Suboptimal choice, reward predictive signals, and temporal information

Paul J. Cunningham, Timothy A. Shahan

2018

Elección subóptima es interesante desde varias perspectivas:

- Forrajeo óptimo, debido a que aleja a los animales de la maximización. Ha motivado la inclusión de perspectivas evolutivas en el estudio de la toma de decisiones maladaptativas en los animales.
- Modelos mecanísticos o descriptivos de elección, pues es inconsistente con la premisa de que los organismos concentran la conducta en las alternativas que proveen la mayor cantidad de reforzamiento (e.g., ley de igualación).
- Como análogo de la conducta de juego patológico, pues puede revelar procesos conductuales básicos involucrados en la conducta de apuestas.

Una propuesta para la explicación de la conducta subóptima está en el efecto desproporcionado que el reforzamiento condicionado parece tener en la elección. Se ha argumentado que la función de los reforzadores condicionados es informativa, pero nunca se formaliza exactamente qué se quiere decir con eso. Por ello, Shahan y Cunningham (2015) proponen una explicación que fomraliza la relación entre la información y el reforzamiento condicionado según la cual los reforzadores condicionados adquieren su habilidad para atraer y mantener la conducta dada la información temporal que proveen acerca de la comida.

Reforzamiento condicionado y elección subóptima

Uno de los hallazgos más robustos en elección subóptima es que los estímulos de los TL de la alternativa subóptima necesitan ser predictivos de las consecuencias finales para guiar la elección de las palomas hacia sí.

Aunque no hay consenso sobre cómo caracterizar el impacto de los TL, la perspectiva más aceptada es que estos estímulos funcionan como reforzadores condicionados que soportan las respuestas de elección por esa alternativa. Así, es necesario definir por qué los estíumlos de los TL de la alternativa subóptima son preferidos sobre los de la alternativa óptima.

La hipótesis SiGN

Se basa parcialmente en la hipótesis de reducción de la demora. Argumenta que los TL funcionan como reforzadores condicionados en la medida que señalan una reducción en la demora hacia la comida con respecto a la respuesta de elección. Presupone que un estímulo que no anuncia comida no tiene efecto en la elección. El TL que anuncia comida se asocia con una mayor reducción en la demora que la respuesta misma de elección, pues ésta va seguida de comida intermitentemente, mientras el TL la anuncia el 100 % de las ocasiones, por lo que se vuelve un reforzador condicionado más robusto que los TL de la alternativa óptima, que no tienen una reducción mayor que aquella de la respuesta de elección. Sin embargo, la hipótesis SiGN tiene varios problemas conceptuales.

La hipótesis del valor predictivo

Según esta hipótesis de Zentall, la elección depende solo del valor predictivo (la probabilidad y cantidad de comida señalada por los TL) de los reforzadores condicionados. Tampoco le da importancia al inhibidor. Así, el TL con la mayor probabilidad o magnitud sesga la elección hacia sí, sin importar que su complemento sea un TL que anuncia no-recompensa.

El modelo ecológico

Este modelo se desvía de los anteriores y sugiere que la elección subóptima refleja una estrategia de búsqueda de información adaptativa en su ambiente natural, pero que resulta en malas decisiones al descontextualizarla. Argumenta que, al haber evolucionado en ambientes en los que la información negativa permite abandonar búsquedas improductivas, los organismos desarrollaron mecanismos que toman en cuenta solamente las buenas noticias y no atienden las malas. Formalmente, cuando una alternativa tiene TL informativos, las demoras asociadas con los ensayos en los que no se entrega comida no son tomadas en cuenta en los cálculos de ingesta de energía de la alternativa, lo que hace que su tasa funcional de reforzamiento sea mayor.

