An Associability Decay Model of Paradoxical Choice

Carter W. Daniels, Federico Sanabria

2018

Llaman "2ABT" a subcho (por *two-armed bandit task* y también por payasos).

Buscan avanzar un modelo que integra perspectivas de modelos asociativos de atención en un modelo que, de otro modo, escogería de forma óptima para explicar los resultados de palomas y ratas en 2ABT.

Las palomas escogen consistentemente TL_{info}, y su preferencia parece ser relativamente insensible a las probabilidades de reforzamiento de TL_{+r}. Las ratas, en cambio, escogen $\mathrm{IL}_{noninfo}$, y la evidencia sugiere que escogerán IL_{info} solo si TL_{+r} y TL_{-g} vienen de modalidades sensoriales distintas (Chow, 2017). Cuando las ratas escogen TL_{info}, su elección disminuye con la probabilidad $P(\mathrm{TL}_{+r})$.

Si los estímulos de TL no son discriminativos, las palomas escogen $\mathrm{IL}_{noninfo}$, de modo que no escogen IL_{info} por aversión a la incertidumbre de $\mathrm{IL}_{noninfo}$.

A diferencia de las ratas, las palomas son insensibles a la frecuencia y duración de TL_{-g} y la modalidad de TL_{+r}, y la inhibición condicionada atribuida a TL_{-g} disminuye con el entrenamiento. Además, las palomas escapan de TL_{-g}. En las ratas, al contrario, la inhibición condicionada atribuida a TL_{-g} no disminuye con el entrenamiento mientras ambos TL estén en la misma modalidad. *Así, la elección en 2ABT probablemente involucra la interacción entre el valor de la información y las propiedades de inhibición condicionada de* TL_{-g}

Vasconcelos et al. (2015) propusieron un modelo derivado del modelo de elección secuencial (SCM) que plantea que las palomas escogen IL_{info} debido a que (1) provee información de forma temprana y (2) no atienden al estímulo TL_{-g} debido a que anuncia no-reforzamiento, y en ambientes naturales indicaría una búsqueda infructuosa que se debe abandonar.

Iigaya (2016) propuso el modelo de utilidad anticipatoria, que supone que el valor de TL se puede descomponer en un valor descontado de la consecuencia y un valor anticipatorio para esa consecuencia; además, se asume que la anticipación se potencia en presencia de estímulos que anuncian el resultado, sea positivo o negativo. Este modelo solo explica la elección de las palomas si se asume que ignoran al inhibidor condicionado. Así, las palomas escogen TL_{info} debido a (1) la anticipación positiva de TL_{+r}, (2) la potenciación de esa anticipación dada la discriminabilidad el estímulo TL_{+r}, y (3) que ignoran TL_{-g}.

Ambos modelos asumen que la alocación de la atención es crítica, pero no especifican cómo esto sucede. Además, no esta claro cómo se adaptan a los resultados de las ratas y su sensibilidad a la frecuencia, duración y modalidad de TL_{-g} sin cambiar supuestos fundamentales. Tampoco aclaran cómo las palomas adquieren la preferencia, pues ambos modelos requieren conocimiento a priori sobre el tiempo hasta una consecuencia. Zhu (2017) propuso un modelo alternativo para abordar algunas de esas limitaciones: el *anticiparoty prediction error model*. Sin embargo, también presupone una alocación diferencial de la atención sin explicarla con dinámicas ensayo por ensayo.

Proponen el modelo de decaimiento de la asociabilidad, que especifica cómo las palomas aprenden a ignorar el TL_{-g}. Asume que la atención, y por lo tanto, la asociabilidad de cada TL y su consecuencia con respecto a IL, es una función inversa de la certeza de la consecuencia, señalada por el TL, y no una función positiva de la predictabilidad de la recompensa. Incorpora la noción de la certidumbre por medio de los errores de predicción de recompensa (RPEs): los RPE son la diferencia entre la recompensa obtenida y la esperada. Si la recompensa obtenida es mayor que la esperada, el RPE es positivo, y a la inversa. El modelo asume que conforme los valores absolutos de RPE se hacen más pequeños (disminuye la diferencia entre lo esperado y lo obtenido), el valor de IL se vuelve menos maleable. El modelo es una variación del modelo de atención de Pierce y Hall (1980), según el cual la atención se centra en estímulos que señalan resultados inciertos.

