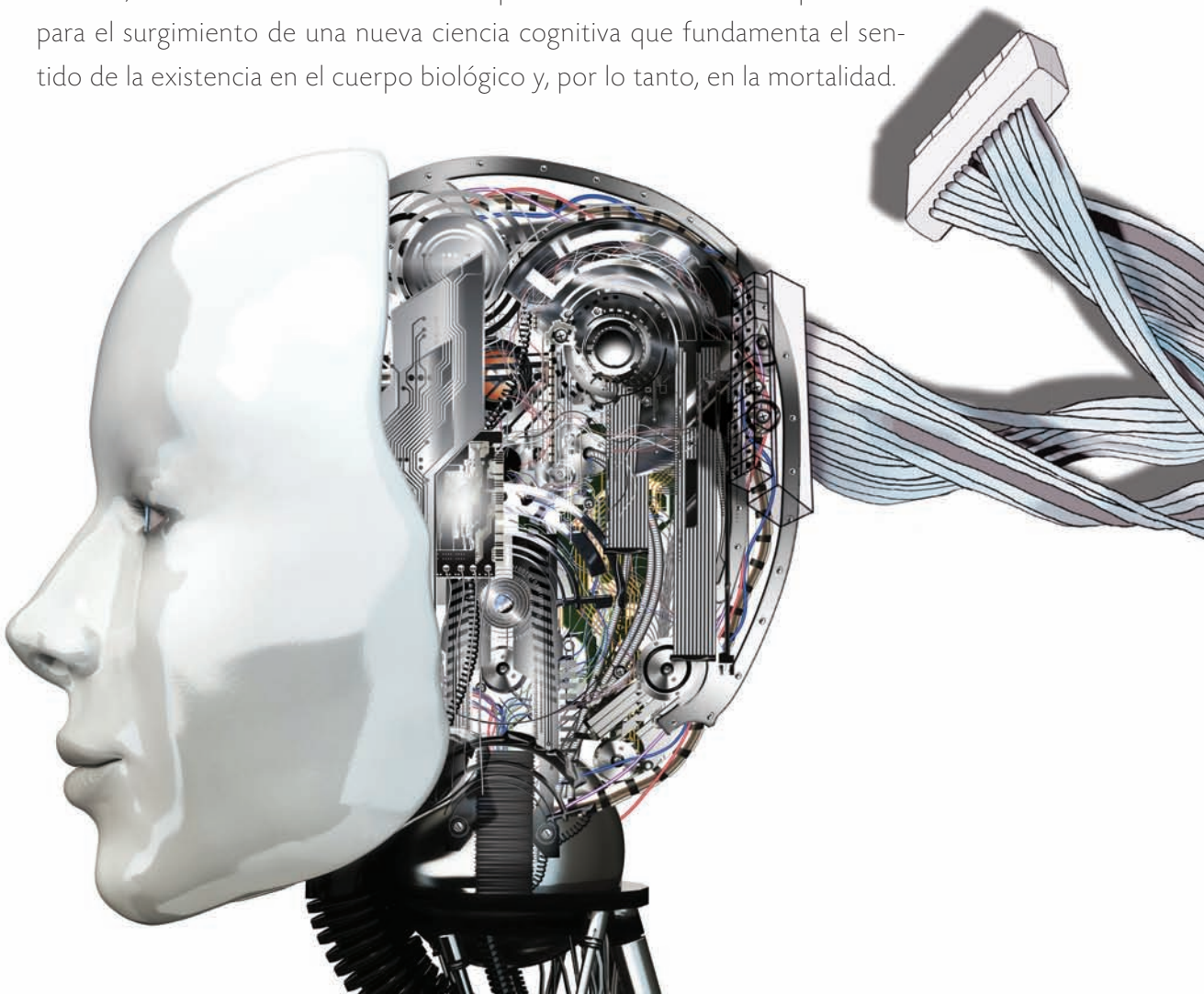
Tom Froese

# De la **cibernética** a la nueva **ciencia cognitiva**

El cibernético mexicano Rosenblueth y sus colegas Wiener y Bigelow argumentaban que el comportamiento dirigido a metas puede ser explicado por la retroalimentación negativa. Esta propuesta revolucionaria implicaba que nuestra experiencia al actuar intencionadamente podía hacerse compatible con una visión del mundo estrictamente científica, en la cual la naturaleza física no sigue ningún propósito. Años después, Wiener fundaría la cibernética bajo el principio de *autogobierno*, por ejemplo, con el uso de "bucles" de retroalimentación negativa para el control de máquinas. Sin embargo, los seres vivos no sólo son autogobernantes, sino que también, a través del metabolismo, son individuos físicamente autoproductivos. Esto es de importancia para el surgimiento de una nueva ciencia cognitiva que fundamenta el sentido de la existencia en el cuerpo biológico y, por lo tanto, en la mortalidad.

