**Trabajo final teoría econométrica. Luis Luarte.**

**Introducción.**

Relevancia del problema. El objetivo central de este trabajo es generar un punto de inicio para la investigación del mecanismo subyacente al comportamiento de toma de decisiones. Se basa en la hipótesis de que este mecanismo comparte un origen evolutivo con el forrajeo, siendo este último el que daría lugar a un proceso de exaptación (Gould, 1991) —cuando un rasgo que evoluciona para una función es cooptado para otra sin ser necesariamente sujeto a mutación— para la toma de decisiones. Este problema ha sido típicamente abordado utilizando paralelos entre ambos tipos de comportamiento (Hills et al., 2012, 2015), e incluso se han propuesto estructuras cerebrales que potencialmente podrían estar detrás de ambos (Chen et al., 2024; Nassar, 2024). Sin embargo, se ha considerado en menor medida la posibilidad de que ambos comportamientos estén modulados por un tercer mecanismo evolutivamente anterior: el comportamiento espontáneo.

Esta idea es plausible, ya que se ha demostrado que el comportamiento espontáneo tiene una estructura altamente adaptativa (Maye et al., 2007), organizando ensambles neuronales que determinan el comportamiento de búsqueda de presas (Romano et al., 2015) y el control del sistema dopaminérgico, el cual se relaciona directamente con la toma de decisiones (Markowitz et al., 2023). Así, la relevancia de este trabajo es entregar una primera aproximación a la conexión entre el comportamiento espontáneo y la toma de decisiones, definiendo un marco de análisis para comprobar la hipótesis —y en un futuro también para relacionar el comportamiento espontáneo con el de forrajeo— bajo un marco evolutivo coherente.*Pregunta de investigación.* ¿Puede la variabilidad del comportamiento espontaneo, predecir la estocasticidad del comportamiento exploratorio de la toma de decisiones económicas, sugiriendo un origen computacional común?

*Modelo teórico o mecanismo que explica la relación entre las variables que se van a analizar e hipótesis.* Los modelos clásicos de la función cerebral típicamente describen a este órgano como primariamente reactivo, estableciendo relaciones input-output donde el comportamiento es una respuesta determinística a los estímulos sensoriales externos. Dentro de este marco conceptual, la variabilidad del comportamiento es típicamente considerada como “ruido”, resultante de implementaciones biológicas imperfectas. Sin embargo, este visión ha sido desafiada, proponiendo que el cerebro es un órgano activo y predictivo en naturaleza, el cual incluso en ausencia de input sensorial o alguna tarea explicita, continua en su función explorando un repertorio de estados de conectividad preferidos o recurrentes (Avitan & Stringer, 2022; Desbernats et al., 2023; Maye et al., 2007).

Esta actividad intrínseca recurrente no es puramente aleatoria. Estudios a lo largo de múltiples especies han demostrado que el comportamiento espontaneo exhibe una estructura temporal compleja. Por ejemplo, las maniobras de vuelo de la Drosophila siguen una distribución marcada por colas pesadas (distribución tipo Lévy), un patrón que ha mostrado ser óptimo para el forrajeo (Viswanathan et al., 1999), sugiriendo que este es un rasgo que evoluciono y se adaptó, no simplemente “ruido”. Mas aun, este comportamiento espontaneo puede ser comprendido como una representación de alta dimensionalidad del comportamiento del animal y su contexto interno (Avitan & Stringer, 2022). La pregunta relevante es, entonces, como esta estructura presente en la actividad espontanea surge y que rol funcional sirve particularmente en cognición del alto nivel.

Los origines funcionales del comportamiento espontaneo aparecen profundamente arraigados en el problema fundamental de la exploración. En el pez cebra, la actividad neural espontanea en el colículo superior se encuentra organizado en ensambles funcionales que están sincronizados para la detección de estímulos vitalmente relevantes como las presas (Romano et al., 2015). La emergencia espontanea de estos circuitos de detección de presas, aun sin input retinal, indica que el cerebro esta intrínsecamente organizado para explorar el ambiente en búsqueda de oportunidades. Lo anterior refleja la posición activa del cerebro en la exploración mostrando una solución filogenética al problema del forrajeo.