El modelo ADM

Los valores de IL y TL se actualizan en cada ensayo de acuerdo con

$$\begin{aligned}\Delta V_t\left(TL_i\right) &= V_t\left(r_k \mid TL_i\right) - V_t\left(TL_i\right)\\ V_{t+1}\left(TL_i\right) &= \begin{cases} V_t\left(TL_i\right) + \alpha\Delta V_t\left(TL_i\right), & \text{si } V_t\left(r_k \mid TL_i\right) > 0,\\ V_t\left(TL_i\right) + \beta\Delta V_t\left(TL_i\right), & \text{si } V_t\left(r_k \mid TL_i\right) = 0\end{cases}\end{aligned}$$

y

$$\begin{aligned}\Delta V_t\left(IL_j\right) &= \left[V_t\left(TL_i\right) + \gamma V_t\left(r_k \mid TL_i\right)\right] - V_t\left(IL_j\right)\\ V_{t+1}\left(IL_j\right) &= \begin{cases} V_t\left(IL_j\right) + \alpha\Delta V_t\left(IL_j\right), & \text{si } V_t\left(r_k \mid TL_i\right) > 0,\\ V_t\left(IL_j\right) + \beta\Delta V_t\left(IL_j\right), & \text{si } V_t\left(r_k \mid TL_i\right) = 0\end{cases}\end{aligned}$$

donde k, i, y j indican la recompensa obtenida, el TL obtenido y el IL escogido; t indica el ensayo; $V_t(r_k \mid TL_i)$, $V_t(TL_i)$ y $V_t(IL_j)$ son los valores de la recompensa obtenida dado el TL obtenido (expresado como porporción del valor de la recompensa más grande), el valor del TL obtenido, y el valor del IL escogido en el ensayo t; α y β son las tasas de aprendizaje cuando $V_t(r_k \mid TL_i) > 0$ y cuando $V_t(r_k \mid TL_i) = 0$, respectivamente; y γ es el factor de descuento de la recompensa. $0 \leq \alpha, \beta, \gamma \leq 1$.

En cada ensayo, se asume que los sujetos escogen el IL con el valor más alto dado cierto ruido. La probabilidad de escoger $\text{IL}_{noninfo}$ en el ensayo t es

$$p_t(IL_{Noninfo}) = \frac{1}{1 + e^{\tau\left[V_t\left(IL_{Info}\right) - V_t\left(IL_{Noninfo}\right)\right]}},$$

que es una función *softmax* de $V_t(IL_{info}) - V_t(IL_{Noninfo})$ con un parámetro de ruido inverso $\tau \geq 0$. El error aumenta cuando τ tiende a 0, y disminuye cuando tiende a infinito.

Sin embargo, eso predice preferencia por $\text{IL}_{noninfo}$. Para explicar la preferencia por IL_{info} el modelo asume que el aprendizaje de IL esta modulado por la certeza que un sujeto tiene por un TL y su recompensa. La asociabilidad de un IL con su TL y consecuencia (descontada) disminuye cuanto más predictivo es el TL, e incrementa a un máximo tras un resultado inesperado:

$$H_{t+1}\left(TL_i\right) = \begin{cases} H_t\left(TL_i\right)\delta, & \text{si } \left|\Delta V_t\left(TL_i\right)\right| < \theta \\ H_{MAX}, & \text{si } \left|\Delta V_t\left(TL_i\right)\right| \geq \theta \end{cases}$$

Donde $H_t(TL_i)$ es la asociabilidad de un TL y su resultado con su IL correspondiente en el ensayo t; H_{MAX} es la asociabilidad máxima; δ es la proporción de asociabilidad obtenida de ensayo a ensayo ($1-\delta$ es la proporción en la cual la asociabilidad decae por ensayo) cuando el valor absoluto de TL RPE está debajo del umbral θ de decaimiento de asociabilidad; y $H_i(TL_i) = H_{MAX}$ al comienzo del entrenamiento y cuando $TLRPE \geq \theta; 0 \leq \theta, \delta \leq 1$. El parámetro θ es el grado de certeza requerido para que la asociabilidad decaiga. Cuando θ es pequeña, se requiere más certeza para que decaiga (menores valores absolutos de TL RPE).