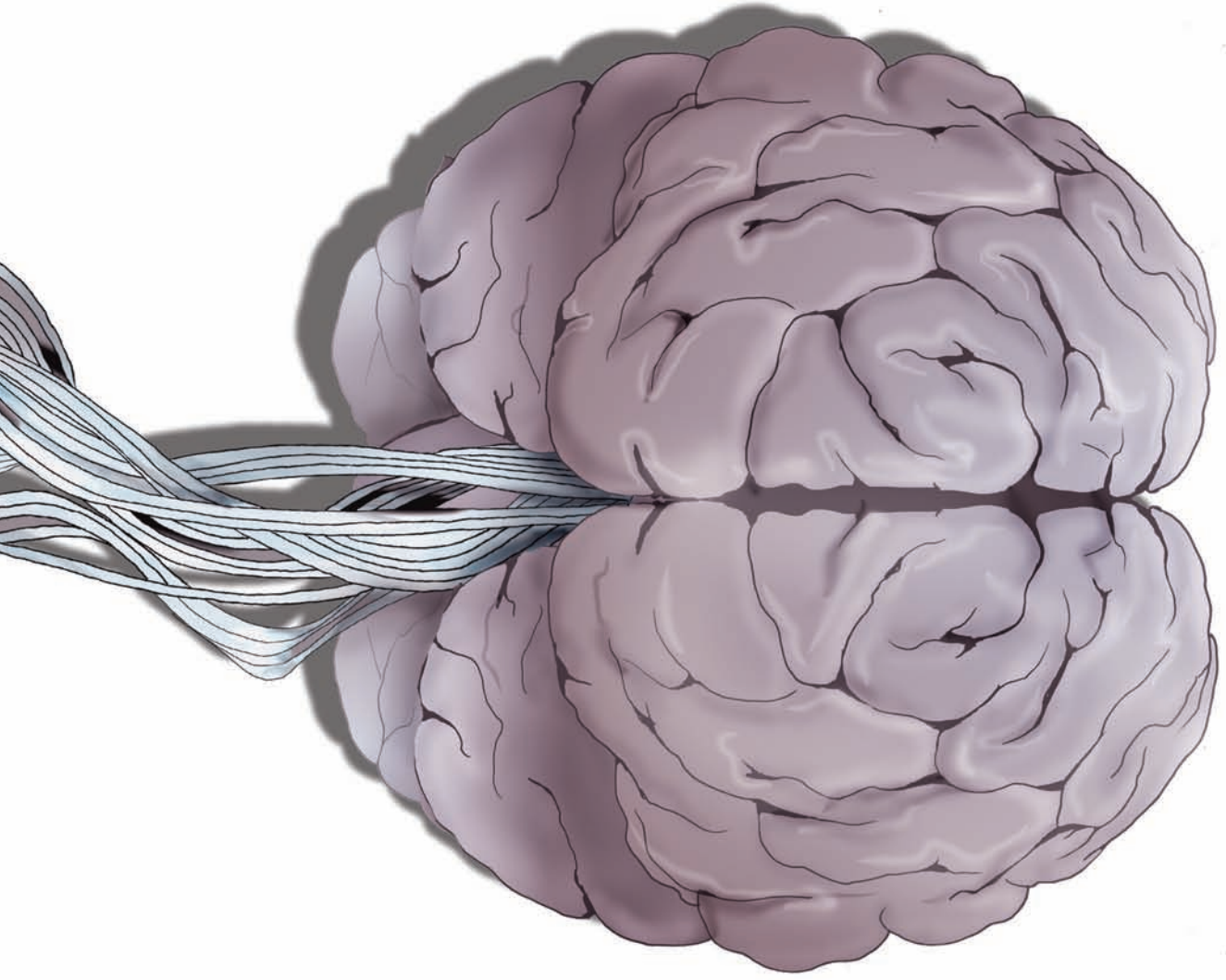


## Introducción

En 1943 Arturo Rosenblueth, Norbert Wiener y Julian Bigelow publicaron uno de los artículos científicos más famosos de la época (Rosenblueth *et al.*, 1943). Su contribución fue parte fundacional de una agrupación de investigadores increíblemente transdisciplinaria que desde 1948 se denominaría *cibernética*, en seguimiento a la propuesta hecha por Wiener. El artículo presentaba un análisis teórico del comportamiento orientado a una meta y sobre cómo éste se manifiesta en todos los seres vivos y también en algunos dispositivos diseñados para actuar automáticamente.

La gran innovación fue acercar la ciencia, con sus métodos y teorías, al estudio del fenómeno psicológico y subjetivo de seguir un fin u objetivo. En general, las metas son asociadas a condiciones de satisfacción y de la posibilidad de fracaso, algo que no se aplica a fenómenos del dominio de la física pura, como por ejemplo, la trayectoria de un planeta. Con base en la aplicación de la teoría de la retroalimentación negativa, Rosenblueth y sus colegas fueron más allá del reduccionismo clásico para fundar el estudio de la mente como tal. Éste fue un paso importante para superar el gran reto de investigar la relación mente-cuerpo.

Sin embargo, en retrospectiva, la aplicación de la teoría de la retroalimentación negativa era demasiado general. No existe duda de que podemos experimentar en nosotros mismos el fenómeno de tener metas propias y de preocuparnos