Este marco centrado en la exploración puede ser extendido del espacio físico (forrajeo) al espacio abstracto de la toma de decisiones. El dilema de exploración-explotación es el desafío central en el aprendizaje por refuerzo (Sutton & Barto, 2020), donde un agente debe balancear explotar opciones conocidas de alta recompensa con explorar aquellas menos conocidas, para ganar información y conocer alternativas potencialmente superiores. Este balance típicamente se formaliza en modelos en donde las decisiones son guidas por el valor estimado de cada una de las opciones y la incertidumbre asociada a las mismas. Dicha incertidumbre puede servir como un “bono de exploración”, motivando al agente a elegir opciones más informativas (Stojić et al., 2020). Esta forma de exploración presente en espacios abstractos para reducir la incertidumbre, podría ser un caso de exaptacion (Gould, 1991) de programas motores de exploración evolutivamente mas antiguos utilizados para el forrajeo en espacio físicos (Hills et al., 2012, 2015).

Propongo un mecanismo computacional común que gobierna la variabilidad en el comportamiento motor espontaneo y el comportamiento estructurado dirigido a objetivos con estimación de valor. El sustrato para este mecanismo seria un sistema central neuromodulador, que provee de una señal de instrucción (o recompensa) la cual es independiente de recompensas externas explicitas.

Recientemente, se ha observado evidencia a favor de la existencia de dicho sustrato, mostrando como las fluctuaciones de dopamina (espontaneas) en el estriado dorsolateral actúan como una señal de refuerzo intrínseca que, causalmente, determina la secuencia y selección módulos de comportamiento espontaneo in ratones en esquemas de movimiento libre (Avitan & Stringer, 2022). Este proceso opera como un algoritmo de aprendizaje por refuerzo, donde el cerebro busca maximizar la señalización interna de dopamina, lo cual lleva por consecuencia la estructura del comportamiento en condiciones libres (Markowitz et al., 2023).

El mecanismo que propongo conecta lo anteriormente mencionado con el comportamiento de toma de decisiones en humanos:

1. La variabilidad de los patrones motores espontáneos son representaciones del contexto interno exploratorio. Alta variabilidad en estos patrones motores refleja un estado donde la actividad conjunta del cerebro estaría sesgada hacia la exploración, consistente con un modo cercano forrajeo.
2. El mismo estado interno directamente modula la estocasticidad de las decisiones (en espacios abstractos) subsiguientes. En el contexto de aprendizaje por refuerzo esto estaría capturado por el parámetro τ de la regla de decisión, la cual determina explícitamente el balance entre exploración (más estocasticidad) y explotación (más determinista).
3. Las mediciones de la variabilidad del comportamiento espontaneo deberían actuar como predictores del nivel de estocasticidad en la toma de decisiones. Alta variabilidad en el comportamiento espontaneo induciría una mayor τ, reflejándose en un fenotipo conductual de decisión mas aleatorio y menos determinista. Lo que sugiere que la decisión de cuanto explorar, no estaría completamente determinado por cada decisión, sino que estaría continuamente modulado por un estado interno general mantenido por las dinámicas del generador del comportamiento espontaneo (potencialmente siendo el control de los niveles circulantes de dopamina).

Hipotetiza que la variabilidad en el comportamiento espontaneo, medido a través de las sacadas, durante periodos de tiempo donde no existe ninguna instrucción o estimulo saliente, van a predecir positivamente el grado de estocasticidad cuantificado por los niveles del parámetro τ obtenido por máxima verosimilitud para un modelo de aprendizaje de refuerzo “Q-learning” (Sutton & Barto, 2020), en las decisiones posteriores. Esto proveería de un mecanismo computacional compartido, el cual gobernaría los niveles de exploración tanto en el forrajeo (trabajo futuro) como en la toma de decisiones económica (presente trabajo).

**Metodología.**

*Descripción de los datos y la muestra.* Todos los datos primarios fueron obtenidos del estudio de (Stojić et al., 2020) desde la open science framework (<https://osf.io/539ps>). Brevemente, el estudio original consistía en la presentación de 6 opciones las cuales, al momento de la elección, entregaban una recompensa proveniente de una distribución normal con μ = 6 y δ = 2.75 para la mejor opción, las demás opciones variaban su recompensa esperada según:

Donde j es el ranking de la opción, lo que permitía que la dificultad de discernir entre dos opciones estuviese controlada. Adicionalmente, la estructura de varianzas determinaba dos tipos de juegos, el primero tenía una estructura de reducción desde la mejor opción (en términos de recompensa esperada) a la peor en -0.4, mientras que en el segundo bajaba desde la mejor a la tercera mejor opción en -0.4 y luego aumentaba su varianza hasta la peor opción. Antes de que los participantes escogieran su opción, se les presentaba una cruz de fijación durante ~5 segundos mientras los movimientos oculares eran registrados. A partir de estos datos se generaron dos sets de datos (a) variabilidad comportamiento espontaneo (VESP) y (b) estocasticidad de la toma de decisiones (VDEC).

