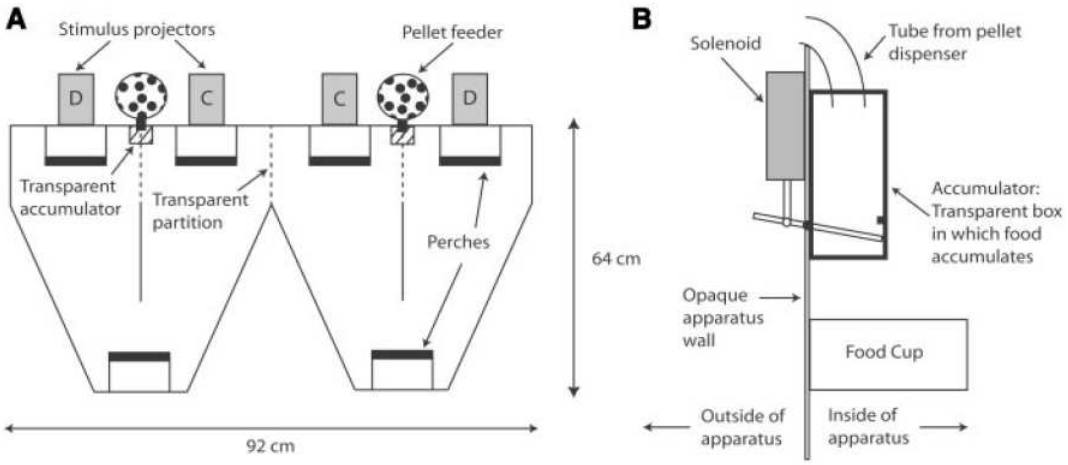

Figura 6.15: (A) Vista superior del aparato. El aparato consiste en dos compartimientos solidarios, cada uno de los cuales tiene la forma de una V. Cada compartimiento cuenta con tres perchas, equipadas con un microswitch para realizar el control del experimento. Cada una de las perchas se encuentra debajo de una luz . La percha del vértice de la V define la posición inicial para los experimentos. (B) El acumulador. La caja plástica transparente, que se encuentra al centro y al frente de cada compartimiento, recibe la comida de la expendedora de pellets. El fondo de dicha caja tiene una tapa que dependiendo del esquema de acumulación deja o no caer directamente el alimento en el comedero. En el caso de los tratamientos acumulados, el sujeto solo puede ver la comida pero no consumirla hasta que la tapa inferior del acumulador se abra. [Fuente: Stephens 2006]

1. El sujeto espera en la percha inicial durante un tiempo fijo que es el intervalo inter ensayo (ITI).

2. Ambas luces de las perchas delanteras se encienden simultáneamente, dando lugar a las dos opciones. Una opción otorga una pequeña cantidad de alimento de manera inmediata, y la otra otorga una mayor cantidad de alimento pero retrasada (las dos opciones ocupan un lugar fijo a lo largo de todo el experimento).

3. El sujeto elige una opción al situarse en una de las dos perchas. La luz de la percha que no fue elegida se apaga y comienza el esquema de retrasos de alimento programado.

4. Cuando se cumple el tiempo de retraso en la entrega de alimento, se lo deja caer y el ensayo comienza nuevamente desde el paso 1.

El programa de los retrasos puede ser acumulado o no. En el caso acumulado, el alimento permanece visible en el acumulador, pero no accesible hasta que se finalizan 4 ensayos. En el caso de los no acumulados, la tapa del acumulador permanece abierta y permite el acceso directo al alimento luego que la expendedora lo proporciona.

Para hacer más evidente el efecto de la acumulación, se realiza el agrupamiento de un número de ensayos sucesivos de manera que el tiempo entre ensayos pertenecientes a un mismo grupo (ITI intra-ensayo del grupo) sea pequeño. Luego se define un intervalo de tiempo grande entre los grupos de ensayos (ITI intra-grupo). Este hecho facilita el reconocimiento del comienzo de un grupo de ensayos. Si se definiese un ITI uniforme, posiblemente se pierda la distinción entre un conjunto de ensayos acumulado y otro.

El resultado del experimento es que efectivamente la acumulación tiene un efecto positivo en la reducción de la impulsividad mediante el agrupamiento de ensayos.

6.7.1. Simulación

El CS_0 representa la luz asociada a la opción inmediata, que es la R_0, y el CS_2 representa la luz asociada a la opción retardada, que es R_2. R_1 es una respuesta equivocada, que (no otorga ningún refuerzo).

