¿Es el principio de la energía libre una teoría normativa o descriptiva de la cognición?

Is the Principle of Free Energy a Normative or Descriptive Theory of Cognition?

Eduardo A. Aponte[◈]
Universidad de Zúrich e Instituto Tecnológico de Suiza (Zúrich)

Recibido: 2014-07-19
Envío a pares: 2014-08-08
Aprobado por pares: 2014-09-10
Aceptado: 2014-10-23

Pensamiento y Cultura | ISSN: 0123-0999 | eISSN: 2027-5331
pensam.cult | Vol. 18-1 | Junio de 2015 | pp. 6-45
DOI: 10.5294/pecu.2015.18.1.1

◈ aponteeduardo@gmail.com

¿Es el principio de la energía libre una teoría normativa o descriptiva de la cognición?

Resumen: las últimas dos décadas han visto un resurgimiento de la estadística bayesiana, la cual fue vista como una disciplina marginal durante la mayor parte del siglo XX. Este fenómeno ha tenido un profundo efecto en la neurociencia, no solo en cuanto al tipo de métodos usados para analizar datos experimentales, sino también en la forma en que la percepción y la acción son conceptualizadas desde un punto de vista teórico. Este giro puede ser resumido en la hipótesis bayesiana del cerebro, según la cual una de las funciones centrales de este órgano es realizar inferencias estadísticas bayesianas. En este contexto, el principio de la energía libre, propuesto por Karl Friston, ha surgido como un posible candidato a una teoría unificada de la cognición. Son dos los propósitos de este artículo: primero presentar el principio de la energía libre desde una perspectiva filosófica y segundo aclarar si este principio debe ser visto como una teoría normativa de la cognición o si, al contrario, este puede realizar predicciones empíricas acerca del tipo de procesos computacionales que caracterizan a la cognición humana. En conclusión, el principio de la energía libre, como es frecuentemente presentado por Friston, corresponde a una teoría descriptiva del tipo de algoritmos computacionales implementados por el cerebro. Más aún, no hay todavía suficiente evidencia empírica en su favor y sí un gran número de hallazgos que apuntan en la dirección contraria.

Palabras clave: neurociencia; energía libre; Bayes; estadística; percepción; cognición.

Is the Principle of Free Energy a Normative or Descriptive Theory of Cognition?

Abstract: The last two decades have witnessed a resurgence of Bayesian statistics, which was regarded as a marginal discipline during most of the twentieth century. This phenomenon has had a profound effect on neuroscience, not only in terms of the kinds of methods used to analyze experimental data, but also in the way perception and action are conceptualized from a theoretical standpoint. This shift can be summarized in the Bayesian brain hypothesis, which holds that one of the central features of this organ is to mount Bayesian statistical inferences. In this context, the principle of free energy proposed by Karl Friston has emerged as a possible candidate for a unified theory of cognition. The purpose of this article is twofold. The first is to introduce the principle of free energy from a philosophical perspective; the second is to clarify whether this principle should be seen as a normative

theory of cognition or if, on the contrary, it can be used to make empirical predictions about the sort computational processes that characterize human cognition. In conclusion, the principle of free energy, as often presented by Friston, is a descriptive theory on the type of computational algorithms the brain uses. Moreover, there is not enough empirical evidence in its favor and, in fact, a large number of findings point in the opposite direction.

Keywords: neuroscience; free energy; Bayes; statistics; perception; cognition.

***

Introducción

En las últimas décadas, la estadística bayesiana ha tomado un rol cada vez más central en diferentes campos de la ciencia. Sorprendentemente, el papel de este grupo de métodos no se ha limitado al análisis de datos experimentales, sino que también ha generado interés en la ciencia cognitiva (Griffiths, Kemp y Tenenbaum, 2008) e indirectamente en la filosofía de la mente (Clark 2013; Jones y Love 2011). La razón es que la estadística bayesiana propone una definición concreta y formal, pero igualmente flexible de qué es una inferencia (Griffiths, Kemp y Tenenbaum, 2008). Es precisamente la simplicidad y elegancia de esta definición la que ha permitido modelar formalmente una plétora de fenómenos cognitivos, psicológicos y psiquiátricos.

Uno de los representantes más prominentes en este campo es Karl Friston (Friston, Karl, Kilner y Harrison 2006; Friston y Stephan 2007; Friston 2010). Su propuesta ha recibido gran atención no solo en la neurociencia computacional, sino también en la filosofía (Clark 2013), la ciencia cognitiva (FitzGerald et al. 2015), la fisiología (Bastos et al. 2012) y la psiquiatría (Montague et al. 2012). En particular, partiendo de una teoría bayesiana de la inferencia, Friston ha propuesto lo que considera una teoría unificada del cerebro, basada en principios matemáticos generales:

> The principle of variational free energy minimization has […] been proposed to explain the ability of complex systems like the brain to resist a natural tendency to disorder and maintain a sustained

> homoeostatic exchange with its environment. Since that time, the free energy principle has been used to account for a variety of phenomena in sensory, cognitive and motor neuroscience and has provided useful insights into structure-function relationships in the brain. […] These formulations provide an important link between information theory (in the sense of statistical thermodynamics) and general formulations of adaptive agents in terms of utility theory and optimal decision theory (Friston 2012).

El principio de la energía libre ha sido presentado por Friston como un marco conceptual que ofrece una imagen unificada de la función del cerebro. En este artículo, mi objetivo es hacer una presentación crítica del principio de la energía libre desde una perspectiva filosófica y darlo a conocer a la comunidad filosófica colombiana.

Antes de iniciar mi presentación, es necesario hacer varias salvedades respecto de esta tarea. Primero, me limitaré a discutir el principio de la energía libre como una teoría de la percepción y la inferencia inductiva, lo cual implica que no me ocuparé de su interpretación más general, según la cual esta teoría define las propiedades básicas de agentes adaptativos. Más particularmente, no me ocuparé del problema de cómo el principio de la energía libre puede ser entendido en el contexto de la teoría de la decisión.

Una segunda salvedad es que uno de mis objetivos fundamentales aquí es mostrar los presupuestos conceptuales del principio de la energía. En particular, trataré de separar los elementos de esta teoría, que deben ser considerados puramente normativos de aquellos que son descriptivos, es decir, aquellos elementos que pueden ser objeto de verificación empírica. Más adelante, definiré en detalle los conceptos de teoría normativa y descriptiva.

Una última salvedad es que, si bien mi presentación está dirigida a una audiencia general, no evadiré conceptos matemáticos formales. Esto es importante pues un entendimiento profundo del principio de la energía libre requiere capturar el formalismo que le da sentido. Sin embargo, la mayor parte de conceptos usados aquí serán explicados de una manera didáctica.

El presente artículo está organizado de la siguiente manera. Primero, me enfocaré en presentar lo que ha venido a llamarse la hipótesis bayesiana del cerebro (Knill y Pouget 2004). Mi objetivo será demostrar que esta hipótesis debe ser entendida, en primera instancia, como una teoría normativa de las inferencias inductivas que cualquier agente inteligente debe realizar. Enseguida trataré de presentar el concepto de 'energía libre' desde una perspectiva termodinámica, para, a continuación, demostrar su relación con la teoría bayesiana de la inferencia y su relación con la teoría de la percepción. Para concluir, trataré de demostrar que hay una tensión entre dos posibles interpretaciones de esta teoría: una interpretación normativa y una algorítmica o representacional de la inferencia. También discutiré brevemente algunos de los méritos de ambas interpretaciones.

La conclusión central de este artículo es que la presentación de Friston sugiere interpretar el principio de la energía libre como una teoría empírica del tipo de computaciones realizadas por el cerebro. El problema central de esta posición es que, si por un lado carece de evidencia empírica directa, por otro, hay evidencia empírica en su contra. A pesar de esto, el principio de la energía libre puede ser entendido como una teoría normativa de la cognición, que da cuenta del tipo de limitaciones computacionales que la teoría bayesiana del cerebro debe explicar.

En la siguiente sección, presentaré el marco general en el cual el principio de la energía libre puede ser expresado: la hipótesis bayesiana del cerebro.

La hipótesis bayesiana del cerebro

La teoría estadística bayesiana tiene su origen en la formulación y aplicación de un sencillo teorema que se deriva de los axiomas generales de la probabilidad (por ejemplo en la axiomatización de Kolmogorov). Este teorema fue inicialmente formulado por el sacerdote inglés Thomas Bayes en el siglo XVIII, pero sus implicaciones formales fueron ignoradas en gran parte del siglo XX debido a consideraciones filosóficas y epistemológicas que no serán discutidas en el presente escrito. Sin

embargo, como mencioné, el interés hacia esta teoría ha crecido drásticamente en las últimas décadas. Parte de este resurgimiento se ha reflejado en el creciente interés en utilizar el lenguaje y los métodos de esta teoría para explicar un gran número de procesos cognitivos. Combinar la teoría estadística bayesiana y la ciencia cognitiva ha conducido a la hipótesis bayesiana del cerebro, la cual puede ser resumida bajo la tesis de que una de las funciones fundamentales del cerebro es realizar inferencias bayesianas.

Inferencias bayesianas

Antes de discutir la teoría bayesiana del cerebro, es necesario reconstruir el concepto bayesiano de inferencia. Para ello, utilizaré un ejemplo proveniente de la literatura psicológica que demuestra de manera didáctica el concepto de 'inferencia'. Se trata originalmente (Phillips y Edwards 1966) de un experimento usado en una población de pacientes esquizofrénicos, que ha sido replicado y extendido en un gran número de ocasiones desde su publicación original (Fine et al. 2007). En lo que resta de esta sección exploraré este ejemplo en detalle.

La premisa de este experimento[1] es que en una habitación hay un gran número de urnas, de las cuales tres cuartas partes son de color rojo y el resto de color azul. En cada urna hay fichas rojas y azules. Las urnas rojas contienen 80 % de fichas rojas, y las urnas azules contienen 80 % de fichas azules. Los participantes son informados de las proporciones de urnas rojas y azules y del número de fichas de cada color en cada tipo de urna. En este experimento, un número de fichas son extraídas de una urna elegida aleatoriamente, permaneciendo la urna oculta a los participantes. Estos son, entonces, informados del color de las fichas que han sido extraídas, después de lo cual deben inferir el color de la urna de la cual las fichas provienen.