por su éxito; ¿pero qué tal un termostato diseñado para mantener la temperatura por medio de la retroalimentación negativa? ¿Es correcto decir que el termostato también tiene fines propios, o bien fines *intrínsecos*, como dijeron Rosenblueth y sus colegas? Algunos investigadores en el campo de la inteligencia artificial contemporánea siguen estando de acuerdo en que no hay ninguna diferencia esencial entre un termostato y un ser vivo (incluidos los seres humanos). Sin embargo, un movimiento importante en la nueva ciencia cognitiva –a veces llamado *enactivismo*, el cual sigue la propuesta del biólogo chileno Francisco Varela y sus colegas (1992)– plantea que existen diferencias fundamentales que separan la organización de los seres vivos y la de los sistemas que son definidos básicamente por la retroalimentación negativa.

Desde la década de 1970, los biólogos chilenos Humberto Maturana y Francisco Varela han destacado que los seres vivos deben estar definidos como seres *autopoiéticos* (su término *autopoiesis* equivale a *autoproducción*) (Maturana y Varela, 1973). Además, según el nuevo enactivismo –por la necesidad de crear y mantener constantemente su organización física y evitar la desintegración y la muerte–, un organismo también genera una perspectiva de preocupación sobre su ambiente, con lo cual se ubica en un mundo significativo. En este sentido, la nueva ciencia cognitiva argumenta que solamente los seres vivos tienen fines propios de una manera intrínseca y subjetiva. La implicación es que somos más que nuestras máquinas porque somos individuos mortales (Di Paolo, 2015).