Se realizó la simulación de 500 experimentos para cada caso: acumulado/no acumulado. Cada experimento consiste en 100 ensayos de aproximación sucesiva, que permite extinguir a la respuesta equivocada (R_1), y 400 ensayos de la parte principal del mismo.

Los ensayos presentan la acumulación del alimento usando al CS_{VE} que modela al hecho de ver al alimento que resulta inaccesible durante los 4 ensayos sucesivos. El ITI intra-ensayo es de 5 pasos entre los sucesivos ensayos del grupo, y el ITI entre los grupos de ensayos es de 345 steps. Estos valores fueron obtenidos de los materiales y métodos del trabajo de Stephens [36].

Además, los valores de US se definen 6 iteraciones para el refuerzo inmediato y 18 iteraciones para el refuerzo retardado. Este valor representa al US asociado como 1 y 3 pellets, respectivamente. La demora del refuerzo inmediato se define como 5 pasos, y la demora del retardado se define como 45 pasos.

6.7.2. Resultados

El resultado obtenido se muestra en la figura 6.16. En dicha figura se exhibe la probabilidad de elegir la respuesta asociada al refuerzo retardado bajo el esquema de ensayos agrupados.

En la figura 6.16(a) se observa que hay una clara preferencia por el refuerzo inmediato según los resultados experimentales obtenidos. El hecho de hacer visible al alimento permitió realizar una reducción de los descuentos temporales en la venida del refuerzo [36]. Por otra parte, la figura 6.16(b) muestra el efecto análogo. Si bien, los valores obtenidos son más elevados que los obtenidos a nivel experimental, cualitativamente se corresponden con los primeros dado que existe una tendencia hacia un mayor nivel de impulsividad para el caso no acumulado que para el acumulado. La acumulación resulta ser es un mecanismo válido para poder reducir los descuentos temporales en los refuerzos.

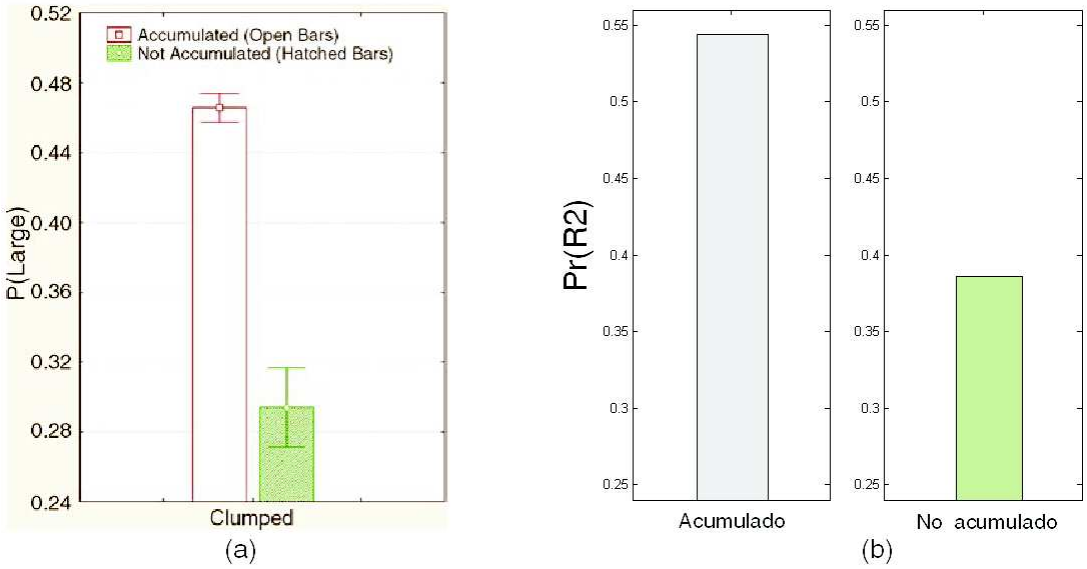

Figura 6.16: (a) Se observa los efectos de la acumulación. En el caso no acumulado se ve una clara preferencia para la opción pequeña-inmediata [Fuente: Stephens 2006]. (b) Los efectos de la no acumulación muestran una tendencia hacia la respuesta de menor refuerzo pero inmediata, mientras que se produce el efecto contrario para el caso acumulado.

6.8. El dilema del prisionero iterado

Como se mencionó en el capítulo 3, la reciprocidad en el DPI permite obtener un beneficio mayor a largo plazo, mientras que la deserción de un individuo egoísta incrementa su beneficio solo a corto plazo si el contrincante tiene una estrategia que no sea la estrategia trivial ALLC, de cooperar siempre.