*Variabilidad del comportamiento espontaneo*. Las manipulaciones de datos fueron realizadas con el script *extract\_saccades*, primeramente, se consideraron los puntos de fijación de los sujetos obtenidos mediante eye-tracker, para cada sujeto, en cada trial se calculó la distancia euclideana entre los puntos de fijación en base a la proyección en los ejes XY de una pantalla (las unidades serán consideradas en pixeles) lo cual correspondería al trayecto de la sacada. Posteriormente, se excluyeron las sacadas que tenían valores superiores a 1.5 veces el rango intercuartílico, los cuales probablemente corresponde a errores de medición del aparato. Finalmente, la variabilidad de las sacadas fue determinada por la desviación estándar de las mismas en una base de trial a trial, sujeto a sujeto, juego a juego.

*Estocasticidad de la toma de decisiones.* Las manipulaciones de los datos fueron realizadas con el script *extract\_rl\_params*. Primeramente, se consideraron los datos del comportamiento de toma decisiones, los cuales se componen de la opción elegida y la recompensa obtenida para cada trial de cada juego. En base a ello se ajustaron modelos de Q-learning (Sutton & Barto, 2020), los cuales apuntan a describir el comportamiento de un agente que balancea la exploración-explotación en un contexto de maximización de recompensas, en este caso, al procedimiento se le determina el problema inverso de refuerzo, esto es, determinar los parámetros mas plausibles del modelo en base a una serie de decisiones y recompensas que fueron empíricamente obtenidas. El modelo cuenta con dos partes centrales:

1. Estimación de los Q-values: Para cualquier trial t, luego de seleccionar una opción at y recibir una recompensa rt, el modelo computa el error de predicción δt. Este error representa la diferencia entre la recompensa recibida y la recompensa esperada (el Q-value) para la acción elegida:

Este error de predicción luego es utilizado para actualizar el Q-value para at. Esta actualizada es escalada por el primer parámetro α o tasa de aprendizaje, que es un parámetro libre entre 0 y 1, determinando que tan relevante es la influencia del error de predicción en los valores existentes.

Los Q-values para las acciones restantes (elección de las otras opciones) no se actualizan y permanecen idénticos para el siguiente trial, esta decisión es únicamente para evitar mayor complejidad del modelo.

1. Selección de la acción en base a la función softmax: para traducir los Q-values en un vector de probabilidad de acción, el modelo utiliza una función softmax o Boltzmann. Esta función asigna una probabilidad a cada acción basado en los Q-values de cada una, permitiendo un balance entre la exploración y la explotación de la opción con mayor valor estimado. La probabilidad P de seleccionar una determinada acción i del conjunto de todas las acciones posibles (6 en este caso) en cada trial t, esta determinado por:

El procedimiento así está determinado por un único parámetro libre τ o temperatura. Cuando τ tiende a 0, la probabilidad de elegir la opción con el Q-value mas alto se acerca a 1, generando un comportamiento determinista o explotativo. Cuando τ tiende a infinito, las probabilidades de todas las acciones se vuelven idénticas, generando un comportamiento más aleatorio o exploratorio.

Todos los parámetros libres son estimados utilizando un proceso de estimación por máxima verosimilitud (MLE). Lo central del procedimiento es encontrar α y τ para maximizar la probabilidad de observar la secuencia de elecciones realizada por cada sujeto, dado el modelo Q-learning, minimizando la log-verosimilitud negativa. La optimización iterativa se realizó ocupando optim (R Core Team, 2025) con el método “L-BFGS-B” para optimización restringida α = [0, 1] y τ = [0.001, 10]. El procedimiento de optimización fue realizado 100 veces con valores de inicio aleatorizados para cada sujeto, para cada juego. Los parámetros finales corresponden a aquellos con la menor log-verosimilitud.