## De la retroalimentación negativa a la retroalimentación doble

La retroalimentación negativa puede usarse para corregir errores, lo cual es útil si queremos crear sistemas autorreguladores. Consideremos de nuevo el caso del termostato. Por ejemplo, imaginemos que la llegada del verano provoca en la casa un incremento en la temperatura, que es medida por el termómetro. El termostato compara esta medición con el valor deseado para el cual ha sido programado. En este caso, la medida va a ser muy alta. Un regulador contrarresta esta divergencia, por ejemplo, al incrementar la salida del sistema de enfriamiento. La influencia del calor externo es, por lo tanto, reducida de forma automática.

Desde la perspectiva de los habitantes de la casa, este sistema se está comportando intencionadamente y siguiendo una meta: mantener la temperatura alrededor de cierto punto establecido. Pero ¿comparte el sistema este punto de vista sobre sus operaciones? Una manera de revelar el hecho de que en realidad al sistema no le importa esto es exponiéndolo a una fuerte perturbación, que tiene el efecto de frustrar su "meta" y ver cómo reacciona.

Imaginemos que abrimos la caja con el sistema de control del clima en la casa y cambiamos un par de cables, de tal manera que el sistema envía exactamente la señal opuesta al sistema de enfriamiento. Al siguiente día, cuando el calor en la casa empieza a volverse demasiado alto, el termostato otra vez detecta una divergencia del valor deseado y activa el regulador. Pero ahora esto va a tener el efecto de reducir, en lugar de incrementar, la salida del sistema de enfriamiento. Así, aumentará más la temperatura. Por lo tanto, el termostato va a enviar una respuesta aún mayor al regulador, y esto va a reducir todavía más la salida del sistema de

enfriamiento, y así de manera reiterada. Podemos imaginar que el "bucle" va a continuar hasta que en cierto momento, el sistema se sobrecaliente y se quiebre.

La situación es muy diferente en el caso de los organismos vivos. Nuestra temperatura es una variable esencial que el cuerpo debe regular para mantener ciertos estados fisiológicos. Por ejemplo, las personas que tuvieron que enfrentar las consecuencias del sistema de clima reconfigurado deben tener calor. Eventualmente, esto debe de haber desencadenado un caso biológico de retroalimentación negativa. Una vez que nuestro cuerpo se calienta demasiado, empieza a sudar, y la evaporación del sudor tiene un efecto de enfriamiento en nuestra piel. ¿Pero qué pasa cuando esta retroalimentación negativa no sirve? Si el cuerpo se sobrecalienta, va a llegar un punto en el que colapse e incluso muera. Sin embargo, antes de este momento final vamos a empezar a sentir malestar y vamos a tratar de reducir el impacto del calor de otra forma. Podemos abrir ventanas, apagar el sistema de control de la temperatura que funciona mal o simplemente salir de la casa. Para explicar esta notable capacidad indefinida de adaptación, necesitamos más que la retroalimentación negativa.

El psiquiatra y cibernético inglés W. Ross Ashby (Figura 1) dedicó una gran parte de su vida profesional a explicar cómo es posible que los animales puedan adaptar su conducta a circunstancias imprevistas. Proponía que cuando el comportamiento no es efectivo para contrarrestar una perturbación, la organización interna podría ser ajustada automáticamente hasta encontrar un nuevo comportamiento efectivo. En lugar de mera estabilidad, Ashby estaba interesado en lo que llamó *ultraestabilidad* (Ashby, 1960).

Así, logró diseñar y construir un sistema mecánico –hecho con materiales excedentes de la Segunda Guerra Mundial– que podía adaptar espontáneamente su organización interna de modo que sus variables esenciales volvieran a estabilizarse cuando excedían sus límites. Este aparato innovador es conocido como *homeostato*. No necesitamos entrar en los detalles de su funcionamiento aquí; es suficiente que tengamos una comprensión general del concepto de *ultraestabilidad* de Ashby, como se ilustra en la Figura 2.