En el experimento anterior de Stephens, mediante el mecanismo de permitir acumular el alimento de manera que éste sea visible pero no accesible, se pudo apreciar que es posible controlar en las urracas azules el grado de preferencia de la recompensa menor inmediata por sobre la mayor retrasada.

En el presente experimento, la clave para lograr establecer niveles sostenidos de cooperación es reconocer que el beneficio de cooperar se encuentra retrasado con respecto al beneficio de la deserción. Entonces el descuento temporal puede hacer la diferencia entre cooperar de manera sostenida y desertar. Este principio no es difícil de entender desde un punto de vista teórico. Lo notable del trabajo de Stephens [36] es que para lograr un menor impacto de los descuentos temporales experimentalmente, las urracas tienen una menor preferencia por la respuesta inmediata si pueden ir acumulando sus beneficios retrasados.

Si bien este efecto produjo un efecto positivo en los niveles de cooperación, aún cuando los descuentos temporales resultaban altos, algunos animales alcanzaron niveles altos de cooperación mientras que otros no lo hicieron. En consecuencia, existe una variabilidad considerable en sus resultados que pone en evidencia que aún existen otros factores que determinan que pueda evolucionar la cooperación en el DPI que no fueron considerados.

El experimento consiste en situar a dos urracas en cada compartimiento del aparato V de la figura 6.15. Nuevamente, los animales son entrenados mediante aproximación sucesiva para llevarlos a las condiciones iniciales del experimento anterior.

El análisis de la cooperación o de la deserción sólo se hará con el individuo situado en el compartimiento de la izquierda, dado que el otro (el compañero) seguirá una estrategia fija dispuesta por un programa que encendiendo la luz conveniente en cada caso va a inducir la estrategia del mismo. La misma será TFT o ALLD (en 2.8 se mencionó el hecho de que ambas estrategias resultan evolucionariamente estables).

De ahora en más, las posiciones de las perchas definen la opción de los individuos y

Suj. \ Comp.	C	D
C	4	2
D	0	0

Figura 6.17: Matriz de pagos del tratamiento de mutualismo para el Sujeto a estudiar (Suj.) y el compañero (Comp)

en consecuencia la paga de sus recompensas. Las perchas del centro del aparato V tendrán asociadas la acción de cooperar, mientras que la de los extremos tendrá la acción de desertar.

Para probar la estabilidad de la cooperación, los individuos son sometidos a un tratamiento de mutualismo antes de comenzar con los experimentos del DP. La matriz de pagos es la que se presenta en la figura 6.17.

Los sujetos son expuestos a esa situación hasta que se logra mantener los niveles de cooperación (es decir, los sujetos del compartimiento de la izquierda eligen la opción de la percha de la derecha) con un criterio del 80 % de las veces. Durante este tratamiento, la opción de desertar no otorga ninguna recompensa.

Una vez superada esta etapa, comienza el experimento del DPI. Para el compañero la estrategia que está siguiendo (sea TFT o ALLD) resulta transparente, dado que solo se limita a seguir la secuencia de encendido de luces indicado por un programa. En el caso de ALLD, en cada ensayo solo se enciende la luz asociada a la percha de desertar, y en el caso de seguir la estrategia TFT solo se enciende la luz que corresponde a la decisión de cooperar o desertar que realizó en el ensayo anterior el sujeto estudiado (por lo definido en la sección 3.2, el compañero comienza cooperando). Para evitar que la decisión del compañero se vea influenciada por la estrategia asignada, siempre se le entrega la misma cantidad de pellets al ejecutar una respuesta.

En cuanto a un ensayo para el sujeto estudiado, la secuencia de eventos es la misma que para el experimento anterior. Ambas luces se encienden, y dependiendo de la opción elegida, su acción será la de cooperar o desertar. Ahora la matriz de pagos es la del DP y está representada en la figura 6.18.

Las aves son sometidas a 1000 ensayos del DP. Los ensayos se encuentran agrupados y considerados los tratamientos de acumulación tal como en el experimento anterior. El resultado de los niveles de cooperación obtenidos para el sujeto estudiado para este

Suj. \ Comp.	C	D
C	4	0
D	6	2

Figura 6.18: Matriz de pagos de DP para el Sujeto a estudiar (Suj.) y el compañero (Comp)