Este tipo de inferencias se caracterizan por ser *inductivas* y *probabilísticas*, es decir, más de una respuesta tiene una probabilidad mayor a cero de ser correcta. La teoría estadística bayesiana es una respuesta

1 En el experimento original (Phillips y Edwards 1966), la tarea usada es ligeramente distinta de la versión presentada aquí.

a la pregunta por cuál es la inferencia correcta por realizarse en este contexto, esto es, cuál es la probabilidad de que cierta hipótesis sea correcta *dada* cierta información empírica.

Consideremos por un momento el problema que debe ser resuelto en nuestro ejemplo: la tarea de los participantes consiste en inferir la probabilidad de que la urna sea de cierto color dado el color de las fichas obtenidas de esta y la distribución de las urnas. Una vez la probabilidad de que la urna sea de cierto color ha sido calculada, basta "adivinar" que el color de la urna es aquel que tiene la mayor probabilidad de ser correcto. Este último paso —decidirse por una de las opciones— es objeto de una segunda teoría —la teoría bayesiana de la decisión—. Aquí me limitaré a considerar el problema de inferir distribuciones de probabilidad e ignoraré el problema de cómo estas dan lugar a decisiones (Robert 2007).

La inferencia necesaria para resolver este problema puede ser formalizada de la siguiente manera. Primero podemos definir la variable randomizada (*random variable*) U que tiene como espacio de muestreo (outcome space), el conjunto {0,1}. La hipótesis U=0 representa la posibilidad de que la urna sea de color rojo, y la hipótesis U=1 representa la posibilidad opuesta. La probabilidad de que U=0 antes de observar una ficha es igual a

$$p(U=0)=0.75. \quad (1)$$

Esta probabilidad es llamada la probabilidad *a priori*, ya que su valor es independiente de las fichas extraídas, esto es, de las observaciones hechas.

Ahora consideremos una serie de variables randomizadas $F_1,\ldots, F_N$ que corresponden a las observaciones hechas, esto es, al color de las fichas que han sido observadas. El espacio de muestreo de estas variables randomizadas es otra vez el conjunto {0,1}, donde F_n=0 corresponde al caso de que la ficha n sea de color rojo, y F_n=1 al caso de que esta ficha sea de color azul. Por simplicidad, podemos asumir que la probabilidad del color de una ficha depende solo del color de la urna de la cual ha sido extraída, de tal manera que la probabilidad de que una ficha roja haya sido obtenida de una urna roja es 0.8 y la probabilidad de que una ficha roja provenga de una urna azul es 0.2 (esto es, las fichas extraídas son reemplazadas después de ser extraídas). La relación entre observaciones

e hipótesis está definida por la función de verosimilitud (likelihood function) que establece la probabilidad condicional de una observación dada cierta hipótesis. En nuestro caso, el espacio de muestreo de U corresponde al espacio de las hipótesis posibles. La función de verosimilitud es la probabilidad de que una ficha de cierto color (azul o rojo) sea obtenida de una cierta urna, y puede ser definida de la siguiente manera:

$$p(F=0 \mid U=1)=0.2: \tag{2}$$

Probabilidad de una ficha roja dada una urna azul.

$$p(F=0 \mid U=0)=0.8: \tag{3}$$

Probabilidad de una ficha roja dada una urna azul.

El teorema de Bayes establece la relación entre la probabilidad *a priori*, la función de verosimilitud y la probabilidad *a posteriori*, esto es, la probabilidad condicional de una hipótesis dada una serie de observaciones:

$$p(U \mid F) = \frac{(p(F \mid U)\, p(U))}{(\Sigma_u p(F \mid U=u)\, p(U=u))} \tag{4}$$

El término a la izquierda de la igualdad corresponde a la probabilidad *a posteriori*. El dominador en el término de la derecha no depende de la hipótesis y corresponde a una constante de normalización que garantiza que la probabilidad *a posteriori* esté bien definida. El teorema de Bayes establece que la probabilidad de una hipótesis dada una serie de observaciones es proporcional al producto de la probabilidad *a priori* y la función de verosimilitud.

En nuestro ejemplo, la probabilidad *a posteriori* es la probabilidad de que la urna sea de color rojo, dada una serie de observaciones.

$$p(U \mid F_1,\ldots,F_N)=\frac{}{z}\, p(F_1,\ldots, F_N \mid U)\, p(U), \tag{5}$$

$$z = \sum_{z} p(F_1,\ldots, F_N \mid U=u)\, p(U=u), \tag{6}$$

donde z es una constante de normalización que garantiza que $\Sigma_u\, p(U = u \mid F_1,\ldots, F_n)=1$.

Para concretizar nuestro ejemplo, imaginemos ahora que tres fichas han sido extraídas de una urna. Dos fichas son de color azul y una de color rojo. La función de verosimilitud es igual al producto de la probabilidad de cada ficha por separado, de tal manera que la probabilidad condicional es igual a:

$$p\left(U|F_1=0,F_2=1,F_3=1\right)=\frac{1}{z}p\left(F_1=0|U\right)\quad p\left(F_2=1|U\right)\quad p\left(F_3=1|U\right)\quad p\left(U\right) \tag{7}$$

$$p\left(U=0|F_1=0,F_2=1,F_3=1\right)=\frac{1}{z}0.8\quad 0.2\quad 0.2\quad 0.75=\frac{1}{z}0.024 \tag{8}$$

$$p\left(U=1|F_1=0,F_2=1,F_3=1\right)=\frac{1}{z}0.2\quad 0.8\quad 0.8\quad 0.25=\frac{1}{z}0.032 \tag{9}$$

$$z=0.024+0.032=0.056 \tag{10}$$

En este caso, la probabilidad de que la urna se de color azul es aproximadamente 0.57.

Este pequeño ejemplo demuestra tres de los postulados centrales de la estadística bayesiana. Primero, el objetivo de la inferencia bayesiana es obtener la probabilidad *a posteriori* de una hipótesis o variable latente. Las variables latentes corresponden a variables randomizadas que no han sido observadas directamente. En nuestro ejemplo, la variable latente corresponde al color de la urna de la cual las fichas han sido extraídas. Inferir el color de la urna no es otra cosa que calcular la probabilidad *a posteriori* de que esta sea de un color u otro.

Segundo, la teoría bayesiana de la inferencia asume que información a priori es usada para realizar inferencias. En nuestro ejemplo, la información a priori es la información sobre la proporción de urnas rojas y azules en la habitación en la cual estas se encuentran. El concepto de 'probabilidad *a priori*' es a veces presentado como el conjunto de creencias sobre una hipótesis antes de obtener cualquier tipo de evidencia respecto de esta. El teorema de Bayes ofrece entonces una explicación de cómo las creencias acerca de una hipótesis (o variable latente) deben cambiar después de recolectar evidencia acerca de esta (Griffiths, Kemp y Tenenbaum, 2008). La noción de 'información *a priori*' puede ser relajada para explicar el origen de estas creencias: estas pueden ser formadas

mediante un gran número de experiencias anteriores. El punto clave es que a priori es aquí un término relacional: una creencia es *a priori* respecto de una serie de observaciones si esta creencia no depende (en el sentido probabilístico) de tales observaciones.

El último postulado fundamental de la teoría bayesiana es que toda inferencia se basa en un modelo generativo de cómo las observaciones son producidas por las variables latentes o hipótesis. El modelo generativo simplemente está compuesto de la probabilidad *a priori* y la función de verosimilitud. El adjetivo generativo proviene de que estos modelos establecen una serie de dependencias entre observaciones e hipótesis acerca de estas.

La hipótesis bayesiana del cerebro postula que el tipo de inferencias inductivas que cualquier agente inteligente debe realizar son, al menos de forma aproximativa, inferencias bayesianas (Kersten, Mamassian y Yuille, 2004). Esta proposición tiene dos implicaciones directas: para realizar inferencias bayesianas es necesario representar, al menos de manera aproximativa, una probabilidad *a priori* y una función de verosimilitud, esto es, un modelo generativo (Knill y Pouget 2004). La segunda implicación es que el agente debe estar en capacidad de, al menos aproximadamente, calcular y representar la probabilidad *a posteriori* de la hipótesis sobre la cual desea realizar una inferencia o, al menos, representar alguna estadística respecto de esta distribución, como su valor esperado o varianza.

Es posible imaginar cómo este concepto de 'inferencia' puede ser traducido en una teoría de la percepción (Kersten, Mamassian y Yuille, 2004). En oposición a la teoría de la percepción directa (Gibson 1978), la hipótesis bayesiana del cerebro afirma que percibir es equivalente a inferir las causas de la evidencia que es disponible a un organismo a través de sus órganos perceptuales. En este contexto, las posibles causas de una observación son simplemente las hipótesis o variables latentes postuladas en el modelo generativo.

Esta teoría de la percepción ha tenido un gran éxito experimental en la última década, en la cual una gran diversidad de estudios ha mostrado que ciertas leyes psicofísicas pueden ser explicadas como el resultado de inferencia bayesiana (Knill y Pouget 2004). Por ejemplo,

Petzschner y Glasauer (2011) probaron que si sujetos son expuestos a un gran número de estímulos, estos tienden a subestimar la magnitud de estímulos que demuestran grandes desviaciones del promedio. Esto es, después de percibir un gran número de objetos pequeños, un objeto de gran tamaño tiende a ser percibido como más pequeño que su tamaño real. Los autores de dicho estudio interpretaron este resultado como la consecuencia de creencias *a priori* desarrolladas mediante la exposición a estímulos anteriores. Sorprendentemente, estas desviaciones pueden ser modeladas con gran exactitud utilizando un modelo de inferencia bayesiano. Más aún, este tipo desviaciones no se limitan a la estimación del tamaño, sino que pueden ser observadas en diferentes paradigmas experimentales que incluyen la estimación consecutiva de propiedades físicas, como distancia o peso. En una sección posterior del texto, consideraré posibles contraargumentos empíricos en contra de la teoría bayesiana del cerebro.