Sin embargo, aunque se basa en gran medida en la información provista por los TL, no hay una definición formal de lo que se entiende por información. Una aproximación hecha por Hendry (1969) para explicar la conducta de observación podría ser funcional. Define cuánta información provee un estímulo discriminativo en bits mediante

$$H = \Sigma_i p_i log_2 \left(\frac{1}{p_i}\right),$$

en donde H se refiere a la incertidumbre promedio (entropía), y p_i se refiere a la probabilidad del estado i entre un conjunto de estados posibles. En el procedimiento de respuestas de observación, los estados se refieren a los períodos de reforzamiento primario (S^+) y extinción (S^-). Así, en un procedimiento en el que los ensayos con y sin comida ocurren con $p = 0.5$, una presentación de estímulo reduce la incertidumbre en un bit. En esta perspectiva, la información es reforzante en sí misma, por lo que las señales de ausencia de reforzamiento deberían poder mantener conducta por sí mismas, y la información debería tener máxima efectividad cuando la probabilidad de comida es de 0.5. Ninguna de esas predicciones se cumple, lo que indica que no es la información acerca de si un reforzador primario será entregado lo que aporta el valor de los reforzadores condicionados, como implica el modelo ecológico.

Esto no indica que una aproximación desde teoría de información sea incorrecta. Los autores sugieren que solo estaba guiada hacia la información incorrecta: no es la información sobre si la recompensa ocurrirá o no, sino la información sobre *cuándo* ocurrirá lo que define al reforzamiento condicionado.

Información temporal, condicionamiento pavloviano, y reforzamiento condicionado

La aproximación al condicionamiento pavloviano se basa en la idea de que un estímulo condicionado (EC) influye en las respuestas por medio de la información que provee acerca de *cuándo* esperar el siguiente estímulo incondicionado (EI). Para calcular la información temporal de un EC hay tres pasos: se calcula la incertidumbre promedio de cuándo esperar un EI independientemente de cualquier otro evento usando la distribución de posibles intervalos EC-EC. Para una distribución exponencial de probabilidades de intervalos, el cálculo de la entropía es

$$H_c = log_2 C + k,$$

en donde k es una constante $(e/\Delta\tau)$ que depende de la resolución temporal del animal. H_c es la incertidumbre basal sobre cuándo esperar comida en la sesión experimental, que pone un límite superior a cuánta información se puede proveer por la entrega de la comida misma. Después, la incertidumbre promedio sobre cuándo esperar un EI en presencia del EC (i.e., H_t) se calcula usando una distribución de probabilidad de posibles intervalos EC-EI,

$$H_t = log_2 t + k,$$

donde t se refiere a la duración promedio EC-EI, y k es igual que antes. H_t cuantifica la incertidumbre sobre cuándo esperar la comida en presencia de EC. Por último, la información temporal dada por EC (i.e., H) se mide por el grado en el que la presentación de EC reduce la incertidumbre sobre cuándo esperar EI:

$$H = (log_2 C + k) - (log_2 + k),$$

que se reduce a

$$H = log_2(C/t).$$

Esto formaliza la noción de que un EC influye en las respuestas mediante la información que provee. Cuando un EC señala una mayor reducción en la demora al EI relativa a la demora promedio al EI, ese EC provee mayor información, y las respuestas condicionadas se aprenden más rápido cuando EC provee más información temporal.

Si, como muchos han argumentado, la habilidad de un estímulo para funcionar como reforzador condicionado es gobernada por condicionamiento pavloviano, entonces estas ecuaciones dictan el grado en que un estímulo puede servir como reforzador condicionado.

Shahan y Cunningham (2015) aplicaron estas ecuaciones en el procedimiento de respuestas de observación y encontraron que la cantidad de información temporal proveída parece gobernar el grado en que las señales mantienen la conducta de observación.

Esta aproximación de información temporal tiene como ventaja sobre otras aproximaciones que especifica por qué los intervalos temporales son críticos para el reforzamiento condicionado: más allá de nombrar la reducción en la demora, el motivo detrás de su

importancia es que esos intervalos temporales son necesarios para hacer el cálculo de la información proveida por las señales predictoras de comida.

Esta aproximación ofrece un solo marco de referencia que unifica al apredizaje pavloviano y al reforzamiento condicionado.

Información temporal en el procedimiento de elección subóptima

Se demuestra el cálculo de la información temporal en elección subóptima, con probabilidades de reforzamiento de 0.2 y 0.5. Consideran solo un estímulo para cada alternativa, ignorando al S^-, como otros modelos, y condensando a los dos de la alternativa óptima en uno solo, pues son funcionalmente equivalentes.

Inicialmente, se debe calcular el tiempo promedio hasta la entrega de comida (*cycle time*, C), y después, el tiempo hasta la entrega de comida en presencia de un TL asociado con ella (*trial time*, t).

