Resumen: **"Cooperation in the IPD is learned by operant conditioning mechanism"**

- resumen

El Dilema del Prisionero es la metáfora que lidera la evolución de comportamientos cooperativos.[38]

...

Mecanismos de aprendizaje como el Condicionamiento Operante permiten a animales adquirir características de su entorno en el orden que reciben, obtienen, realimentaciones (refuerzos) y evaden castigos (penalizaciones).

En este estudio se evaluó mediante el Dilema del prisionero el rol del Condicionamiento operante en el aprendizaje de cooperación.

Se encontró que el mecanismo Operante permite jugar el Juego del Dilema del Prisionero iterado y fue contrastado con otras estrategias.

- intro

Surge una paradoja dentro de la teoría evolutiva, ya que la Selección Natural no explica firmemente la cooperación entre individuos. Se esboza una posible solución, estableciendo que la selección natural se aplique tanto a familias como a individuos.

Hamilton [24] en 1964 da una posible solución a la paradoja generada por la cooperación en la teoría de la Selección Natural, the**KinSelection**, que es el sacrificio de producciones propias en colaboración con producciones de otros individuos, en aras de obtener un mayor promedio de copias genéticas de su familia.

Otras teorías que explican la evolución de comportamientos cooperativos son:

[17]

[70] trait group selection

[13] by-product mutualism

[64,65] **reciprocal altruism**

El altruismo reciproco involucra un intercambio de beneficios entre dos o más individuos y un costo asociado a cada acto altruista.

Si no hay ningún individuo que falle en la reciprocidad durante un largo tiempo, cada individuo experimentará un beneficio neto. Este mecanismo de cooperación será seleccionado si existe una protección contra la falta de reciprocidad en los actos altruistas. [5]. Utilizando teorías de juego (PD) **Trievers** fue el primero en formalizar los principales rasgos.

A description...El PD fue estudiado en diversas áreas como la biología evolutiva, la sociología, la psicología y la economía. En el juego, un jugador puede o cooperar (C) o rechazar (D) y cada jugador recibe una paga en base a la tabla.

El equilibrio cooperativo se mantiene cuando se establecen las siguientes relaciones:

**T > R > P > S**

**2xR > T + S**

Dos jugadores han adoptado un equilibrio de Nash, si cada jugador está utilizando una estrategia que es la mejor respuesta a las estrategias de los otros. Así, si el PD es jugado solo una vez el equilibrio de Nash es la Deserción (Defect), porque es el mejor pago independientemente de la elección del oponente. Ellos obtiene un 1 punto, ellos podrían haber obtenido 3 puntos si hubieran optado por la cooperación.

En el *iteratedprisioner’s dilema* (IPD), los jugadores se enfrentan a su oponente repetidas veces. Con el fin de maximizar la paga, los jugadores pueden cambiar sus movimientos en base a las estrategias de sus oponentes.

Si existe un número fijo y conocido de iteraciones entre los pares de jugadores, entonces el comportamiento *defect* seguirá siendo la única estrategia evolutivamente más estable. Pero si el número no es fijo de antemano, significará que no existe una única mejor estrategia entro los comportamientos de los oponentes. [5]

En el *evolutionaryprisioner’s dilema* (EPD, Dilema del prisionero evolutivo), los puntajes de cada jugador ha obtenido en un IPD, son utilizados para simular la evolución de una población. Cada nueva generación tiene un diferente arreglo de jugadores proporcional al total de los puntos obtenidos entre todos los adversarios. Esto lleva a que el total de los puntos para cada estrategia cambie en cada nueva generación, porque la proporción de oponentes cambia.

Triever fue el primero en relacionar el altruismo reciproco con el IPD. Axelrod y Hamilton [5] usan simulación por computadora para evaluar el desempeño de un grupo de estrategias, buscando que sean evolutivamente estables (ESSs). Ellos no solo consideraban la estabilidad final de una dada estrategia, sino también la viabilidad de cada una en un ambiente dominado por estrategias egoístas. Ellos encontraron que si la probabilidad de encontrar un patrón era por encima de un determinado umbral, luego junto a la exitosa ALLD (defectalways), otra estrategia emerge, la TFT (titfortat – toma y deja).