Valga resaltar que la teoría bayesiana del cerebro es fundamentalmente una teoría sobre inferencias confirmativas *e inductivas* realizadas por el cerebro, esto es, una teoría sobre inferencias que dan lugar a conclusiones basada en información empírica, opuestas a *deducciones* realizadas a partir de premisas. Más aún, no todas las inferencias empíricas pueden ser formalizadas usando el lenguaje formal que subyace a la teoría bayesiana de la inferencia. Un ejemplo emblemático es el de *inferencias causales*, pues no es posible capturar en el lenguaje puramente bayesiano la idea de que si A causa B, intervenir experimentalmente en A debe tener un efecto en B, mas intervenir en B no debe tener un efecto en A. La razón es que, en ciertos contextos, la noción de 'intervención experimental' no puede ser capturada por el concepto bayesiano de 'probabilidad condicional'. Judea Pearl (2000) ofrece un examen cuidadoso de por qué estos conceptos no pueden ser expresados satisfactoriamente en la teoría bayesiana de la inferencia.

Como nota final hay que observar que la teoría bayesiana del cerebro no debe ser entendida como una teoría del razonamiento científico que afirma que argumentos de carácter científico son fundamentalmente argumentos probabilísticos. Según esta posición, que autores como Clark Glaymour (1981) identifica con la teoría de la confirmación

bayesiana, el tipo de argumentos que dan base a teorías científicas están basados en la probabilidad condicional de una hipótesis dadas ciertas observaciones empíricas. La teoría bayesiana del cerebro no es una teoría del razonamiento científico, mas sí una teoría de cómo creencias surgen a partir de observaciones desde una perspectiva *personal*, esto es, la teoría bayesiana del cerebro es una teoría del aprendizaje desde la perspectiva personal de un agente cognitivo. En otras palabras, la teoría bayesiana del cerebro no es una explicación histórica o normativa sobre argumentos de carácter científico.

Optimalidad

Volvamos por un momento al ejemplo de las urnas. La tarea consiste ahora no en inferir cuál es color de la urna, sino en determinar cuál es la cuota (odds) de que la urna sea de cierto color —esto es, la razón (ratio) entre la probabilidad de las dos hipótesis posibles— y realizar una apuesta basada en esta cuota. Es posible probar que no hay ningún sistema de apuestas que en promedio obtenga mayores ganancias que aquel basado en la probabilidad *a posteriori* de las urnas. En términos más generales, la inferencia bayesiana es el sistema óptimo para determinar la probabilidad de un evento en el sentido de que ningún otro tipo de inferencia puede en promedio ofrecer mejores ganancias en este tipo de apuestas (Freedman y Purves 1969). Esta noción de optimalidad fue propuesta inicialmente por Ramsay (1926-1964) y desarrollada por Carnap (1950). Según el argumento de estos autores, en condiciones idealizadas, un agente racional debe siempre elegir la opción que maximice su utilidad esperada, lo cual implica que, asumiendo que la noción de 'racionalidad' puede ser equiparada (en condiciones idealizadas) con la maximización de la utilidad esperada en un problema de decisión, solo la teoría bayesiana de la inferencia puede ofrecer un esquema de inferencia que satisfaga este concepto de racionalidad. Valga añadir que, incluso, Ramsay vislumbró las posibles objeciones que pueden ser hechas a este argumento: en condiciones no idealizadas, un sin número de factores independientes a la utilidad esperada afectan las decisiones de un agente sin que por esto consideremos que estas sean fundamentalmente irracionales. El punto aquí es que esta teoría es normativa y se basa en dos

postulados fundamentales: 1) la única variable que debe ser optimizada es la utilidad esperada y 2) el agente tiene recursos infinitos (en términos de tiempo y memoria) para resolver este tipo de problemas.

Dejando a un lado las posibles deficiencias de esta definición de racionalidad, es esta probablemente la piedra fundacional del argumento central para utilizar la estadística bayesiana para modelar la cognición humana: la estadística bayesiana ofrece un marco normativo de qué es una inferencia inductiva correcta. En este punto es necesario clarificar la noción de 'normativa' o, si se quiere, norma epistémica. Aquí seguiré la definición de John Pollock (1987):

> Norms are general descriptions of the circumstances under which various kinds of normative judgments are correct. Epistemic norms are norms describing when it is epistemically permissible to hold various beliefs. A belief is justified iff it is licensed by correct epistemic norms. Assuming that what justifies a belief is the reasoning underlying it ("reasoning" constructed broadly), epistemic norms are norms governing "right reasoning" (Pollock 1987, 81).

El concepto bayesiano de 'inferencia' es normativo en cuanto define qué tipo de inferencias son admisibles bajo ciertas premisas. Como mencioné, es posible demostrar que —aceptado el concepto de 'racionalidad' propuesto por Ramsay y Carnap— no hay ninguna otra definición de inferencia que pueda ser considerada racional. Consecuentemente, la inferencia bayesiana es óptima con respecto a cualquier otra definición normativa de inferencia inductiva en contextos probabilísticos. La hipótesis bayesiana del cerebro simplemente afirma que una de las funciones centrales del cerebro es realizar este tipo de inferencias, al menos de manera aproximativa. Esta última salvedad es importante porque revela el carácter normativo de esta hipótesis: como veremos más adelante, hay razones formales para asumir que en la mayor parte de condiciones el cerebro solo puede aproximar el tipo de inferencias que la hipótesis bayesiana del cerebro postula como óptimas.

En oposición a una teoría normativa, una teoría descriptiva caracteriza cómo el cerebro humano realiza inferencias inductivas, indepen-

dientemente de si estas son correctas o no bajo cierta normatividad. Por tanto, este tipo de teorías son empíricas y pueden o no ser falsificadas. En este sentido, la distinción entre una teoría normativa y una teoría descriptiva de la cognición es análoga a la misma distinción hecha en la ética, donde cómo se deben actuar y cómo se actúa en la realidad son preguntas independientes.

Ahora bien, determinar cuáles inferencias deben ser realizadas por un agente inteligente no implica ningún algoritmo o implementación biológica de los procesos necesarios para realizar tales inferencias. No habiendo ninguna restricción respecto del tipo de implementaciones que pueden ser usadas, es difícil imaginar cómo es posible hacer predicciones empíricas basadas directamente en la hipótesis bayesiana del cerebro. Esta observación ha sido formulada de manera más pesimista por Spratling (2013, 232):

> Bayes' theorem states that the posterior is proportional to the product of the likelihood and the prior. However, it places no constraints on how these probabilities are calculated. Hence, any model that involves multiplying two numbers together, where those numbers can be plausibly claimed to represent the likelihood and [prior], can be passed off as a Bayesian model. This has led to numerous computational models which lay claim to derive "probabilities" that are as ad-hoc and unprincipled as the non-Bayesian model they claim superiority over.

La posición de Spratling es una manera crítica de resumir el argumento que he bosquejado: una teoría bayesiana de la inferencia no ofrece ninguna restricción acerca de cómo el cerebro realiza inferencias. Esta teoría ofrece solo un marco normativo del tipo de inferencia que el cerebro debe realizar.

El carácter normativo de la hipótesis bayesiana del cerebro es aún más patente si se tiene en cuenta que debemos aceptar, por principio, que las inferencias realizadas por seres humanos se desvían necesariamente de la definición bayesiana de inferencia, lo cual se origina en que, en la mayor parte de casos, calcular una probabilidad *a posteriori* es prohibitivamente difícil. Hay al menos dos razones para esto.

La primera razón está relacionada con la constante de normalización z mencionada anterioremente. Esta cantidad consiste en la probabilidad marginal de un modelo, esto es, la probabilidad de los datos empirícos después de integrar todas las variables latentes. Si esta integral no puede ser resuelta de forma analítica —lo cual es comúnmente el caso—, realizar inferencia bayesiana de manera exacta no es posible. Otra forma de presentar esta dificultad es considerar casos en que las variables latentes tienen un espacio de muestreo finito. Por ejemplo, si las variables randomizadas asumen valores binarios (cero o uno), el número total de realizaciones posibles de las variables randomizadas crece exponencialmente con el número de variables n. En el peor de los casos, para poder representar la distribución *a posteriori* es necesaria una cantidad de memoria proporcional a 2^n; para marginalizar esta distribución, $n2^n$-1 sumas son necesarias. Esta es entonces la segunda razón por la cual es imposible realizar inferencias bayesianas perfectas: no solo es necesario realizar un número exponencial de multiplicaciones, sino que también es necesario representar un número exponencial de valores.

Este circunstancia lleva nuevamente a la dicotomía mencionada: si bien una teoría bayesiana del cerebro ofrece una definición normativa de que es una inferencia empírica, esta no puede implicar ningún algoritmo particular para realizar tal inferencia. La hipótesis bayesiana del cerebro deja sin resolver —por su naturaleza puramente normativa— dos problemas computacionales: cómo representar funciones de probabilidad y cómo utilizar estas representaciones para realizar inferencias.

Esta dicotomía puede ser presentada en términos de los niveles explicativos propuestos por Marr y Poggio (1976): la hipótesis bayesiana del cerebro es una teoría computacional del cerebro que se ocupa de formalizar los problemas que deben ser resueltos por el cerebro. En particular, la hipótesis bayesiana define qué es una inferencia empírica correcta y, por tanto, qué problemas debe resolver un agente inteligente ideal o, en palabras de Putnam (1964), una máquina de aprendizaje. Cómo estos problemas puedan ser resueltos corresponde al nivel algorítmico y representacional de la jerarquía propuesta por Marr.

En el contexto de la teoría bayesiana del cerebro, Friston (2007) ha propuesto en la última década lo que esta ha denominado el principio

de la energía libre, tomando como base descubrimientos en el campo de la neurociencia, la estadística, la teoría de la información y, más generalmente, un entendimiento cada vez más profundo de la relación de estas disciplinas con la termodinámica. La pregunta que trataré de resolver en el resto de este artículo es si el principio de la energía libre debe ser considerado como una teoría puramente normativa de la inferencia o si, al contrario, esta es una teoría que se sitúa en el nivel algorítmico y representacional. En la siguiente sección, introduciré el concepto de 'energía libre' utilizando el concepto termodinámico correspondiente y a continuación mostraré su relación con el concepto bayesiano de inferencia. Seguidamente presentaré dos variantes interpretativas del principio de la energía libre.