El procedimiento de estimación de los parámetros genera un resumen de la conducta a lo largo de toda la tarea, sin embargo, mi objetivo era realizar comparaciones trial a trial, para ello genere un procedimiento simple.

1. Se determina un parámetro α óptimo para cada sujeto, para cada juego. Este α se asume como un rasgo estable a lo largo de la tarea y por lo tanto como globalmente optimo. Si bien existen propuesta de α dinámicos (Roesch et al., 2012), dada la corta duración de la tarea (60 trials por juego) asumir una tasa de aprendizaje constante es conservador.
2. Utilizando un α constante, el modelo Q-learning se corre nuevamente de manera determinista. Esta simulación no genera nuevas decisiones, utiliza la secuencia empírica obtenida por el participante para reconstruir la estimados de la recompensa de cada opción (Q-values) trial a trial usando:
3. Con la secuencia de Q-values reconstruida, el parámetro τ se estima. Para cada trial t, el parámetro de temperatura τt es tratado como una variable desconocida. Se aplica un procedimiento de estimación MLE para encontrar el valor τt que maximiza la probabilidad de at del sujeto, dado los valores Q reconstruidos:

Para construir una buena representación de los Q-values se generaron 500 corridas, para las cuales el parámetro fue estimado trial a trial, para finalmente tomar un promedio a lo largo de todas las 500 corridas (procedimiento realizado en el script *analysis*)

Para los sets de datos finales se ocuparon los datos de 12 mujeres y 11 hombres, la descripción de la muestra se encuentra en la Tabla 1, las distribuciones de las variables en Ilustración 1, Ilustración 2 e Ilustración 3.

Para testear la hipótesis de que VESP predice la VDEC subsiguientes, decidí utilizar un modelo lineal mixto (LMM). Esta aproximación estadística fue seleccionada para considerar de manera apropiada la estructura jerárquica de los datos, donde múltiples observaciones a nivel de trial se encontraban anidadas al interior de los participantes en cada juego. Esta estructura anidada viola el supuesto de observaciones independientes requerido por la regresión lineal estándar, dado que las observaciones dentro de un mismo participante serán potencialmente mas semejantes entre si que aquellas de diferentes participantes. Para modelar las diferencias individuales, especifique una estructura de efectos aleatorios con un intercepto aleatorio para cada sujeto, lo que consideraría las diferencias en VESP. Si bien teóricamente podría haber incluido una pendiente aleatoria para la variabilidad de las sacadas para cada sujeto, permitiendo que la fuerza de la asociación entre el comportamiento espontaneo y la elección variara entre sujetos a lo largo de los trials, esto complejiza un modelo que es principalmente exploratorio por lo cual no poseo ninguna hipótesis concreta para determinar ese tipo de variabilidad y adicionalmente el modelo muestra problemas de convergencia al realizar esa adición. Un modelo lineal simple (sin efectos aleatorios) también fue ajustado como análisis complementario. Una de las limitaciones de este modelo es que la variable dependiente τ, es estrictamente positiva (por definición) lo que eventualmente viola el supuesto de normalidad de los residuos, lo cual puede llevar a predicciones invalidas. En caso de que esto quisiera ser remedado se podrían hacer uso de modelos lineales generalizados con una distribución tipo Gamma con función logarítmica como enlace o al menos generando una transformación logarítmica directa a la variable τ. Finalmente, el modelo será especificado con τ como variable dependiente, la desviación estándar de las sacadas como variable independiente principalmente, y edad, sexo, tiempo medio de respuesta y máxima cantidad de recompensa obtenida como controles, principalmente para atender a fenómenos de orden motor y de niveles de recompensa.

**Resultados.**

*Caracterización de la muestra (estadísticas descriptivas).* Las estadísticas principales de la muestra se encuentran en la Tabla 1. Respecto a las variables más relevantes, no encontré diferencias significativas para τ entre los juegos (βjuego1 – juego2=0.09±0.17, p = 0.59, Ilustración 4), y de manera importante testee si la variabilidad de las sacadas era modificada en términos generales a lo largo de los trials, ya que esto podría invalidar el considerar la estructura de trials propuesta. La variabilidad permanece constante a lo largo de los trials (β=0.0006±0.0003, p = 0.06, Ilustración 3). Adicionalmente, no detecte diferencias entre el tiempo medio de reacción entre sujetos y entre juegos (p > 0.05, para ambos, modelo lineal simple), ni para la máxima cantidad de recompensas obtenidas (p > 0.05, para ambos, modelo lineal simple). Para el tiempo de reacción y recompensas ocupe estadística inferencial, ya que la descriptiva es difícil de visualizar.