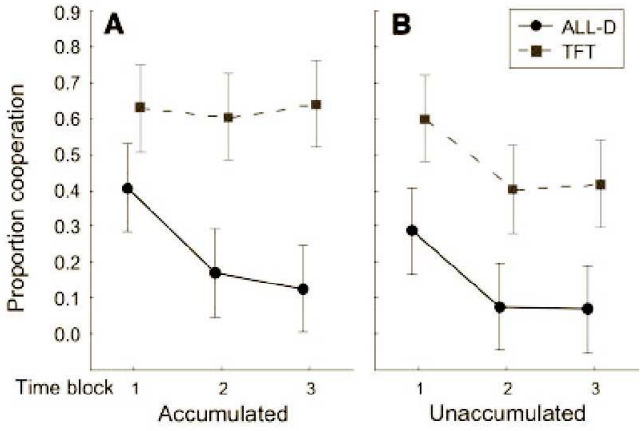

Figura 6.19: Estabilidad de la cooperación en cada uno de los diferentes tratamientos. El eje y muestra la frecuencia relativa de la respuesta de cooperar. El eje x divide los ensayos totales en tercios de 333 trials. (A) Dato para el caso acumulado (descuento reducido). (B) Caso no acumulado. Las lineas punteadas y continuas muestran las estrategias de TFT y ALLD respecivamente. [Fuente: Stephens 2002]

experimento se resume en la figura 6.19.

Cuando el oponente toma la estrategia ALLD, los niveles de cooperación disminuyen notablemente sin importar cuál sea el tratamiento de acumulación. Cuando hay reciprocidad con el compañero (en el caso TFT), se observan niveles de cooperación más elevados. Sin embargo, existe una diferencia notable para el tratamiento de acumulado y no acumulado. Para el caso acumulado, donde en teoría existe una reducción de los efectos del descuento, los niveles de cooperación resultan altos y estables. En cambio, para el caso de los no acumulados, los niveles declinan con un patrón análogo al de los de la estrategia ALLD. De esta forma, la acumulación muestra ser un mecanismo capaz de lograr estabilidad en los niveles de cooperación en un paradigma que experimentalmente

ha mostrado una tendencia a la fragilidad de la misma.

6.8.1. Simulación

Los valores de ITI son los mismos que en el caso anterior.

Ahora El CS_0 representa la luz asociada a la acción de desertar, que es la R_0, y el CS_2 representa la luz asociada a la acción de cooperar, que es la R_2. Al igual que antes, se activa el CS_{VE} para realimentar al modelo acerca de la presencia de la acumulación.

Inicialmente se realiza la aproximación sucesiva inicial que permite extinguir a R_1, que es la respuesta equivocada. El paso siguiente es la ejecución del tratamiento base con el mismo criterio de Stephens. Finalmente se ejecutan los 1000 ensayos del DPI.

Las matrices de pago son respetadas para definir el valor de US asociado. Para cada valor definido en la matriz se lo multiplica por un factor 6, que al igual que en el caso anterior, representa al valor del US asociado un pellet.

6.8.2. Resultados

El resultado obtenido por medio de las simulaciones se expresa en la figura 6.20

Cualitativamente se corresponde uno a uno con los resultados obtenidos experimentalmente. El efecto de la acumulación modelado por medio del CS_{VE} propuesto, permite lograr niveles estables de cooperación para el caso TFT Acumulado. Por otra parte, los demás casos tienen una evidente tendencia a la deserción (en la figura 6.21 se muestran los resultados obtenidos superpuestos con los resultados experimentales de Stephens).