Energía libre: termodinámica e inferencia

Para introducir el concepto de 'energía libre', su relación con la teoría de la inferencia, y mostrar cómo este ha sido usado para formular un principio general de la cognición, me basaré en la presentación de Ortega y Braun (2013) y Feynman (1998). Consideremos el sistema físico en la figura 1. En este sistema, un gas ideal sumergido en un foco calórico a temperatura T se encuentra en un cilindro. Este gas es comprimido por un pistón desde el estado A hasta alcanzar el estado. El volumen del cilindro en estado A y B es V_A y V_B respectivamente. Para simplificar la presentación, podemos asumir que una única partícula se encuentra en la cámara y que se trata de un proceso isotérmico (i. e., la temperatura T permanece constante). La energía interna del sistema es igual en las configuraciones A y B, ya que la energía interna de un gas depende solo de su temperatura.

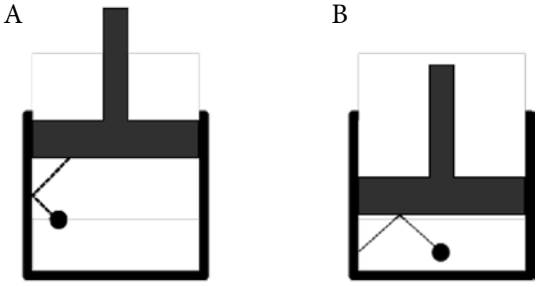


Figura 1. A. Sistema inicial. Un gas ideal (compuesto de una única partícula) en un foco calórico (*thermal bath*) a temperatura *T*. El gas se encuentra en una cámara, donde una de las paredes es operada por un pistón. **B.** El gas después de un ser comprimido de manera isotérmica. El cambio en energía libre entre ambos estados es proporcional a la diferencia del logaritmo del volumen de la cámara antes y después de la compresión. Al disminuir el volumen de la cámara, las locaciones que la partícula puede asumir disminuyen, incrementado la información acerca de la posición de esta.

La definición formal de la energía libre de Helmholtz es la diferencia entre la energía interna de un sistema y el producto de la temperatura y la entropía:

$$F=U-TS, \tag{11}$$

donde F es la energía libre, U es la energía interna del sistema, T la temperatura y S la entropía. Es posible probar que en el caso del sistema presentado, el cambio de energía libre entre el estado A y B es descrito por la ecuación

$$dF = -kT\frac{1}{V}dV \tag{12}$$

$$\int dF = -\int_{V_A}^{V_B} kT\frac{1}{V}dV \tag{13}$$

$$F_B - F_A = -kT \ln V_B / V_A \tag{14}$$

donde k es la constante de Bolzman. En palabras sencillas, el cambio de energía libre es proporcional al cambio del logaritmo del volumen ocupado por el gas en los estados A y B. Dado que U permanece constante, es fácil observar que el cambio de entropía del sistema es proporcional a la diferencia de la energía libre:

$$F_B - F_A = \left(U - TS_B\right) - \left(U - TS_A\right) \propto S_A - S_B \tag{15}$$

También es posible probar que el cambio entre los dos estados requiere la utilización de trabajo; la energía libre (negativa) no es otra cosa que el límite inferior (*lower bound*) del trabajo necesario para comprimir el gas desde el estado A al estado B[4].

Sorprendentemente, hay una relación directa entre este sistema físico y el concepto bayesiano de 'inferencia'. Consideremos una partícula moviéndose aleatoriamente en el sistema en estado A. La probabilidad de que se encuentre en una sección del cilindro x con volumen $V_A(x)$ es equivalente a $V_A(x)/V_A$. Al comprimir el gas, se incrementa la información acerca de la localización de la partícula al reducir el volumen de la región en la cual esta puede encontrarse. Es decir, al comprimir el gas se genera nueva información acerca de la localización de la partícula. Este proceso reduce la ignorancia (o entropía) respecto del sistema. La entropía de un sistema es definida por

$$S = -\sum_i p_i \log p_i \tag{16}$$

donde p_i es la probabilidad de la configuración i del sistema. Esta cantidad es una métrica de la ignorancia o incertidumbre asociada con el sistema. Un ejemplo de esta situación es presentado en figura 2.

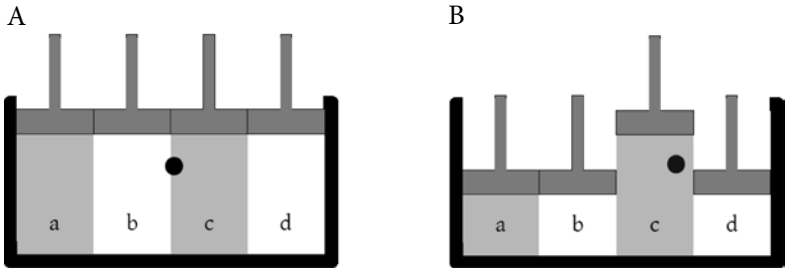


Figura 2. En este ejemplo, la recámara puede ser dividida en cuatro secciones de igual volumen. El volumen de cada división puede ser manipulado de manera independiente por un pisto. Ahora, considérese la probabilidad de que una partícula $\emptyset$ se encuentre en cada uno de los estados o configuraciones posibles *a, b, c, d.* En A la entropía $-S_A$ es igual al valor esperado del logaritmo de la probabilidad de cada una de las configuraciones. Dado que el volumen de cada una de las configuraciones es igual (en equilibrio), la probabilidad de que $\emptyset$ se encuentre en cualquiera de los estados es homogénea, de tal forma que la entropía de A es igual a $S_A = -4\ 0.25\ \ln 0.25 = -\ln 0.25 = 0.6$. En la configuración B asumiendo que $p_B(\theta=a) = p_B(\theta=b) = p_B(\theta=s) = 0.5 \times p_B(\theta=c)$, la entropía S_B es igual a $-3 \times 0.2\ \ln 0.2 - 0.4\ \ln 0.4 = 0.56$. La entropía representa la incertidumbre asociada con un función de probabilidad, en este caso p_A y p_B

Ortega y Braun (2013) formularon la idea de que la transición entre dos estados en equilibrio es una generalización de la noción bayesiana de 'inferencia'; el estado inicial corresponde a la probabilidad *a priori* y el estado final corresponde a la probabilidad *a posteriori*. Para entender este punto, podemos considerar un potencial energético inicial ϕ_0 que define la energía del sistema en la configuración *θ*. En el caso del ejemplo del pistón, una configuración del sistema corresponde a la localización de la partícula y el potencial energético es proporcional al logaritmo del volumen de tal localización. La probabilidad de que el sistema se encuentre en la configuración *θ* —una vez el sistema se encuentra en equilibrio— esz^{-1} exp $-T\phi_0\ (\theta)$[5]. Ahora consideremos un segundo potencial energético ϕ^*, tal que el estado final del sistema está caracterizado por $\phi = \phi_0 + \phi^*$. La interpretación termodinámica de las inferencias bayesianas es asumir que el potencial energético inicial corresponde al logaritmo de la probabilidad *a priori* $p(\theta)$ y el logaritmo de la función de verisimilitud corresponde a ϕ^*, de tal manera que ϕ no es otra cosa que el logaritmo de la probabilidad *a posteriori*. La energía libre es simplemente la mínima cantidad de trabajo que puede ser aplicado para llevar a un sistema desde la configuración ϕ_0 hasta ϕ. La particularidad de la regla de Bayes es que el estado final ϕ es equivalente a $\phi_0 + \phi^*$. y que la temperatura T es igual a 1.

Otra forma de ver este punto, en nuestro ejemplo, es que la posición exacta de la partícula es una variable latente, esto es, la posición de la partícula no puede ser observada directamente. Los dos estados del sistema corresponden a dos distribuciones distintas de esta variable: la distribución *a priori*, y la distribución *a posteriori*. Es importante notar que esta no es una metáfora o simplemente un isomorfismo entre dos teorías abstractas: la interpretación termodinámica de la inferencia establece un límite físico y real de la cantidad de trabajo necesaria para realizar una inferencia. En otras palabras, el concepto bayesiano de 'inferencia' es equivalente al concepto termodinámico de 'reducción de energía libre'. Realizar una inferencia consiste en utilizar trabajo para disminuir la ignorancia promedio respecto de cierta variable randomizada (donde la ignorancia puede ser definida a través del concepto de 'entropía').

Bajo esta simple reflexión, inicialmente expuesta por Feynman y redescubierta y profundizada por Ortega y Braun (2013), hay una profunda intuición que relaciona dos campos de estudio aparentemente disimilares: la termodinámica y la teoría de la inferencia.[2] El punto de conexión es la interpretación de conceptos termodinámicos desde la teoría de la información. La intuición fundamental es que la noción de 'información' y la de 'ignorancia' pueden ser formuladas de manera puramente formal y tienen una contraparte en propiedades físicas que son el objeto de estudio de la termodinámica. Así, no resulta sorprendente que en su exposición Feynman estuviera más interesado en un tratamiento teórico de un problema práctico: ¿cuál es el límite de la eficiencia energética que un transistor (la unidad básica de procesamiento de información inventada por el hombre) puede alcanzar? La energía libre ofrece una respuesta teorética a esta pregunta. Ortega y Braun (2013) dieron el siguiente paso al reconocer que la teoría de la información es también una teoría de la inferencia y, en consecuencia, los principios de la termodinámica deben aplicar también a la cognición humana y la inferencia bayesiana en particular.

La idea de inferencia bayesiana aproximativa cobra aquí un nuevo significado: realizar inferencias requiere trabajo. La razón por la cual algunas de estas no pueden ser exactas en el sentido bayesiano es que en muchas circunstancias un sistema cognitivo no puede acceder a la energía suficiente o no puede usarla de manera eficiente para realizar una inferencia exacta.

Desde la perspectiva de un agente cognitivo, realizar una inferencia consiste entonces en la transición desde ciertas creencias iniciales (descritas por las creencias *a priori*) hasta una serie de creencias justificadas por información empírica (descritas por la distribución *a posteriori*). Este proceso requiere la aplicación de trabajo, esto es, un cambio en la energía libre del sistema. La conclusión aquí es que es posible reinterpretar la hipótesis bayesiana del cerebro como afirmando que

2 Jaynes (1957) ya había notado la relación entre la teoría de la información, la termodinámica y el concepto de 'inferencia'. Sin embargo, el objetivo de esta teoría es fundamentar la termodinámica desde una teoría de la inferencia.

la función del cerebro es disminuir la energía libre de *un* sistema que está compuesto de las representaciones subjetivas de un agente inteligente. La regla de Bayes estable cuál es el estado final del sistema, una vez nueva evidencia empírica es tomada en cuenta.