*Pasos para el análisis y resultados en cada etapa, considerando las posibles violaciones a los supuestos de OLS en su análisis.*

*Modelo 1.* Este modelo considera únicamente VESP y juego como independientes, VDEC como variable dependiente e intercepto aleatorio. El modelo muestra que τ aumenta en 2.18 unidades ante el aumento de una unidad de desviación estándar en el largo de las sacadas (esto es significativo para juego considerado como variable continua, y con p = 0.056 para juego como variable dummy), adicionalmente, la interacción entre sacadas y juego muestra que esta asociación se revierte para el juego 2, sin embargo las “estimated marginal trends” indica que la pendiente para el juego 2 no es diferente de 0 (βjuego2=-0.472±0.47, p = 0.31), por lo tanto la significancia obtenida por el modelo indica que a diferencia del juego 1 en el juego 2 la pendiente no es positiva.

Modelo 2 (Ilustración 6). Este modelo incluye las variables control de trial, para atender a cualquier efecto de tipo “fatiga” o “habituación”, edad en tanto, en términos generales, la variabilidad del comportamiento tiende a reducir sistemáticamente con la edad, también se atendieron a potencialmente diferencias por sexo, tiempo de reacción típico y recompensas obtenidas, por razones anteriormente mencionadas. El coeficiente de la variabilidad de sacadas se redujo considerablemente en este modelo a 1.16, manteniendo la significancia estadística. Bajo este modelo, τ se reduce en aproximadamente 0.004 unidades de trial a trial, si bien esto alcanza la significancia estadística la magnitud del coeficiente es marginal en la escala del parámetro τ, y responde en parte a un efecto esperado de que los participantes, a medida que acumulan información, opten por decidir de manera mas determinista. La cantidad máxima de recompensa obtenida también obtuvo significancia estadística, τ se vio reducido en 0.02 por cada unidad arbitraria de recompensa (en total), si bien el coeficiente es relativamente bajo, esto puede indicar que los sujetos que detectaron de mejor manera la estructura de recompensas generan decisiones más deterministas en términos generales. La interacción con juego como dummy mantiene básicamente la misma interpretación que para el modelo anterior.

**Discusión/conclusiones.**

Los resultados de este estudio respaldan la hipótesis de que un mecanismo computacional compartido subyace tanto al comportamiento espontáneo como a la toma de decisiones económicas. Se encontró que la variabilidad de los movimientos sacádicos espontáneos (VESP) predice de la estocasticidad de la elección subsiguiente (τ). Esta relación es flexible y sensible al contexto, como lo demuestra la interacción significativa con el tipo de juego, lo que indica que la estructura de recompensas del entorno modula la conexión entre la exploración motora y la cognitiva. Estos hallazgos son consistentes con un marco evolutivo de exaptación, donde los mecanismos de decisión de alto nivel cooptan sistemas filogenéticos más antiguos, originalmente desarrollados para el forrajeo y la búsqueda espacial.

A pesar de su significancia, estos resultados deben interpretarse considerando las limitaciones del modelo estadístico. El uso de un Modelo Lineal Mixto (LMM) estándar, si bien es adecuado para la estructura anidada de los datos, asume una normalidad de los residuos que probablemente es violada por la naturaleza estrictamente positiva y asimétrica del parámetro τ. Esta violación podría sesgar los errores estándar e inflar la significancia estadística de los coeficientes, por lo que los hallazgos deben ser considerados como una evidencia preliminar sólida. Para confirmar la robustez de estas conclusiones, investigaciones futuras deberían emplear un Modelo Lineal Generalizado Mixto (GLMM) con una distribución Gamma, que se ajusta mejor a este tipo de datos. Además, será crucial investigar las características del entorno que impulsan la interacción observada para comprender cómo se regula este vínculo fundamental, además de buscar controles experimentales para la cantidad de recompensa obtenida. Finalmente, estos modelos y análisis pueden ser usados para explorar la parte restante del modelo conceptual, esto es, la asociación entre VESP y el comportamiento de forrajeo, potencialmente utilizando ambientes virtuales.

Figuras y tablas.

A screenshot of a table

AI-generated content may be incorrect.