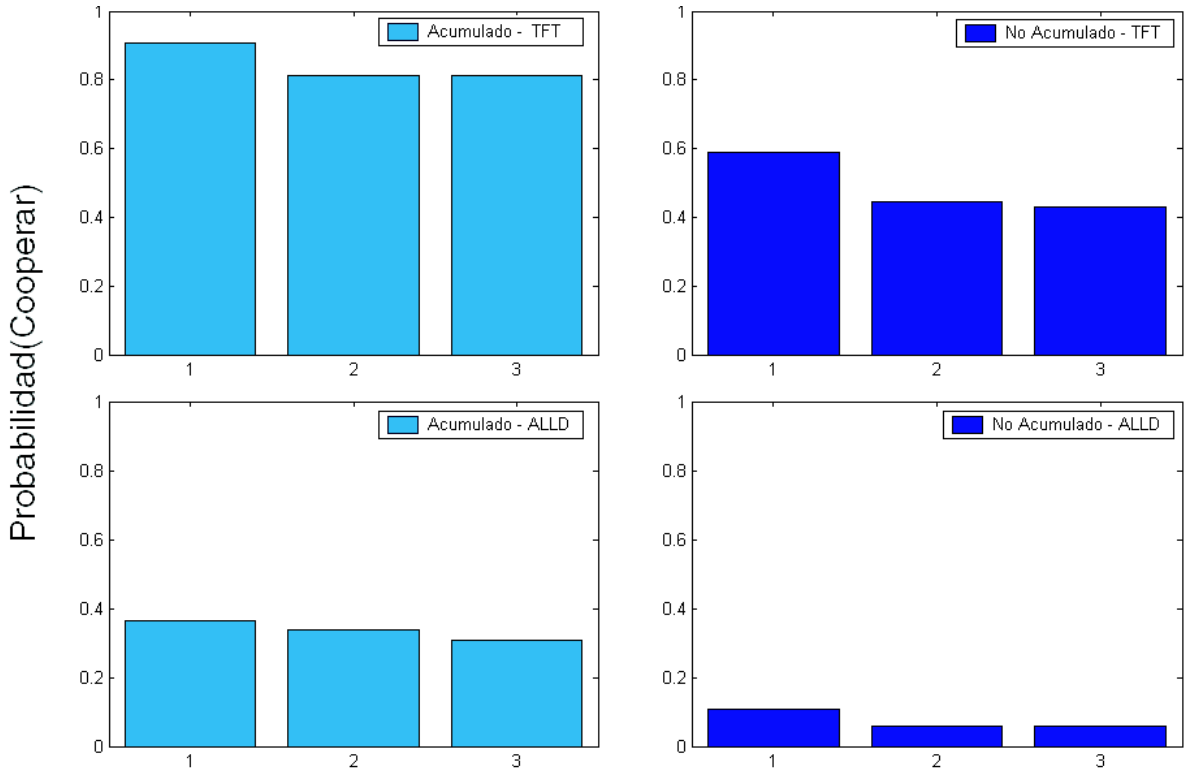

Figura 6.20: Resultados obtenidos por el modelo para los distintos tratamientos de acumulación y las estrategias TFT y ALLD. El eje y muestra la frecuencia relativa de la respuesta de cooperar. El eje x divide los ensayos totales en tercios de 333 trials.

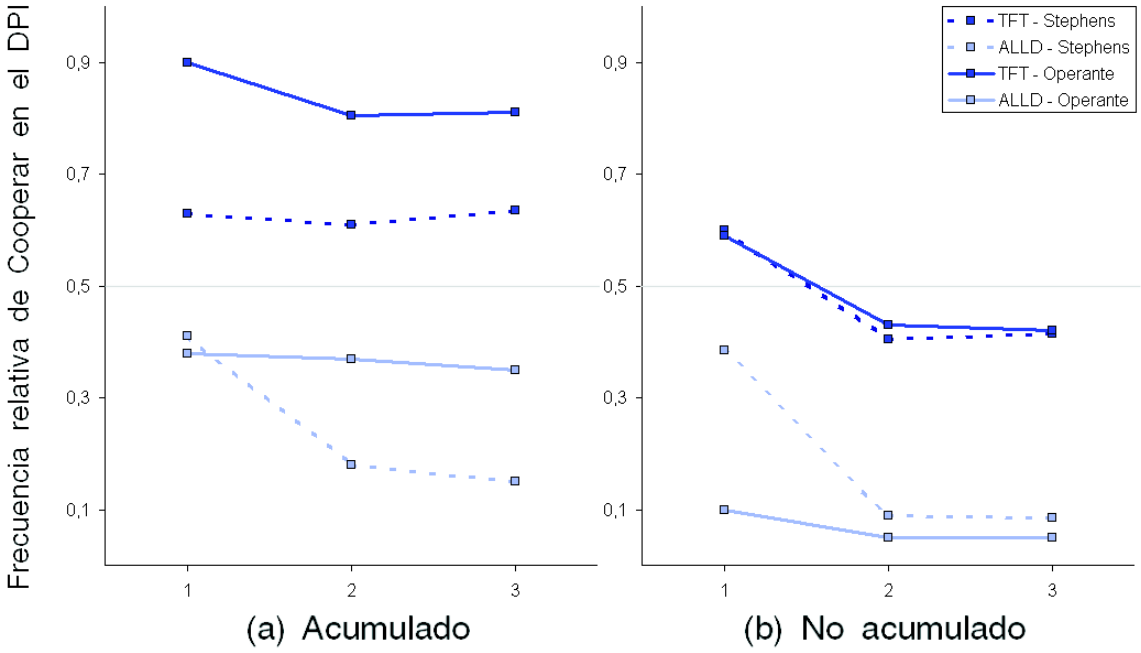

Figura 6.21: Resumen de las figuras 6.19 y 6.20. (a) Caso acumulado (b) Caso no acumulado.

Capítulo 7

Discusión

En este trabajo se realiza un estudio del rol del condicionamiento operante en el altruísmo recíproco.