Cycle time

Dado un ITI de 10s , un TL de 10s, y una latencia de respuesta asumida de 1s, el tiempo entre consecuencias es de 21s. La alternativa suboptima tiene 0.2 de probabilidad de entrega de comida; y la óptima, 0.5; y las alternativas se presentan con igual frecuencia (considerando solo ensayos forzados). Así, el tiempo promedio esperado hasta la entrega de comida es

$$C = \frac{ITIs + ILs + TLs}{.5(pSr_{sub} + pSR_{opt})} = \frac{10 + 1 + 10}{.5(0.2 + 0.5)} = \frac{21}{0.35} = 60s,$$

Nótese que el tiempo pasado en presencia del S^- también se incluye en los cálculos, pues también es necesario como parámetro temporal para definir la incertidumbre basal. Sin embargo, esto no significa que S^- influya directamente en la elección. El modelo rechaza la noción de que los S^- tengan un efecto en la elección. Más bien, su tiempo de duración es relevante para el cálculo de C.

Además, C depende de las probabilidades, no de la señalización, por lo que permanecerá constante sin importar que la función predictora de los estímulos cambie.

Trial time

El tiempo a la entrega de comida en presencia de un TL asociado con ella está dado por

$$t = \frac{TLs}{pSr|TL},$$

donde $pSr|TL$ es la probaiblidad de entrega de comida tras el encendido del TL asociado con ella. Si es una condición señalada, entonces $pSr|TL = 1$, y t será igual a la duración del TL. Si no es señalada, entonces $pSr|TL = pSr$ y t será mayor que la duración de TL en un factor de $1/pSr$. Dado que t depende de $pSr|TL$, la condición de señalización afecta directamente al hecho de que los estímulos de TL den más o menos información temporal. Debe notarse, además, que los TL en la opción óptima sí proveen alguna información temporal, aun cuando se les suele considerar no-informativos.

Ya en el procedimiento de elección subóptima, si solo la alternativa subóptima es señalada, entonces t para ella será $10s/1 = 10s$, y para la óptima $t = 10s/0.5 = 20s$. Convirtiendo C/t en bits de información usando la ecuación previa

$$H = log_2(C/t)$$

obtenemos que el estímulo de TL de la alternativa subóptima entrega 2.6 bits de información, mientras que el de la óptima entrega 1.6. Así, el estímulo del TL de la alternativa subóptima es más temporalmente informativo que el de la alternativa óptima.

Este modelo sugiere que, si ambas alternativas fuesen señaladas, proveerían la misma cantidad de información temporal; y al ser señalada solamente la alternativa óptima o al ser ambas no señaladas, la alternativa óptima proveería más información.

En el caso del procedimiento de magnitudes, en el que ambas alternativas tienen TL señalados, el modelo predeciría indiferencia, cuando los resultados indican preferencia por la alternativa óptima. Se puede sugerir que la magnitud de las recompensas juega igualmente un papel fundamental en la elección. Los autores unen un modelo de Killeen que relaciona la magnitud con el valor, con el modelo de información temporal:

$$V = H(1 - e^{\lambda A})$$

donde V es el "valor de la señal. El valor asintótico de un estímulo es determinado por la información temporal que provee (i.e., H), y magnitudes mayores generan valores más cercanos a la asíntota, en función de λ.

La proporción de elección predicha es igual a

$$pSub = \frac{V_{sub}}{V_{sub} + V_{opt}}$$

Para ajustarse a lo extremas que son las preferencias (quizá dado el procedimiento de ensayo discreto y el entrenamiento independiente en cada alternativa), se agrega un parámetro de sensibilidad

$$pSub = \frac{V_{sub}^{a}}{V_{sub}^{a} + V_{opt}^{a}}$$

donde a es un parámetro que corresponde a la sensibilidad de la elección subóptima al valor relativo de la señal. Esta ecuación fue ajustada a todos los resultados de aves de Gipson (2009) en adelante, y hubo un buen ajuste.

Evidencia de los primeros experimentos en subcho sugiere que hay una relación postivia entre elección subóptima y la duración del TL, y una relación negativa entre elección subóptima y la duración del IL. El modelo indica que las duraciones de TL e IL no deberían modificar la preferencia, de modo que la información temporal relativa por sí misma no puede explicar todos los resultados. Para sortear esta limitación, los autores proponen la idea de que las duraciones de IL y TL podrían afectar el grado de competencia entre reforzamiento primario y el valor de las señales.