Esta interpretación parece ser sugerida en ciertas ocasiones por Friston, Kilner Harrison (2006, 71):

> The free energy principle states that systems change to decrease their free energy. The concept of free-energy arises in many contexts, especially physics and statistics. In thermodynamics, free energy is a measure of the amount of work that can be extracted from a system. It is the difference between the energy and the entropy of a system. […] It is this sort of free energy, which is a measure of statistical probability distributions; we apply to the exchange of biological systems with the world. The implication is that these systems make implicit inferences about their surroundings.

Esta interpretación tiene dos desventajas: si la admitimos, el principio de la energía libre es simplemente una reformulación termodinámica del concepto bayesiano de 'inferencia', lo cual conduce al segundo problema: si bien las dos formulaciones son equivalentes en un nivel formal, presentar el concepto de 'inferencia' en términos termodinámicos oscurece la noción de 'inferencia' que se requiere para modelar el tipo de operaciones inferenciales realizadas por agentes adaptativos. La razón es que el concepto termodinámico de 'energía libre' hace hincapié en que toda inferencia requiere trabajo; la noción bayesiana de 'inferencia' resalta que este proceso es una forma de computación. Más aún, la teoría bayesiana de la inferencia sugiere un lenguaje descriptivo y un concepto claro de 'optimalidad' que parece diluirse en el concepto termodinámico de 'energía libre'. Por ejemplo, el concepto de 'observación empírica' —fundamental en la presentación de la teoría bayesiana de la inferencia— no puede ser expresado de manera explícita en términos termodinámicos. La razón de esto es que en termodinámica no existe el concepto de 'probabilidad condicional' —el cual expresa la relación entre evidencia empírica e hipótesis—. Sin embargo, ya que

una función de densidad de probabilidad es simplemente una función, desde la perspectiva termodinámica el logaritmo de una probabilidad condicional es simplemente interpretado como la energía asociada con una configuración posible de un sistema.

Junto con esta interpretación general (que denominaré interpretación débil) del principio de la energía libre, una interpretación más particular descansa en una serie de métodos computacionales que pueden ser derivados usando ecuaciones provenientes de la termodinámica. En la siguiente sección revisaré en detalle esta posibilidad.

Maximización de la energía libre como un algoritmo inferencial

Anteriormente mostré que, si bien el concepto de 'inferencia' bayesiano puede ser utilizado como un concepto normativo de la inferencia, en la mayoría de los casos este tipo de inferencia no puede ser realizada de manera analítica. Una forma de ver esta limitación es observar, primero, que para realizar una inferencia ciertos recursos cognitivos son necesarios y, segundo, que estos no siempre están disponibles para un agente. Esta observación está estrechamente relacionada con métodos de inferencia aproximativa que han sido vistos en la última década como uno de los posibles algoritmos que el cerebro humano utiliza para realizar inferencias (Gershman y Daw 2012). Estos métodos son usualmente llamados inferencia variacional bayesiana (variational Bayes). En lo que sigue haré una presentación de estos métodos y mostraré cómo han influenciado ciertas perspectivas en el área de la neurociencia computacional. Esta sección tiene un carácter técnico y puede ser ignorada por un lector no interesado en el aspecto formal de la teoría.

El algoritmo variacional de Bayes

En lo que sigue utilizaré θ para referirme a las variables latentes o causas y y para referirme a las observaciónes empíricas. La idea fundamental de la inferencia variacional bayesiana consiste en aproximar la función de probabilidad *a posteriori* $p(\theta|y)$ con una distribución arbitraria $q(\theta)$[7]. Dayan el al. (1995) llamaron a $q(\theta)$ una distribución de reconocimiento,

bajo la premisa de que si el cerebro utiliza un modelo generativo codificado por la distribución conjunta $p(\theta|y)\ p(\theta)$, la distribución $q(\theta)$ es utilizada paralelamente para hacer inferencias aproximativas acerca de la variable randomizada θ. Ahora bien, el algoritmo variacional de Bayes se basa en reformular la diferencia de la energía libre respecto de la distribución de reconocimiento q. Esta derivación requiere solo el teorema de Bayes:

$$\ln p(y) = \ln \frac{p(y|\theta)p(\theta)}{p(\theta|y)} \tag{17}$$

$$= \int q(\theta) \ln \frac{p(y|\theta)p(\theta)}{p(\theta|y)} \frac{q(\theta)}{q(\theta)} d\theta \tag{18}$$

$$= \int q(\theta) \ln \frac{p(y|\theta)p(\theta)}{q(\theta)} d\theta \ - \int q(\theta) \ln \frac{p(\theta|y)}{q(\theta)} d\theta \tag{19}$$

$$Energia\ libre: F[q] = \int q(\theta) \ln \frac{p(y|\theta)p(\theta)}{q(\theta)} d\theta \tag{20}$$

$$Divergencia: KL[q(\theta)|p(\theta|y)] = \int q(\theta) \ln \frac{p(\theta|y)}{q(\theta)} d\theta \ \geq 0 \tag{21}$$

El valor de la ecuación 19 es vincular una distribución arbitraria q con la distribución *a posteriori* $p(\theta|y)$. La intuición fundamental aquí es que es posible aproximar la distribución *a postiori* sin conocerla, al maximizar la energía libre negativa de q :

$$F[q] = \int q(\theta) \ln \frac{p(y|\theta)p(\theta)}{q(\theta)} d\theta = \int q(\theta) \ln p(y|\theta)p(\theta)\, d\theta \ + \int q(\theta) \ln q(\theta)\, d\theta \tag{22}$$

$$Energía\ interna\ U: \int q(\theta) \ln p(y|\theta)p(\theta)\, d\theta \tag{23}$$

$$Entropía\ S: \ - \int q(\theta) \ln q(\theta)\, d\theta \tag{24}$$

Cuatro consideraciones son importantes aquí. Primero $p(y)$ es una constante, lo cual implica que maximizar la energía libre negativa $F[q]$ (ecuación 20) supone minimizar el término que llamé divergencia. La

segunda observación es que la divergencia (ecuación 21) es siempre positiva o igual a cero. Tercero, este término es igual a cero si y solo si $q(\theta) = p(\theta|y)$, es decir, si q es una aproximación perfecta de la probabilidad *a posteriori*. Este término es llamado frecuentemente la divergencia de Kullback-Leibler (*Kullback-Leibler divergence*), porque puede ser visto como una métrica de la distancia entre dos distribuciones, con la salvedad de que no es un operador simétrico y, por tanto, no es una distancia en sentido estricto. La última observación es que, si la energía libre negativa en la ecuación 20 es maximizada con respecto de q, la divergencia entre $q(\theta)$ y $p(\theta|y)$ disminuye. En otras palabras, la diferencia entre ambas distribuciones disminuye en la medida en que la energía libre negativa se incrementa, lo cual implica que, al maximizar la energía libre negativa, la distribución de reconocimiento se acerca a la distribución *a posteriori*.

Las observaciones anteriores suponen que la energía libre en la ecuación 20 siempre es igual o menor a la constante de normalización $p(y)=z$; la diferencia entre la energía libre y la constante de normalización está dada por la divergencia de Kullback-Leibler. La característica central de los métodos variacionales es que reformulan el problema de computar la probabilidad *a posteriori* $p(\theta|y)$ como un problema de optimización, esto es, como maximizar el valor de la función objetiva $F[q]$.

Para entender cómo la ecuación 19 puede llevar a un algoritmo inferencial, podemos considerar un modelo probabilístico que es ampliamente usado en aplicaciones como reconocimiento de voz: el modelo oculto de Markov. Este modelo se usa para dar cuenta de un proceso causal $\theta=\{\theta_1,\ldots,\theta_N\}$ del cual solo tenemos observaciones estocásticas $y_1,\ldots,y_N$.

Para construir un ejemplo sencillo, imagínese que el problema es decodificar un grabación sonora que contiene cinco palabras proveniente de la oración "Santander, salve usted la patria". Ese es precisamente el problema perceptual de inferir palabras —en el sentido lingüístico— a partir de vibraciones sonoras. Cada palabra $\theta_1,\ldots,\theta_5$ está asociada a un fragmento de una grabación $y_1,\ldots,y_5$. La primera observación es que la palabra en la posición n depende solo de la palabra directamente anterior. La segunda observación es que el fragmento sonoro asociado con

una palabra depende solo de la palabra en cuestión. La probabilidad de la palabra *patria* dada la palabra *la* esta codificada por:

$$p(\theta_n = patria | \theta_{n-1} = la). \tag{25}$$

Dado que la grabación está contaminada por ruido estocástico, el objetivo de la inferencia es obtener las probabilidades marginales condicionales de las causas θ:

$$p\,(\theta_1,\ldots,\theta_5 | y_1,\ldots,y_N) \tag{26}$$

En otras palabras, el objetivo es determinar la distribución de las causas $\theta_1,\ldots,\theta_N$ dadas las observaciones $y_1,\ldots,y_N$. En el ejemplo en cuestión, el objetivo es determinar la distribución de cada palabra, con la esperanza de que, por ejemplo, el evento θ_1=*Santander* tenga la más alta probabilidad. La propiedad fundamental de los modelos de Markov ocultos es que la observación y_t depende solo del estado θ_t y que este, a su vez, depende solo del estado θ_{t-1}. Esta propiedad es llamada propiedad de Markov de primer orden. Una representación gráfica de este modelo es dada en la figura 3, donde las conexiones representan condicionamiento en el sentido estadístico. Usando la figura 3, es claro que la probabilidad conjunta del modelo puede ser descompuesta de la siguiente manera.

$$p(\theta_1, \ldots, \theta_N | y_1, \ldots, y_N) = p(\theta_1) \prod_{t=1}^{N-1} p(\theta_{t+1} | \theta_t) \prod_{t=1}^{N} p(y_t | \theta_t) \tag{27}$$

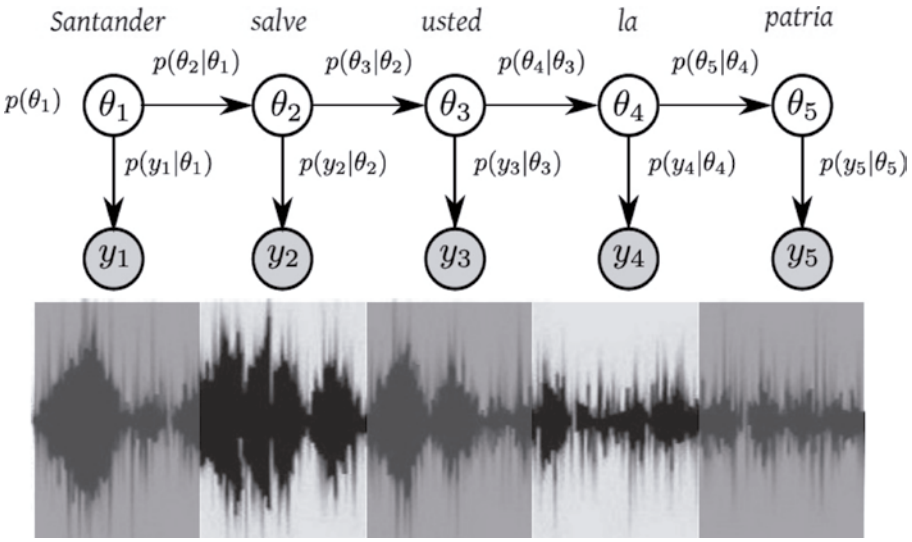


Figura 3. El modelo oculto de Markov. Este modelo asume que un proceso temporal $\theta_1,\ldots,\theta_N$ se desarrolla sin que observaciones directas sobre los estados de este proceso puedan realizarse. Las causas dan origen a observaciones $y_1,\ldots,y_N$. Una de las aplicaciones más comunes de este modelo es el reconocimiento de voz, donde las causas corresponden a palabras, mientras las observaciones corresponden a los fonemas distorsionados por ruido estocástico.