Tabla 1 Descripción de la muestra finalmente utilizada. Las variables “Mean RT” es el tiempo promedio de respuesta en segundos durante toda la tarea, “Max. Reward” es la cantidad de recompensa obtenida durante cada juego.

A graph of a graph

AI-generated content may be incorrect.

Ilustración 1 Densidades del largo de las sacadas en pixeles para cada sujeto a lo largo de toda la tarea.

A graph of a temperature

AI-generated content may be incorrect.

Ilustración 2 Distribución global del parámetro τ a través de todos los sujetos y trials.

A graph of a graph

AI-generated content may be incorrect.

Ilustración 3 Desviación estándar de las sacadas a lo largo de los trials, a lo largo de ambos juegos.

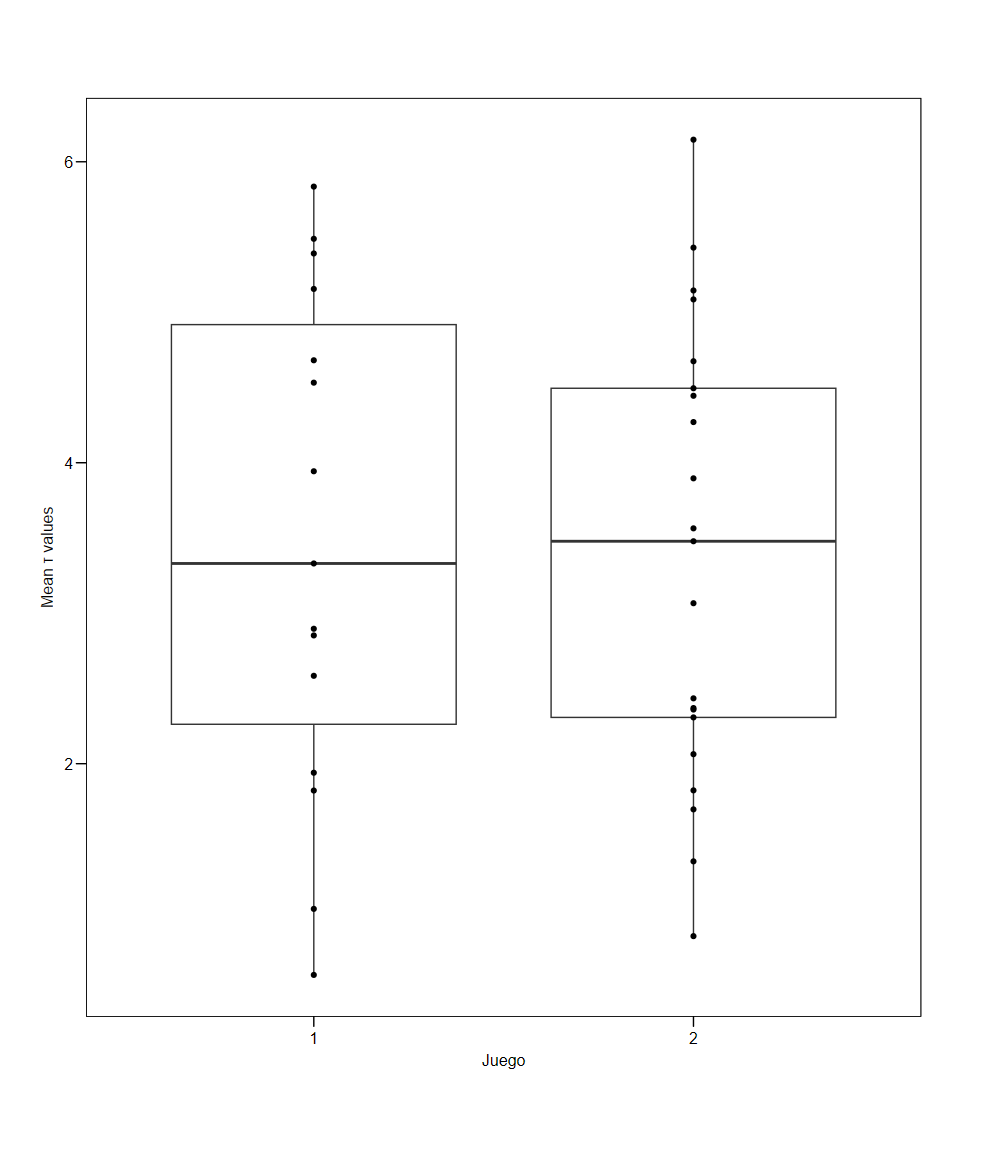


Ilustración 4 Valores promedio del parámetro tau en los diferentes juegos.