Existen distintos paradigmas para estudiar cooperación, pero el dilema del prisionero iterado ha sido el prototipo de problema para analizar este tipo de conductas.

El paradigma del dilema del prisionero ha sido abordado desde diversas disciplinas como la teoría de juegos, la economía, la psicología. En particular, en el presente trabajo interesa analizarlo desde un punto de vista biológico.

Una estrategia posible es aquella propuesta por Gutnisky y Zanutto [14], quienes utilizan un modelo de condicionamiento operante, mostrando que dicho modelo resulta ser una estrategia robusta frente a diversas estrategias de sus oponentes. Sin embargo, el enfoque de dicho estudio es distinto al del presente trabajo, donde en el último se realiza el análisis de ciertos aspectos que hasta el momento no han sido tenidos en cuenta.

En este trabajo se toman como punto de partida los resultados biológicos actuales, con el objetivo de obtener un modelo de cooperación que se comporte de manera más realista de acuerdo a los resultados experimentales existentes. Se propone una teoría computacional para el dilema del prisionero iterado, bajo los efectos del retraso y la acumulación utilizando la evidencia experimental provista por Stephens [36]. Dicho resultado experimental es sumamente importante, dado que es el primer paso en la búsqueda de mecanismos que hagan posible la cooperación entre animales, que es una conducta

presente en diversas situaciones en condiciones naturales, pero que es difícil de lograr en estudios de laboratorio.

La teoría computacional propuesta está basada en modelos previos de condicionamiento operante [20]. La respuesta a ejecutar se elige de forma estocástica, tal como lo realizan los modelos biofísicos actuales [34]. Se involucra el efecto de ver al alimento como un estímulo condicionado (CS_{VE}) que varía su saliencia en función de la magnitud del refuerzo observado.

El modelo resulta capaz de verificar los resultados experimentales analizados en [36] [18] y [20]. A su vez predice que el hecho de ver el alimento, aunque resulte momentáneamente inaccesible, genera una predicción en su entrega (en este sentido, ver el alimento es interpretado como un CS).

El modelo propone un marco de trabajo computacional para el análisis de la cooperación entre individuos que incluye resultados experimentales existentes, la devaluación de refuerzo por retardos y además resulta una potencial herramienta para el diseño de nuevos experimentos entre individuos y a nivel poblacional.

7.1. Trabajo a futuro

Como trabajo a futuro,

- Sería interesante comprobar experimentalmente la predicción del modelo en cuanto al hecho de ver el alimento, reemplazando la acumulación de alimento con otro CS.
- No fue necesario considerar, para este experimento, la presencia del animal oponente para que el modelo pueda predecir los resultados obtenidos por Stephens [36]. Sin embargo, sería interesante analizar las predicciones del modelo con respecto a ver al oponente y testear experimentalmente el hecho de no poder ver al oponente.
- Para ciertos rangos de refuerzos, el modelo predice los efectos de la devaluación por retrasos en la entrega de los mismos, y en consecuencia, los niveles de cooperación observados en animales bajo el DPI. Sin embargo, sería conveniente hacer una exploración exhaustiva del espacio de parámetros a fin de incluir todos los resultados experimentales conocidos para la Ley de matching generalizada, es decir, variando el valor del refuerzo, el tiempo y la probabilidad de entrega.

- El estudio se realizó simulando pares de individuos sometidos al DPI, sería importante estudiar las predicciones que arroja el modelo en poblaciones de individuos y compararlos con los escasos datos experimentales actuales [5].

- También sería importante incluir en estos trabajos, poblaciones de neuronas de integración y disparo que reemplacen a las neuronas únicas que representan a estos grupos del modelo presentado, siguiendo las estrategias de los modelos biofísicos existentes [34]. Estos últimos dos puntos propuestos resultan contrapuestos dado que el costo computacional requerido por el segundo se elevaría exponencialmente para realizar un análisis poblacional de individuos. Es importante estudiar diferentes estrategias para el abordaje computacional de este problema. Una posible estrategia podría ser trabajar sobre distintos niveles de complejidad: simulando modelos biofísicos que den pautas sobre el comportamiento macroscópico a modelos simplificados que interactuarían a nivel pòblacional. A su vez las variables macroscópicas observadas en estos modelos simplificados, deberían ser predichas constantemente por el modelo biofísico.

Capítulo 8

Conclusión

En el presente trabajo se ha abordado el estudio del rol del condicionamiento operante con respecto a la cooperación en el dilema del prisionero iterado.