Para calcular la probabilidad condicional de la secuencia "Santander, salve usted la patria", basta evaluar las siguientes probabilidades:

$$p(\theta_1 = General, \ldots, \theta_5 = patria \,|y_1, \ldots, y_5) = \tag{28}$$
$$\frac{1}{z} p(\theta_1 = General) p(y_1|\theta_1 = General) p(\theta_2 = salve|\theta_1 = General)$$
$$p(y_2|\theta_2 = salve) \ldots$$

Para utilizar la ecuación 19 en este contexto, basta observar que la probabilidad $q(\theta)$ toma una forma arbitraria, de tal manera que puede ser expandida en probabilidades independientes:

$$q(\theta) = \prod_i q(\theta_i) \tag{29}$$

Es decir, el conjunto de variables randomizadas θ puede ser dividido en una partición que da como resultado los conjuntos$\theta_1,\ldots,\theta_N$, tal que $\bigcup_{i=1}^{N} \theta_i = \theta$ y para todo i y j si $i \neq j$, entonces $\theta_i \bigcup \theta_j = \theta$. Este método es llamado la aproximación del campo promedio (*mean field approximation*). Esta aproximación simplifica la representación y la inferencia sobre las causas θ al asumir que estas son independientes y que, por tanto, la distribución conjunta puede ser factorizada en diferentes distribuciones marginales. En un momento, volveré a las consecuencias de esta aproximación.

En el ejemplo de un modelo oculto de Markov, podemos hacer una partición de las causas o variables latentes de tal manera que:

$$q(\theta_1, \ldots, \theta_N) = \prod_{t=1}^{5} q(\theta_t)\,. \tag{30}$$

Gracias al teorema fundamental del cálculo de variaciones, se puede demostrar que el máximo de la energía libre negativa con respecto de la distribución $q(\theta_t)$ ocurre cuando:

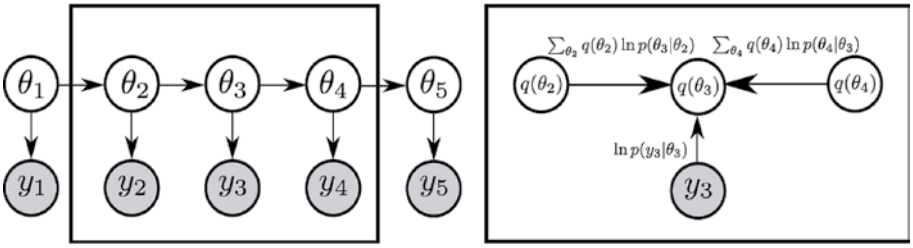
$$\ln q(\theta_i) = \int q(\theta_{I\setminus i}) \ln p(y|\theta) p(\theta)\, d\,\theta_{I\setminus i} - \ln z, \tag{31}$$

donde $\theta_{I\setminus i}$ es el complemento θ_i en θ y z es una constante de normalización. Esta expresión facilita la aproximación de $q(\theta_i)$, lo cual se hace patente cuando consideramos el modelo oculto de Markov (reemplazando integrales por sumas):

$$\ln q(\theta_t) = \sum_{\theta_{t-1}} \sum_{\theta_{t+1}} q(\theta_{t-1}) q(\theta_{t+1}) \ln p(\theta_t|\theta_{t-1}) p(\theta_{t+1}|\theta_{t-1}) p(y_t|\theta_t) \; - \ln z \tag{32}$$

$$\ln q(\theta_t) = \ln p(y_t|\theta_t) + \sum_{\theta_{t-1}} q(\theta_{t-1}) \ln p(\theta_t|\theta_{t-1}) \; + \sum_{\theta_{t+1}} q(\theta_{t+1}) \ln p(\theta_{t+1}|\theta_{t-1}) \; - \ln z \tag{33}$$

La figura 4 demuestra la simplificación del problema inferencial. Para maximizar la energía libre negativa con respecto de $q(\theta_t)$, basta considerar aquellas variables que son ancestros, descendientes directas o ancestros de los descentientes directos de θ_t, esto es, θ_{t-1}, θ_{t+1}, y_t. Este conjunto es denominado la manta de Markov (*Markov blanket*) de θ_t. La aproximación del campo promedio simplifica radicalmente el problema computacional, ya que posibles dependencias estadísticas globales son eliminadas por dependencias puramente locales.



$$\ln q(\theta_3) = \ln p(y_3|\theta_3) + \textstyle\sum_{\theta_2} q(\theta_2) \ln p(\theta_3|\theta_2) + \sum_{\theta_4} q(\theta_4) \ln p(\theta_4|\theta_3) - \ln z$$

Figura 4. La distribución de reconocimiento $q(\theta_3)$. Gracias a la aproximación de campo promedio, las únicas variables relevantes para aproximar la distribución *a posteriori* son las variables locales, lo cual simplifica en gran medida el algoritmo inferencial.La pregunta obvia en este contexto es cómo podemos obtener $q(\theta_t)$ sin conocer $q(\theta_{t-1})$ o $q(\theta_{t+1})$. La respuesta es utilizar un procedimiento iterativo, donde las distribuciones q son inicializadas con una parametrización arbitraria y cada una de estas es actualizada iterativamente. Es posible probar que el punto fijo o punto de equilibrio de este procedimiento maximiza la energía libre negativa (Koller y Friedman 2009).

El algoritmo variacional de Bayes puede ser resumido en tres pasos:

1. Definir una probabilidad de reconocimiento q utilizando algún tipo de factorización aproximativa $q=\prod_i q_i$.
2. Inicializar los factores q_i.
3. Maximizar iterativamente cada uno de los factores q_i.

En la práctica, como se optimice q, depende de las condiciones particulares del problema.

La interpretación fuerte del principio de la energía libre

En la sección anterior, mostré que el algoritmo variacional de Bayes es un procedimiento para aproximar una distribución *a posteriori* desconocida. Este se basa en proponer una función de reconocimiento q que usualmente puede ser factorizada en distribuciones locales (un método llamado aproximación del campo promedio). Usando ciertas identidades variacionales, es posible obtener expresiones analíticas para la función de reconocimiento q. Estas identidades son el resultado de maximizar la energía libre con respecto de la función de reconocimiento q.

¿Cuál es la relación del principio de la energía libre propuesto por Friston y el algoritmo variacional bayesiano? Una interpretación posible del principio de la energía libre —que denominaré interpretación fuerte— afirma que el tipo de inferencias realizadas por un organismo son realizadas a través del algoritmo variacional de Bayes. Bajo esta interpretación, reformular la noción de 'inferencia' en términos de la optimización de la energía libre implica abandonar el territorio de la teoría normativa de la cognición y acercarse a una teoría descriptiva de

esta. La razón es que el algoritmo variacional bayesiano se compromete con el tipo de representación que un agente debe usar —la distribución de reconocimiento *q* — y el tipo de aproximaciones necesarias para resolver este problema.

Respecto del primer compromiso, el principio de la energía libre —en esta interpretación— requiere una teoría representacional muy particular: el resultado de una inferencia es la distribución de reconocimiento *q*. El principio de la energía libre, en su lectura fuerte, tiene también una relación con el tipo de aproximaciones necesarias para poder realizar inferencias bayesiana. Como mencioné —desde un punto de vista puramente algorítmico—, minimizar la energía libre es posible gracias a la aproximación del campo promedio. Esta aproximación asume que subgrupos de variables latentes son independientes, esto es, que su distribución conjunta puede ser factorizada.

En resumen, el principio de la energía libre puede ser visto, bien como una reformulación de la teoría bayesiana del cerebro, bien como una teoría algorítmica de cómo la inferencia bayesiana es implementada en el cerebro. En el primer caso, la interpretación se basa en la equivalencia formal de ciertos conceptos clave en termodinámica y el concepto bayesiano de 'inferencia'. Esta interpretación es poco atractiva, porque la interpretación termodinámica, si bien ilumina ciertos aspectos del concepto de 'inferencia', falla en expresar los problemas computacionales que son fundamentales para dilucidar los procesos algorítmicos que ocurren en el cerebro. La segunda lectura tiene como resultado proponer una teoría representacional y algorítmica de la inferencia: esta teoría propone no solo que el cerebro realiza inferencias bayesianas, sino que para esto optimiza la energía libre de una función de reconocimiento *q*. Según esta interpretación, el principio de la energía libre tiene tres predicciones centrales: las distribuciones de reconocimiento *q* deben ser representadas por el cerebro, estas deben obedecer a algún tipo de aproximación del campo promedio y son aproximadas usando un algoritmo iterativo que *explícitamente* minimiza la energía libre negativa usando identidades variacionales.