Alexrod organizó un torneo en donde la estrategia TFT salió ganadora entre 15 estrategias diferentes, aun luego de que el torneo sea modificado para permitir simulaciones de evolución [6]. Luego un gran número de investigadores aparecieron con interesantes modificaciones en el juego para hacerlo más realista, como ser la posibilidad de realizar movimientos erróneos [10]. Esto generó que la estabilidad de ciertas estrategias cambien considerablemente, por ejemplo se evita la siguiente situación, cuando dos jugadores utilizan TFT y uno cambia de C a D, el oponente lo copia y ellos quedaban alternando entre CD y DC, reduciendo significativamente la performance de la estrategia.

El IPD ha sido el marco para el estudio del potencial de la cooperación y se observa que es muy difícil observar una sustancial reciprocidad en los experimentos de cooperación con animales. [12].

Recientemente, Stephens y co. [61] propone que una posible explicación para esta fragilidad de cooperación en el IPD es un descuento temporal constante, dado que los animales prefieren recibir pequeños estímulos inmediatos en vez de esporádicos y fuertes estímulos. [2, 35, 36, 42]

EL IPD ha sido estudiado por màs de 20 años y los investigadores se han focalizado en estrategias que no contemplan el aprendizaje de comportamiento. Macy y Flache [24] han utilizado modelos basados en agentes para estudiar la dinámica con que una población se mueve de una estado de equilibrio a otro. Encontrando, a diferencia de los supuestos convencionales de cálculos a futuro, que el modelo usado explora retrospectivamente las alternativas basadas en adaptación evolutiva y aprendizaje. La teoría analítica de juegos asume que los jugadores tiene suficientes habilidades cognitivas para realizar predicciones exactas de las consecuencias de sus diferentes decisiones. La teoría de aprendizaje aclara este requisito permitiendo a los jugadores basar sus predicciones en experiencias pasadas más que en deducciones lógicas.

Sandholm and Crites [45] evaluaron como un conocido algoritmo de aprendizaje por refuerzo, Q-Learning, no puede aprender a jugar el IPD. Esto indicó que Q-learning no puede aprender a cooperar con otros Q-learning como adversarios.

Arita and Suzuki [1] analizaron la interacción entre aprendizaje y evolución al incorporar el “Baldwin effect” en al IPD. Mostrando que se obtiene una estrategia Estacionariamente estable y da señales de cómo el aprendizaje puede alterar el curso de la evolución en sistemas dinámicos, donde ni el sistema de aprendizaje ni el mecanismo de evolución tiene soporte biológico.

Una propuesta interesante sería el desarrollo de un modelo de evolución y aprendizaje con hipótesis biológicamente plausibles, con el fin de analizar su interacción.

Aunque existen muchos ejemplos de cooperación entre animales [17], los mecanismos de aprendizaje en cooperación no han sido suficientemente analizados. Sin embargo, el mecanismo de aprendizaje que se ha observado en muchas especies (“Condicionamiento operante”) puede que explique al altruismo reciproco. Por ello, es que debe ser estudiado. Al condicionamiento operante a aparecido tempranamente en la historio de la evolución. Aquí se propone que el aprendizaje de cooperación puede haber aparecido en los animales como un mecanismo para garantizar la alta proporción de estímulos apetitivos.

Además de la propuesta presentada, se plantea que el aprendizaje operante puede llevar al éxito en el en el PD contra otras estrategias conocidas. La propuesta de este trabajo es explorar la posibilidad de que el altruismo reciproco sea explicado mediante un simple mecanismo de aprendizaje por refuerzo.

**Concepto general del condicionamiento operante.**

Staddon [56] “Los organismos son máquinas diseñadas por la evolución para desarrollar un rol particular. Este rol junto al ambiente en que viven es denominado Nicho del Organismo.”