Para ello se propuso una teoría basada en las siguientes hipótesis:

- la existencia de una memoria de capacidad temporal limitada de los estímulos y las respuestas.
- la capacidad de generar una predicción refuerzos futuros.
- la asociación entre estímulos y dicha predicción según la regla de Rescorla-Wagner.
- la asociación hebbiana/anti-hebbiana entre estímulos y respuestas en función al nivel de predicción.

Dichas hipótesis han sido sustentadas en evidencias conductuales y neurofisiológicas. Además de estas hipótesis, se propuso un mecanismo de elección de respuestas a ejecutar utilizando una función probabilística, tal como lo realizan los modelos biofísicos actuales [34].

El modelo propuesto predice algunos de los principales experimentos de la psicología experimental tales como:

- Adquisición.

- Extinción.
- Selección de respuesta.
- Reversión.
- Discriminación.
- Reversión de discriminación.
- Ley de matching (*Matching law*).

Por otra parte, el modelo predice los resultados acerca de la devaluación de refuerzos y los efectos que produce la acumulación de alimento, que permite reducir dicha devaluación.

Finalmente el modelo predice los resultados experimentales de Stephens acerca del aumento en los niveles de cooperación utilizando dicho mecanismo de acumulación. Este resultado permite establecer que dicho recurso propuesto por Stephens [36] (que es la primer alternativa propuesta que posibilita obtener estabilidad en niveles de cooperación entre animales en estudios de laboratorio) puede ser explicado mediante una teoría de condicionamiento operante. De esta manera se demuestra que el condicionamiento operante, bajo estas condiciones, resulta ser un mecanismo posible para lograr el aprendizaje de conductas de cooperación.

Apéndice A

Parámetros de las simulaciones

Cantidad de iteraciones totales de un ensayo básico= 60.[1]

Límite inferior de iteraciones para activar una respuesta en un ensayo = 4.
Límite superior de pasos para activar una respuesta en un ensayo = 9.
Tiempo de ejecución de una respuesta en un ensayo = 5.
Factor de reducción del CS_{VE} = 15.

Umbral de disparo (μ) = 0,35.
Factor del ruido (ρ) = 0,015.
Caida de exploración (ω) = 0,99.

Trazas:

$\varepsilon = 0{,}25$.
$\beta = 0{,}01$.
$\alpha_{DOWN} = 0{,}95$.
$\alpha_{UP} = 0{,}1$.

Respuestas:

$\gamma = 0{,}2$.
$\eta_i = 0{,}01$.
$\eta_i = 0{,}002$.
$\lambda = 0{,}6$.

[1]En el caso de no aclararse en la simulación, ésta será la longitud de un ensayo

$\phi = 0{,}025$.
$\psi = 0{,}998$.

Predicción:

$V_{US} = 0{,}1$.
$\sigma = 1$.
$\upsilon = 3$.

Bibliografía

[1] G.W. Ainslie, *Impulse control in pigeons*, Journal of the Experimental Analysis of Behavior **21** (1974), 485–9.

[2] R. Axelrod and W.D. Hamilton, *The evolution of cooperation*, Science **211** (1981), 1390–6.

[3] J.F. Bates and P.S. Goldman-Rakic, *Prefrontal connections of medial motor areas in the rhesus monkey*, Journal of Comparative Neurolobiology **336** (1993), 211–28.

[4] R. Bshary and A.S. Grutter, *Image scoring and cooperation in a cleaner fish mutualism*, Nature **441** (2006), 975–8.

[5] C.F. Camerer and E. Fehr, *When does "economic man" dominate social behavior?*, Science **311** (2006), 47–52.

[6] C. Catania, *Concurrent performances: Reinforcement interaction and response independece*, Journal of the Experimental Analysis of Behavior **6(2)** (1963), 253–63.

[7] K.C. Clements and D.W. Stephens, *Testing models of non-kin cooperation: Mutualism and the prisoner's dilemma*, Animal Behaviour **50** (1995), 527–49.

[8] M. Flood, K. Lendenmann, and A. Rapoport, *2 x 2 games played by rats: Different delays of reinforcement as payoffs*, Behavioral Science **28** (1983), 65–78.

[9] J.M. Fuster and J.P. Jervey, *Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task*, Journal of Neuroscience **2(3)** (1982), 361–75.

[10] R.M. Gardner, T.L. Corbin, J.S. Beltramo, and G.S. Nickell, *The prisoner's dilemma game and cooperation in the rat*, Psychological Reports **55** (1984), 687–96.

[11] F. Gonon, *Prolonged and extrasynaptic excitatory action of dopamine mediated by d1 receptors in the rat striatum in vivo*, Journal of Neuroscience **17(15)** (1997), 5972–8.