Friston se inclina por una interpretación fuerte del principio de energía libre:

> Over the past decade, a free energy principle has been proposed that explains several aspects of action, perception and cognition in the neurosciences. This principle appeals to variational Bayesian methods in statistics to understand how the brain infers the causes of its sensory inputs based on the original proposal of Helmholtz and subsequent advances in psychology and machine learning. In brief, variational Bayesian methods allow one to perform approximate Bayesian inference, given some data and a (generative) model of how those data were generated. The key feature of these methods is the minimization of a variational free energy that bounds the (negative log) evidence for the model in question. This minimization eschews the difficult problem of evaluating the evidence directly, where evidence corresponds to the probability of some data, given the model. […] Crucially, under some simplifying assumptions, these variational schemes can be implemented in a biologically plausible way, making them an important metaphor for neuronal processing in the brain (Friston 2012, 2001).

En la misma dirección, Friston et al. (2010, 228) sugieren:

> The idea that the brain uses hierarchical inference has been established for years and provides a nice explanation for the hierarchical organization of cortical systems. Critically, hierarchical inference can be formulated as a minimization of free-energy; where free-energy bounds the surprise inherent in sensory data, under a model of how those data were caused. This leads to the free-energy principle, which says that everything in the brain should change to minimize free-energy. […] Free-energy can be minimized by changing perceptual representations so that they approximate a posterior or conditional density on the causes of sensations. In short, the free-energy principle entails the Bayesian brain hypothesis.

Este tipo de afirmaciones y los argumentos usualmente propuestos por Friston sugieren que el principio de la energía libre debe ser vista como una teoría algorítmica y representacional de cómo inferencias bayesianas aproximativas son realizadas por el cerebro.

Desafortunadamente en su interpretación fuerte, no hay mayor evidencia empírica en favor del principio de la energía libre. Si bien recientemente un número de estudios (por ejemplo Iglesias et al. 2013, Diaconescu et al. 2014) han demostrado que las predicciones hechas por los modelos bayesianos que utilizan el algoritmo variacional se correlacionan con cambios en la actividad del cerebro, ninguno de estos estudios ha mostrado evidencia de que otro método aproximativo puede explicar de forma más apropiada los datos experimentales reportados por estos autores. De hecho, estudios anteriores encontraron resultados similares utilizando el mismo modelo al usar métodos de fuerza bruta para producir inferencia bayesiana exacta (Behrens et al. 2007; Behrens et al. 2008).

En general, mostrar que ciertos tipos de computación en el cerebro pueden ser descritos con el concepto bayesiano de 'inferencia' no es evidencia adecuada del principio de la energía libre en su interpretación fuerte. La razón es que el tipo de argumentos que debe ofrecer esta teoría ha de demostrar que el cerebro codifica probabilidades de reconocimiento, que estas utilizan la aproximación del campo promedio y que son optimizadas usando el método variacional. En otras palabras, el tipo de evidencia central en favor de esta teoría debe demostrar que el tipo de errores que surgen de la aproximación del campo promedio corresponden al tipo de errores que humanos comenten rutinariamente.

Un estudio prominente (aunque probablemente el único) en favor del principio de la energía libre en su lectura fuerte es Daw et al. (2008). Este grupo consideró explícitamente un modelo de inferencia bayesiana aproximativa del fenómeno conocido como *condicionamiento*, y demostraron que un modelo óptimo de inferencia no puede predecir una serie de observaciones experimentales, pero una aproximación variacional puede predecir el tipo de errores inferenciales que comenten los sujetos en este paradigma. Este tipo de estudios ponen de manifiesto que, si el objetivo es ofrecer una teoría representacional y algorítmica basada en la hipótesis bayesiana del cerebro, los argumentos experimentales deben explicar cómo las inferencias humanas son subóptimas.

Por otra parte, hay evidencia de que ciertos procesos cognitivos pueden ser explicados con una teoría algorítmica opuesta al prin-

cipio de la energía libre en su interpretación fuerte. En particular, hay evidencia de que, en vez de utilizar distribuciones de reconocimiento para representar distribuciones *a posteriori*, un gran número de muestras (samples) pueden ser usadas para representar la misma distribución (Gershman y Daw 2012a). Este tipo de métodos —llamados métodos de Monte Carlo— se basan en dos observaciones: primero, la ley de grandes números garantiza que con suficientes muestras es posible representar las propiedades fundamentales de una distribución con exactitud arbitraria. La segunda observación es que, sorprendentemente, es posible obtener muestras de una distribución aun cuando no es posible representarla explícitamente (Robert y Casella 2004). De esta manera, y de forma contraria al principio de la energía libre en su interpretación fuerte, esta teoría requiere un tipo de representaciones radicalmente distintas y, en consecuencia, métodos de inferencia distintos.

Un ejemplo del tipo de evidencia experimental en favor de esta posición fue presentada por Gersman, Vul y Tenenbaum (2012b). Este grupo modeló el fenómeno de rivalidad binocular usando el método de Monte Carlo. La rivalidad binocular se genera cuando dos estímulos diferentes son presentados al ojo izquierdo y derecho, respectivamente (Tong, Meng y Blake 2006). La experiencia perceptual de estos estímulos está caracterizada por la experiencia consciente de solo uno de estos, con transiciones estocásticas entre experiencias de cada uno de los estímulos. Gershman, Vul y Tenenbaum (2009) demostraron que las transiciones pueden ser explicadas por un algoritmo llamado Metropolis-Hasting, que se caracteriza por tomar muestras secuenciales de una distribución. Estas muestras están correlacionadas unas con otras, pero en conjunto representan (asintóticamente) la distribución objetivo. Este estudio demostró que la transición entre una experiencia a otra puede ser explicada como el resultado de la transición de la secuencia entre modas (*modes*) de una distribución *a posteriori* multimodal. Más importante aún, Gershman, Vul y Tenenbaum (2009) mostraron que la distribución de los tiempos de transición entre las experiencias perceptuales puede ser también explicada usando el algoritmo de Metropolis-Hastings. La conclusión es que, si bien estudios como el de Gershman, Vul y Tenenbaum (2009) se basan en un principio normativo, esto es, en la hipótesis

bayesiana del cerebro, estos ofrecen evidencia empírica en favor de una teoría opuesta a la interpretación fuerte del principio de la energía libre.

Por último, otro tipo de argumentos proviene de la literatura sobre algoritmos heurísticos, definidos por Newell, Calman y Simon (1958, en Stuart y Norvig 1995, 94) así: "Un proceso que puede resolver un problema, pero no ofrece garantías de hacerlo, es llamado una solución heurística a tal problema". En general, las soluciones heurísticas se caracterizan por resolver un problema explotando las características particulares de este, sin demostrar garantías formales sobre el tipo de soluciones generadas (por ejemplo tiempo máximo de ejecución, completitud, etcétera). En las últimas tres décadas, un gran número de estudios (Gigerenzer y Goldstein 1996; Marsh, Todd y Gigerenzer 2004; Gigerenzer y Gaissmaier 2011; Gingerenzer y Goldstein 2011) ha demostrado que los seres humanos y otros organismos utilizan estrategias heurísticas para resolver problemas inferenciales, sin recurrir, por ejemplo, al algoritmo variacional de Bayes o, de forma más general, a ningún método probabilístico. Un ejemplo típico de este comportamiento es conocido como la heurística de reconocimiento (Marsh, Todd y Gigerenzer 2004), fenómeno donde los sujetos prefieren una opción conocida a una opción desconocida, aun cuando no hay información respecto de la utilidad de ambas.

Optar por estrategias heurísticas es una consecuencia natural de las restricciones en recursos cognitivos que he discutido con anterioridad. Así, por ejemplo, los algoritmos heurísticos han sido caracterizados de la siguiente manera: "A heuristic is a strategy that ignores part of the information, with the goal of making decisions more quickly, frugally, and/or accurately than more complex methods" (Gigerenzer y Gaissmaier 2011, 454). En otras palabras, una estrategia heurística busca una solución adecuada a un problema de decisión dada una cantidad limitada de recursos. Perspectivas como la de Gigerenzer no se oponen directamente a la teoría de la energía libre en su versión débil; al contrario, se basa en las mismas consideraciones y tiene un origen similar en la teoría de racionalidad limitada propuesta por Herbert Simon. Sin embargo, estas se oponen directamente a la versión fuerte de la teoría de la energía libre, y constituyen evidencia en su contra.

Conclusiones

En este artículo, he tratado de mostrar que la hipótesis bayesiana del cerebro debe ser entendida como una teoría normativa del tipo de computaciones que cualquier agente inteligente debe realizar. El valor teórico de esta teoría es que ofrece un gran poder expresivo para modelar las inferencias que un agente inteligente debe realizar. Esta característica conduce a un paradigma que no impone ninguna restricción algorítmica o representacional al tipo de procesos que son necesarios para realizar inferencias. De manera un poco paradójica, es este el paso que parece ser el más interesante en la investigación científica: solo teorías que proponen mecanismos algorítmicos para solucionar problemas inferenciales pueden ser evaluadas de manera empírica. El hecho de que la hipótesis bayesiana del cerebro sea defendida sin considerar esta limitación genera recelos sobre la agenda detrás de esta hipótesis. En particular, no es claro cuál es el valor epistemológico de modelar ciertos procesos cognitivos utilizando el lenguaje de la estadística bayesiana:

> Taken as a psychological theory, the Bayesian framework does not have much to say. Its most unambiguous claim is that much of human behavior can be explained by appeal to what is rational or optimal. […] More importantly, rational explanations for behavior offer no guidance as to how that behavior is accomplished. […] The Bayesian framework is more radical in that, unlike previous brain metaphors grounded in technology and machines, the Bayesian metaphor is tied to a mathematical ideal and thus eschews mechanism altogether (Jones y Love 2011, 173).

El principio de la energía libre es un gran candidato para ser víctima de este tipo de criticismos. Por un lado, este parece ser, en ocasiones, la reformulación de la hipótesis bayesiana del cerebro en términos termodinámicos, esto es, la observación de que todo tipo de inferencia puede ser entendida como un proceso que reduce la energía libre de un sistema. Aunque correcta, esta interpretación de la hipótesis bayesiana del cerebro tiende a olvidar la necesidad de formular teorías algorítmicas y representacionales de este tipo de inferencias.

Una segunda interpretación del principio de energía libre consiste en ver la formulación de la energía libre como una teoría algorítmica basada en métodos variacionales de optimización. Desde esta perspectiva, el principio de la energía libre afirma que los agentes realizan inferencia bayesiana mediante un algoritmo particular que conduce a cierto tipo de suboptimalidades. El problema con esta interpretación es que no hay mayor evidencia en favor de esta; al contrario, hay evidencia en su contra.