En psicología se ha identificado al condicionamiento Clasico y operante como dos formas primerias de aprendizaje que le permite a un animal adquirir características relevantes de un ambiente en orden de obtener mayores recompensas y menos castigos. En condicionamiento clásico, el aprendizaje ocurre por asociación repetitiva de un estímulo condicionado (CS) con un estímulo no condicionado (US), que produce una respuesta no condicionada (UCR). Experimento de Pavlov.

El condicionamiento operante es un procedimiento experimental a lazo cerrado, en el sentido de que el estímulo recibido es supeditado al comportamiento. EL animal aprende a desarrollar una acción que conduce a obtener recompensas con mayor frecuencia.

Existen dos tipos de Estimulo no condicionado, Apetitivo y aversivo.

**Teoría del Condicionamiento Operante.**

Zanutto y Lew presentan una red neuronal de condicionamiento operante para estímulos apetitivos y aversivos.

Hay muchas teorías no matemáticas que explican el condicionamiento operante. No existe un acuerdo firme sobre la hipótesis y el rol de la predicción, especialmente en teorías que explican la evasión y el escape.

…

En teorías cognitivas de evasión [51], se asume que durante la fase de adquisición, el animal desarrolla una expectativa en base a su respuesta. Seligman y co. [39] proponen que bajo ciertas circunstancias el animal desarrolla una expectativa de que su comportamiento no causará efecto en el ambiente y dejará de responder, este efecto se denomina “LearnedHelplessness”. Teorías para estimulos apetitivos [16] o aversivos [46, 71].

**Hipótesis del modelo:**

* Experimentos sobre comportamientos sugieren que el aprendizaje está dado por el cambio de las expectativas sobre un acontecimiento futuro, principalmente recompensas o castigos. Rescorlaand Wagner [43] proponen que un animal aprende comparando las expectativas por una dada situación y lo que sucede actualmente. Staddon [56] habla de expectativa o predicción. Además, existe un sustrato neuronal de predicción y recompensa, como es la participación de las neuronas Dopaminérgicas en el área Ventral Tegmentral (VTA) y sustancia Nigra (SN) [50]. Waelti y co. [66] encontró que el disparo de estas neuronal corresponde con el comportamiento de predicción descripto por Rescorla and Wagner rule.
* El camino de conexión de los diferentes bloques que se proponen como participantes del mecanismos operante tienen sus bases en estudios anatómicos en vertebrados. En primate, la corteza prefrontal se conecta a la corteza premotora y ésta a la corteza motora primaria [8, 67]. Las neuronas VTA están conectadas a la corteza prefrontala través del sistema dopaminérgicomesocortical [41].
* En primates, la corteza orbitofrontal está relacionada con el trabajo de memorización [8, 23]. Las neuronas de esta área continúan disparando luego de que el estimulo fue presentados. Esta memoria de corto plazo se pueden asociar a CSs, acción y UCs.
* El efecto cortical de la interacción lateral en la corteza premotora y motora primaria es simulada por un grupo de neuronas que disparan de forma similar y tiene una alta correlación. También se asume que la respuesta es generada por un cluster de neuronas y este representa a una única neurona que simula la reacción en cascada del conexionado motor.
* En este modelo el aprendizaje es controlado por las neuronas VTA y SN, que están representadas por la neurona de predicción. Schultz and co. [50] encontraron que hay una neurona en la VTA y SN que responde a un error en la predicción de recompensa y luego la asociación entre CS y US es aprendida. La misma neurona dispara concurrentemente en presencia de CS, prediciendo que seguirá el US. Schultz [49] sugiera que la acción de estas señales de enseñanza pueden ser formalizadas aplicado la regla de aprendizaje de Rescorla-Wagner a los pesos sinápticos. Una señal de enseñanza dopaminérgica puede modificar los pesos de la sinapsis prefrontal de acuerdo a la regla de aprendizaje de Hebbian. En este modelo la neurona Dopamina controla el aprendizaje de la neurona prefrontal a través del aprendizaje Hebbiano o anti-Hebbiano acorde si la predicción esta por encima de un determinado umbral.