[12] L. Green, E.B. Fisher, S. Perlow, and L. ShermanMazur, *Preference reversal and self control: Choice as a function of reward amount and delay*, Behavior Analysis Letters **1** (1981), 43–51.

[13] L. Green, P.C. Price, and M.E. Hamburger, *Prisoner's dilemma and the pigeon: Control by immediate consequences*, Journal of the Experimental Analysis of Behavior **64** (1995), 1–17.

[14] D.A. Gutnisky and B.S. Zanutto, *Cooperation in the iterated prisoner's dilemma is learned by operant conditioning mechanisms*, Artificial Life **10(4)** (2004), 433–61.

[15] _____, *Learning obstacle ovoidance with an operant behavior model*, Artificial Life **10(1)** (2004), 65–81.

[16] W.D. Hamilton, *The genetical evolution of social behavior*, Journal of Theoretical Biology **7** (1964), 1–16.

[17] S. Haykin, *Neural networks: A comprehensive foundation*, Prentice-Hall, International Edition, 1998.

[18] R.J. Herrnstein., *Relative and absolute strength of response as a function of frequency of reinforcement*, J Exp Anal Behav. **4** (1961), 267–72.

[19] J. Hertz, A. Krogh, and R.G. Palmer, *Introduction to the theory of neural computation*, Addision-Wesley, Reading, 1991.

[20] S.E. Lew, C. Wedemeyer, and B.S. Zanutto, *Role of unconditioned stimulus prediction in the operant learning: a neural network model*, Proceedings of IJCNN '01, Washington DC, USA (2001), 331–6.

[21] E. Lieberman, C. Hauert, and M.A. Nowak, *Evolutionary dynamics on graphs*, Nature **433** (2005), 312–6.

[22] N.J. Mackintosh, *The psychology of animal learning*, Prentice Hall, San Diego, 1974.

[23] J.E. Mazur, *Test of an equivalence rule for fixed and variable reinforcer delays*, Journal of Experimental Psychology: Animal Behavior Processes **10** (1984), 426–36.

[24] ———, *Learning and behaviour*, Prentice Hall, 1994.

[25] M.A. Nowak and K. Sigmund, *A strategy of win-stay, lose-shift that outperforms tit for tat in prisoner's dilemma*, Nature **364** (1993), 56–8.

[26] ———, *Evolution of indirect reciprocity by image scoring*, Nature **393** (1998), 573–7.

[27] J.P. O'Doherty, *Reward representations and reward-related learning in the human brain: insights from neuroimaging*, Current Opinion in Neurobiology **14(6)** (2004), 769–76.

[28] H. Ohtsuki, C. Hauert, E. Lieberman, and M.A. Nowak, *A simple rule for the evolution of cooperation on graphs and social networks*, Nature **441** (2006), 502–5.

[29] H. Rachlin and L. Green, *Commitment, choice and self-control*, Journal of the Experimental Analysis of Behavior **17** (1972), 15–22.

[30] N. Schmajuk and J. DiCarlo, *Stimulus configuration, classical conditioning, and hippocampal function*, Psychological Review **99(2)** (1992), 268–305.

[31] N.A. Schmajuk and B.S. Zanutto, *Escape, avoidance, and imitation: a neural network approach*, Adaptive Behavior **6(1)** (1997), 63–129.

[32] W. Schultz, *Getting formal with dopamine and reward*, Neuron **36** (2002), 241–63.

[33] W. Schultz, P. Dayan, and P.R. Montague, *A neural substrate of prediction and reward*, Science **275(5306)** (1997), 1593–9.

[34] A. Soltani and X.J. Wang, *A biophysically based neural model of matching law behavior: melioration by stochastic synapses*, Journal of Neuroscience **26(14)** (2006), 3731–44.

[35] J.E.R. Staddon and R.H. Ettinger, *Learning: An introduction to the principles of adaptive behavior*, San Diego: Harcourt, Brace, Jovanovich, San Diego, 1989.

[36] D.W. Stephens, C.M. McLinn, and J.R. Stevens, *Discounting and reciprocity in an iterated prisoner's dilemma*, Science **298** (2002), 2216–8.

[37] R.S. Sutton and A.G. Barto, *Time-derivative models of pavlovian reinforcement*, Cambridge, 1990.

[38] A. Traulsen and M.A. Nowak, *Evolution of cooperation by multilevel selection*, National Academy of Sciences USA **103** (2006).

[39] R.L. Trivers, *The evolution of reciprocal altruism*, Quarterly Review of Biology **46** (1971), 35–57.

[40] P. Waelti, A. Dickinson, and W. Schultz, *Dopamine responses comply with basic assumptions of formal learning theory*, Nature **412** (2001), 43–8.