A screenshot of a paper

AI-generated content may be incorrect.

Ilustración 5 Resultados de los dos modelos, modelo (1) simple, modelo (2) con variables de control.

A graph of different colored lines

AI-generated content may be incorrect.

Ilustración 6 Estimados del modelo (2), para todas las variables control no representadas visualmente se ocuparon sus valores promedio. Las líneas representas las “estimated marginal means” del modelo y las bandas los intervalos de confianza al 95%.

**Referencias.**

Avitan, L., & Stringer, C. (2022). Not so spontaneous: Multi-dimensional representations of behaviors and context in sensory areas. *Neuron*, *110*(19), 3064-3075. https://doi.org/10.1016/j.neuron.2022.06.019

Chen, C. S., Mueller, D., Knep, E., Ebitz, R. B., & Grissom, N. M. (2024). Dopamine and Norepinephrine Differentially Mediate the Exploration–Exploitation Tradeoff. *The Journal of Neuroscience*, *44*(44), e1194232024. https://doi.org/10.1523/JNEUROSCI.1194-23.2024

Desbernats, A., Martin, E., & Tallet, J. (2023). Which factors modulate spontaneous motor tempo? A systematic review of the literature. *Frontiers in Psychology*, *14*, null. https://doi.org/10.3389/fpsyg.2023.1161052

Gould, S. J. (1991). Exaptation: A Crucial Tool for an Evolutionary Psychology. *Journal of Social Issues*, *47*(3), 43-65. https://doi.org/10.1111/j.1540-4560.1991.tb01822.x

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431-440. https://doi.org/10.1037/a0027373

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1), 46-54. https://doi.org/10.1016/j.tics.2014.10.004

Markowitz, J. E., Gillis, W. F., Jay, M., Wood, J., Harris, R. W., Cieszkowski, R., Scott, R., Brann, D., Koveal, D., Kula, T., Weinreb, C., Osman, M. A. M., Pinto, S. R., Uchida, N., Linderman, S. W., Sabatini, B. L., & Datta, S. R. (2023). Spontaneous behaviour is structured by reinforcement without explicit reward. *Nature*, *614*(7946), 108-117. https://doi.org/10.1038/s41586-022-05611-2

Maye, A., Hsieh, C., Sugihara, G., & Brembs, B. (2007). Order in Spontaneous Behavior. *PLoS ONE*, *2*(5), e443. https://doi.org/10.1371/journal.pone.0000443

Nassar, M. R. (2024). Toward a computational role for locus coeruleus/norepinephrine arousal systems. *Current Opinion in Behavioral Sciences*, *59*, 101407. https://doi.org/10.1016/j.cobeha.2024.101407

R Core Team. (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Roesch, M. R., Esber, G. R., Li, J., Daw, N. D., & Schoenbaum, G. (2012). Surprise! Neural correlates of Pearce-Hall and Rescorla-Wagner coexist within the brain. *The European Journal of Neuroscience*, *35*(7), 1190-1200. https://doi.org/10.1111/j.1460-9568.2011.07986.x

Romano, S. A., Pietri, T., Pérez-Schuster, V., Jouary, A., Haudrechy, M., & Sumbre, G. (2015). Spontaneous Neuronal Network Dynamics Reveal Circuit’s Functional Adaptations for Behavior. *Neuron*, *85*(5), 1070-1085. https://doi.org/10.1016/j.neuron.2015.01.027

Stojić, H., Orquin, J. L., Dayan, P., Dolan, R. J., & Speekenbrink, M. (2020). Uncertainty in learning, choice, and visual fixation. *Proceedings of the National Academy of Sciences*, *117*(6), 3291-3300. https://doi.org/10.1073/pnas.1911348117

Sutton, R. S., & Barto, A. (2020). *Reinforcement learning: An introduction* (Second edition). The MIT Press.

Viswanathan, G. M., Buldyrev, S. V., Havlin, S., da Luz, M. G., Raposo, E. P., & Stanley, H. E. (1999). Optimizing the success of random searches. *Nature*, *401*(6756), 911-914. https://doi.org/10.1038/44831