El impacto del reforzamiento primario relativo se expresa como

$$pSub = \frac{R_{sub}^{b}}{R_{sub}^{b} + R_{opt}^{b}}$$

donde R es la tasa de reforzamiento primario para cada alternativa, y b es un parámetro libre que determina la sensibilidad de la elección a la tasa relativa de reforzamiento primario. Para formalizar la competencia entre el valor de las señales y la tasa de reforzamiento,

se puede implementar un parámetro de peso

$$pSub = w\frac{V^a_{sub}}{V^a_{sub} + v^a_{opt}} + (1-w)\frac{R^b_{sub}}{R^b_{sub} + R^b_{opt}}.$$

Cuando $w = 1$, la elección se guía solo por las señales; y cuando $w = 0$, se guía solo por el reforzamiento primario. Esto puede entenderse como competencia entre conducta pavloviana (guiada hacia el EC más potente) e instrumental (guiada hacia la recompensa más densa).

Las longitudes de TL e IL se relacionan con el peso w de este modo:

$$w = \frac{1}{1 + e^{-\beta}\left(\frac{D_f}{D_s} - m\right)}$$

donde D_s es la demora promedio a una señal temporalmente informativa (estímulo TL) en el momento de la elección; D_f es la demora promedio a la comida en el momento de la elección; β es un parámetro libre que determina la sensibilidad de w a la tasa de demoras, y m es un parámetro libre que determina la tasa de demora en la cual $w = .5$. m puede verse como un sesgo a favor de el uso del reforzamiento primario como guía de la elección. Cuando una señal temporalmente informativa está más cercana relativa a la comida misma en el momento de la eleción (cuando IL es pequeño y TL relativamente grande, por lo tanto D_f/D_s es grande), w estará cerca de 1 y la elección se regirá por las señales. Así, ahora el modelo captura los hallazgos de la influencia de las duraciones relativas de IL y TL.

La interpretación de la ecuación es que la elecicón subóptima ocurre cuando le pedimos a los animales escoger en un momento en el que estan fuertemente atraídos por las señales pavlovianas, pero las recompensas finales aun no dejan ver su efecto.

Este modelo extendido fue ajustado con los hallazgos previos a Gipson (2009), que manipulaban las demoras de IL y TL, resultando en buenos ajustes.

El modelo puede aplicarse al problema de la diferencia entre especies si se asume un peso w distinto entre ratas y palomas: las ratas pueden estar mayormente influenciadas por la tasa de reforzamiento primario. Los autores sugieren que la saliencia incentiva podría modular esta preponderancia de un componente sobre otro: quizá las ratas, al no usar una modalidad sensorial predominante en ellas, no se ven sesgadas por las señales predictoras de comida.

Limitaciones

Algunas condiciones (valores muy altos de demora de TL con un valor muy pequeño de C) pueden resultar en valores negativos de información temporal. Esto no tiene sentido, porque implicaría que las señales aumentan la incertidumbre más allá de su máximo basal.

Se asume que un S^- no tendrá impacto en la elección, pero un estímulo que señale una probabilidad ínfima de reforzamiento sí debería tenerlo. No está claro si los animales tratarían distinto a esos dos estímulos.

El peso w se calcula en el momento de elección, que es muy claro con IL FR1, pero no tanto cuando se usan programas más largos, como IF. La valoración del peso podría cambiar conforme transcurre el IL.

Los estímulos podrían señalar probabilidades de reforzamiento distintas de 0 y 1. Una misma alternativa podría tener dos estímulos temporalmente informativos, y no esta del todo claro cómo cuantificar la información temporal en ese caso.

Conclusión

Esta aproximación formaliza el papel que la información temporal tiene en la elección subóptima (y no solo la información acerca de si una consecuencia ocurrirá o no). Proporciona un marco de referencia cuantitavito para entender cómo los estímulos influyen enla elección subóptima de una forma consistente con otras preparaciones y con el condicionamiento pavloviano.

Se desarrolló un modelo basado en (1) valor relativo de las señales, (2) tasa relativa del reforzamiento primario, y (3) competencia entre ambos en el momento de la elección.