La confusión entre estas dos interpretaciones tiene origen en el tipo de argumentos que usualmente son usados en favor del principio de la energía libre. Si bien la teoría es presentada en algunas ocasiones como una teoría general sobre la inferencia, el tipo de argumentos presentados buscan demostrar que un tipo particular de algoritmos puede ser implementado por el cerebro. Sin embargo, estos argumentos fallan en demostrar que las aproximaciones que estos algoritmos suponen son las mismas que las usadas por el cerebro humano.

En este contexto, la propuesta de Friston se inclina en la dirección de la interpretación fuerte de la energía libre, en la medida en que las predicciones acerca de los métodos algorítmicos usados por el cerebro han sido sostenidos en varias de sus publicaciones (por ejemplo, Bastos et al. 2012). Vista como una teoría general de los procesos algorítmicos usados por el cerebro para realizar inferencias inductivas, esta propuesta es poco viable. En efecto, como señalé en la sección anterior, hay una plétora de evidencia que demuestra que el cerebro utiliza estrategias distintas del algoritmo variacional de Bayes para realizar inferencias inductivas. En general, es muy probable que diferentes problemas requieran diferentes estrategias para generar soluciones adecuadas que balanceen la calidad de las respuestas con la cantidad de recursos invertidos en estas. Las presiones evolutivas deben favorecer soluciones *adecuadas* donde tiempo y recursos son limitaciones reales, independientemente del algoritmo particular usado por el cerebro para implementarlas.

En general, si el principio de energía libre ha de convertirse en un cambio teórico fundamental en la neurociencia, una revisión conceptual de este principio desde la perspectiva de la filosofía es necesaria. Dilucidar cuáles aspectos de este principio son de carácter puramente

normativo y cuáles son de carácter empírico puede ayudar a formular experimentos empíricos que contribuyan a nuestro entendimiento del cerebro humano.

Referencias bibliográficas

Bastos, Andre et al. 2012. "Canonical microcircuits for predictive coding". *Neuron* 4: 695-711.

Behrens, Timothy et al. 2007. "Learning the value of information in an uncertain world". *Nature Neuroscience* 9: 1214-1221.

Behrens, Timothy et al. 2008. "Associative learning of social value". *Nature Neuroscience* 7219: 245-249.

Blundell, Stephen y Katherine M. Blundell. 2010. *Concepts in thermal physics*. Oxford, Inglaterra: Oxford University Press.

Carnap, Rudolf. 1950. *The fundations of probability*. Chicago: University of Chicago Press.

Clark, Andy. 2013 "Whatever next? Predictive brains, situated agents, and the future of cognitive science". *Behavioral and Brain Sciences* 3: 181-204.

Daw, Nathaniel D., Aaron C. Courville y Peter Dayan. 2008. "Semi-rational models of conditioning: the case of trial order". En *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*, Nick Chater y Mike Oaksford, 431-452. Oxford, Inglaterra: Oxford University Press.

Dayan, Peter et al. 1995. "The helmholtz machine". *Neural Computation* 5: 889-904.

Diaconescu, Andreea O. et al. 2014. Inferring on the intentions of others by hierarchical Bayesian learning. *The PLOS Computational Biology Staff* 10: e1003952.

Feynman, Richard Phillips. 1998. *Feynman lectures on computation*. Boston, MA: Addison-Wesley Longman Publishing Co., Inc.

Fine, Cordelia et al. 2007. "Hopping, skipping or jumping to conclusions? Clarifying the role of the JTC bias in delusions". *Cognitive Neuropsychiatry* 1: 46-77.

FitzGerald, Thomas H. B. et al. (2015). "Active inference, evidence accumulation, and the urn task". *Neural Computation* 2: 306328.

Freedman, David A. y Roger A. Purves. 1969. "Bayes' method for bookies". *The Annals of Mathematical Statistics* 4: 1177-1186.

Friston, Karl. 2009. "The free-energy principle: a rough guide to the brain?". *Trends in Cognitive Sciences* 7: 293-301.

Friston, Karl. 2010. "The free-energy principle: a unified brain theory?". *Nature Reviews Neuroscience* 2: 127-138.

Friston, Karl. 2012. "A free energy principle for biological systems". *Entropy* 11: 2100-212.

Friston, Karl, James Kilner y Lee Harrison. 2006. "A free energy principle for the brain". *Journal of Physiology* 1: 70-87.

Friston, Karl J. y Klaas E. Stephan. 2007. "Free-energy and the brain". *Synthese* 3: 417-458.

Friston, Karl J. et al. 2010. "Action and behavior: a free-energy formulation". *Biological Cybernetics* 3: 227-260.

Friston, Karl, Spyridon Samothrakis y Read Montague. 2012. "Active inference and agency: optimal control without cost functions". *Biological Cybernetics* 8-9: 523-541.

Friston, Karl et al. 2013. "The anatomy of choice: active inference and agency". *Frontiers in Human Neuroscience* 7: 598.

Gershman, Samuel, Ed Vul y Joshua B. Tenenbaum. 2009. "Perceptual multistability as Markov chain Monte Carlo inference". *Advances in Neural Information Processing Systems* 22: 611-619.

Gershman, Samuel J., Edward Vul y Joshua B. Tenenbaum. 2012a. Multistability and perceptual inference". *Neural Computation* 1: 1-24.

Gershman, Samuel J. y Nathaniel D. Daw. 2012b. Perception, action and utility: the tangled skein. En *Principles of brain dynamics: global state interactions*, editado por M. I. Rabinovich, Karl J. Friston y Pablo Varona, 293-312. Cambridge, MA: The MIT Press.

Gibson, James J. 1978. "The ecological approach to the visual perception of pictures". *Leonardo* 3: 227-235.

Gibson, James J. 2015. *The ecological approach to visual perception*. Nueva York/Londres: Psychology Press.

Gigerenzer, Gerd y Wolfgang Gaissmaier. 2011. "Heuristic decision making". *Annual Review Of Psychology* 62: 451-482.

Gigerenzer, Gerd y Daniel G. Goldstein. 1996. "Reasoning the fast and frugal way: models of bounded rationality". *Psychological Review* 4: 650.

Gigerenzer, Gerd y Daniel G. Goldstein. 2011. "The recognition heuristic: a decade of research". *Judgment and Decision Making* 1: 100-121.

Glaymour, Clark. 1981. "Why I'm not a Bayesian". En *Theory and evidence*, 63-93. Chicago: University of Chicago Press.

Griffiths, Thomas L., Charles Kemp y Joshua B. Tenenbaum. 2008. "Bayesian models of cognition". En *Cambridge Handbook of computational cognitive modeling*. Cambridge: Cambridge University Press.

Iglesias, Sandra et al. 2013. "Hierarchical prediction errors in midbrain and basal forebrain during sensory learning". *Neuron* 2: 519-530.

Jaynes, Edwin T. 1957. "Information theory and statistical mechanics". *Physical Review* 4: 620.

Jones, Matt y Bradley C. Love. 2011. "Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition". *Behavioral and Brain Sciences* 4: 169-188.

Jordan, Michael I. et al. 1999. "An introduction to variational methods for graphical models". *Machine Learning* 2: 183-233.

Kamberova, Gerda. 1992. "Markov random elds: A Bayesian approach to computer vision problems". *Technical Reports (CIS)*. Paper 491.

Kersten, Daniel, Pascal Mamassian y Alan Yuille. 2004. "Object perception as Bayesian inference". *Annual Review of Psychology* 55: 271-304.

Knill, David C. y Alexandre Pouget. 2004. "The Bayesian brain: the role of uncertainty in neural coding and computation". *TRENDS in Neurosciences* 12: 712-719.

Koller, Daphne y Nir Friedman. 2009. *Probabilistic graphical models: principles and techniques*. Cambridge, MA: The MIT Press.

Marr, David y Tomaso Poggio. 1976. *From understanding computation to understanding neural circutry*. Cambridge, MA: The MIT Press.

Marsh, Barnaby, Peter M. Todd y Gerd Gigerenzer 2004. Cognitive heuristics: reasoning the fast and frugal way. En *The nature of reasoning*, editado por Leighton, Jacqueline P. y Robert J. Sternberg, 273-287. Nueva York: Cambridge University Press.

Montague, P. Read et al. 2012. "Computational psychiatry". *Trends in Cognitive Sciences* 1: 72-80.

Newell, Allen, John Calman Shaw y Herbert Alexander Simon, 1958. "Chess-playing programs and the problem of complexity". *IBM Journal of Research and Development* 4: 320-335.

Ortega, Pedro A. y Daniel A. Braun. 2013. "Thermodynamics as a theory of decision-making with information-processing costs". *Proceedings* 2153: 20120683.

Pearl, Judea. 2000. *Causality: models, reasoning and inference*. Nueva York: Cambridge University Press.

Petzschner, Frederike H. y Stefan Glasauer. 2011. "Iterative Bayesian estimation as an explanation for range and regression effects: a study on human path integration". *The Journal of Neuroscience* 47: 17220-17229.

Phillips, Lawrence D. y Ward Edwards. 1966. "Conservatism in a simple probability inference task". *Journal of Experimental Psychology* 3: 346.

Pollock, John L. 1987. "Epistemic norms". *Synthese* 1: 61-95.

Putnam, Hilary. 1964. "Probability and confirmation". US Information Agency, Voice of America Forum.

Ramsay, Frank P. 1964. Truth and probability, 1926. En *Studies in subjective probability*, editado por H. Kyburg y H. Smokler, 23-52. Nueva York: John Wiley.

Robert, Christian. 2007. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Nueva York: Springer Science & Business Media.

Robert, Christian y George Casella. 2004. *Monte carlo statistical methods*. Berlín: Springer Verlag.

Russell, Stuart y Peter Norvig. (1995). *Artificial intelligence: a modern approach*. Englewood Cliffs, NJ: Prentice Hall.

Spratling, Michael W. 2013. "Distinguishing theory from implementation in predictive coding accounts of brain function". *Behavioral and Brain Sciences* 3: 231-232.

Tong, Frank, Ming Meng y Randolph Blake. 2006. "Neural bases of binocular rivalry". *Trends in Cognitive Sciences* 11: 502-51.