Dado este mecanismo de decaimiento de asociabilidad, el aprendizaje de IL se modifica de esta forma:

$$\omega_t\left(TL_i\right) = \tfrac{H_t(TL_i)}{1+H_t(TL_i)}$$
$$V_{t+1}\left(IL_j\right) = \begin{cases} V_t\left(IL_j\right) + \omega_t\left(TL_i\right)\alpha\Delta V_t\left(IL_j\right), & \text{si } V_t\left(r_k \mid TL_i\right) > 0, \\ V_t\left(IL_j\right) + \omega_t\left(TL_i\right)\beta\Delta V_t\left(IL_j\right), & \text{si } V_t\left(r_k \mid TL_i\right) = 0 \end{cases}$$

donde $\omega_t(TL_i)$ es la probabilidad de asociar un TL y su consecuencia con un IL; y $H_t/TL_i)$ es la probabilidad (*odds*) de esa asociación. Mientras $\omega_t(TL_t)$ tiende a cero, la probabilidad de que un TL se actualice decrece, haciendo a IL insensible a los cambios en TLs y consecuencias.

Todas estas ecuaciones forman al *associability decay model*, que es parsimonioso dado que tiene pocos parámetros y éstos tienen consistencia teórica y empírica.

Probando el modelo ADM

El modelo fue ajustado a los datos de varios estudios con ratas y palomas.

Palomas

En casi todos los estudios de palomas $\beta > \alpha$, lo que sugiere que las palomas aprenden más rápido acerca de recompensa que de no-recompensa. τ fue grande en todos los casos,indicando poco ruido en la elección de IL, lo que sugiere gran regularidad en los datos. Las medias de los parámetros de asociabilidad fueron de $\delta = .609$ y $\theta = .364$, lo que indica que la asociabilidad decayó relativamente rápido para todos los TLs, perdiendo la mitad de su valor en dos ensayos, pero no comenzó a decaer para ningún TL sino hasta que el valor absoluto de un TL RPE dado era reducido en cerca del 27 % de su valor inicial (arbitrario) de .5. El parámetro γ pareció depender de la variante de la tarea (probabilidades o magnitudes) que se evaluaba.

El modelo describió adecuadamente los datos de Stagner et al. (2011) [probabilidades], Laude et al. (2014) [magnitudes], Stagner y Zentall (2010) y Zentall y Stagner (2011) [que volvieron no-discriminativa a la alternativa discriminativa].

La dinámica del cambio indicada por el modelo sugiere que, con el entrenamiento, $V_t(TL_{+r})$ y $V_t(TL_{-g})$ dejan de actualizar el valor de IL_{info}, pero $V_t(TL_{0.5y})$ y $V_t(TL_{0.5b})$ continúan manteniendo bajo el valor de $\text{IL}_{noninfo}$. Además, reducir la discriminabilidad de TL_{+r} y TL_{-g} recupera su asociabilidad y les permite reducir el valor de IL_{info}.

El modelo también fue ajustado a los experimentos que probaron los efectos de (1) diferencias individuales en impulsividad, (2) privación de comida, y (3) enriquecimiento ambiental. En cada experimento hicieron una comparación entre los grupos experimentales y control, permitiendo que un solo parámetro variara entre los grupos. Eligieron como parámetro libre a aquel que llevaba a un mejor ajuste en los datos. Sus ajustes sugieren que (1) las palomas impulsivas descuentan *más despacio* que las no impulsivas (wtf), (2) que el decaimiento de asociabilidad es más rápido para las palomas hambrientas, y que (3) las palomas enriquecidas requieren de menor certeza para que la asociabilidad comience a decaer.

Ratas

Se ajustó el modelo a los datos de ratas. Los parámetros indican que para las ratas, el peso de recompensa y no-recompensa dependen de la longitud de IL. El parámetro τ fue grande, indicando poco ruido en la elección de IL. Los parámetros θ, δ y γ indican que, al contrario de las palomas, no se requiere de decaimiento de la asociabilidad para explicar los resultados de las ratas. Un ajuste que realizaron para adecuarse al hallazgo de que, al incrementar la longitud de TL de 10 a 30s, la preferencia por IL_{info} disminuye indicó que, cuando los TL se alargan, las ratas descuentan más las recompensas y se vuelven más sensibles a la no-recompensa.

Discusión

El modelo ADM supone que el decaimiento de la asociabilidad esta inversamente relacionado con la atribución de inhibición condicionada, y que el decaimiento depende de la certeza del animal sobre la consecuencia esperada. La actualización del valor de IL depende de la asociabilidad de cada TL, que decae a cieta tasa una vez que el error de predicción de recompensa (RPE) absoluto cae debajo de cierto umbral.

Otros modelos (Vasconcelos, Iigaya, Zhu) explican los mismos datos que ADM, pero presuponiendo que la asociabilidad es estática o no indicando cómo cambia ensayo por ensayo, de modo que no especifican cómo surgen la diferencias en pesos asociativos.

La explicación ofrecida por el ADM es la siguiente: **la elección sistemática de IL_{info} en elección subóptima se debe a que, primero, el TL_{-g} predice con certeza la ausencia de recompensa, de modo que las palomas aprenden a ignorarlo, lo que evita que el IL_{info} pierda valor. Después, las palomas aprenden a atender a los TLs que predicen probabilísticamente la recompensa, lo que reduce el valor de $\text{IL}_{noninfo}$ relativo al valor de IL_{info}. Finalmente, las palomas aprenden a ignorar cualquier otro TL que prediga con certeza las consecuencias finales.**

Parece que ADM también explica los efectos de el enriquecimiento y la privación de comida. Argumentan que los cambios en los parámetros son consistentes con la noción de que la atención por las recompensas y los estímulos condicionados es una función de la motivación y el enriquecimiento social.

Sin embargo, la dinámica de la asociabilidad no parece explicar la relación de la elección con las diferencias en impulsividad. En lugar de ello, la explicación parece estar en el factor

de descuento γ.

El modelo ADM hace una predicción contraria a otros modelos: dada la relación entre la certeza y la asociabilidad, cuando incrementa la duración de TL, debería incrementar la preferencia por IL_{info} en la variante de probabilidades, pero debería decrementar en la de magnitudes. Además, una vez que la asociabilidad ya ha decaído para un TL, debería ser insensible a las manipulaciones futuras, a menos que la manipulación induzca un TL RPE muy grande.

Para las ratas, el ADM predice que la asociabilidad no decae, lo que es consistente con la noción de que el decaimiento está inversamente relacionado con el grado de atribución de inhibición condicionada.

Sugieren como explicación a la discrepancia en los resultados de Chow (2017) y López (2018) la posibilidad de que, al venir de la misma modalidad sensorial, en el caso de López, los TL+ y TL- no eran lo bastante discriminables, de modo que TL- le restaba valor a TL+ e impedía la elección de IL_{info}, aunque los datos de Lopez no muestran problemas de discriminabilidad. **Una explicación alternativa es que, igual que la saliencia incentiva, la inhibición condicionada dependa de la modalidad sensorial**, aunque se trata de especulación. Si esto se presupone, entonces ADM puede explicar en esencia todos los datos reportados con ratas.

Aun así, la duda que prevalece es por qué, en una sola modalidad sensorial, la asociabilidad decae para las palomas pero no para las ratas.

Limitaciones

Dado que no había datos ensayo por ensayo disponibles, no se pudo ajustar los modelos usando máxima verosimilitud o estimación jerárquica bayesiana. Por lo tanto, los parámetros estimados deben tomarse como aproximaciones y no como valores precisos. Además, la evaluación del ADM se limitó a estudios con datos de adquisición, así que no es posible tomar en cuenta fenómenos como el efecto de las demoras diferenciales al inicio de TL tras la elección de IL_{info}.

Aunque permite explicar la conducta de elección en ratas y palomas, el ADM actual no atiende a todas las diferencias entre ellas. Por ejemplo, las ratas, pero no las palomas, son sensibles a $P(TL_{+r})$.

Este trabajo revela que un modelo otrora óptimo (operadores lineales de Bush Mosteller) puede hacer elecciones paradójicas tras incorporar ideas de modelos asociativos de atención (mecanismo de asociabilidad Pearce Hall). Este modelo explica cómo las palomas aprenden a ignorar el TL- al permitir que la asociabilidad decaiga tras pasar un cierto umbral de certidumbre en los TL.