En el sistema, la parte responsable de regular la interacción sensoriomotora es representada por (*R*), mientras que el ambiente externo es representado por (*E*). Las dos partes están acopladas mediante movimientos (M) desempeñados por el sistema, y las sensaciones (*S*) recibidas por su entorno. Esto puede ser la base de un "bucle" de retroalimentación negativa. La forma precisa de comportamiento está determinada por parámetros (*P*), indicados por una flecha que va de (*P*) a (*R*). El sistema también incluye una variable esencial, que interpretaremos aquí como la temperatura de nuestro cuerpo. Este tipo de variable esencial es

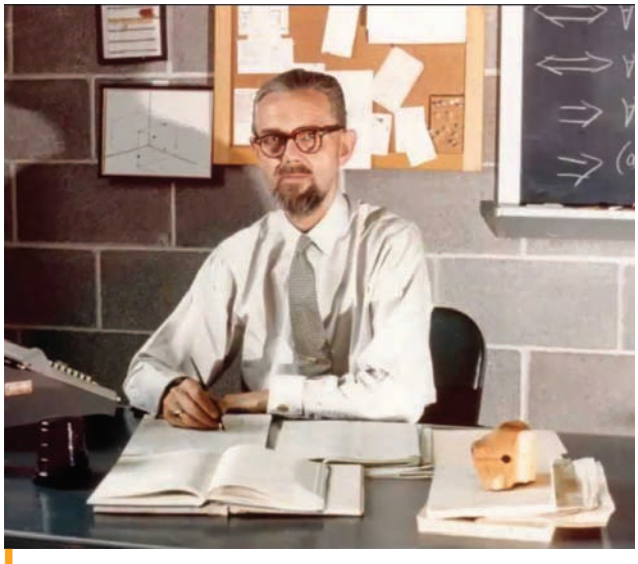


**Figura 1.** W. Ross Ashby (tomada de www.rossashby.info).

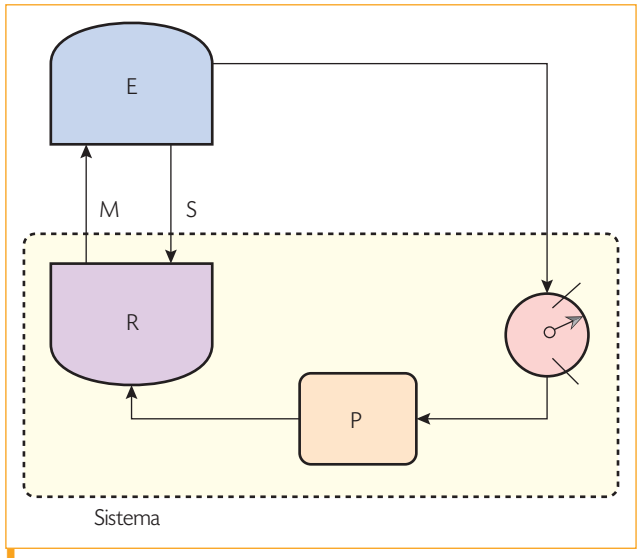



**Figura 2.** Esquema de un sistema caracterizado por la ultraestabilidad.

ilustrada como un reloj comparador con una flecha que debe quedarse entre dos límites marcados en su orilla.

El gran logro en el entendimiento de Ashby fue que la ultraestabilidad podía ser implantada al introducir un segundo "bucle" de retroalimentación, conectando así el estado de la variable esencial con los parámetros (*P*). En particular, cada vez que la variable esencial se escapa de sus límites, el valor de uno de los parámetros (*P*) es modificado aleatoriamente. Este diseño de retroalimentación doble tiene implicaciones profundas en términos de estabilidad.

Imaginemos que el termostato descompuesto ha sido remplazado por un nuevo sistema diseñado para ser ultraestable. ¿Qué pasa si una vez más cambiamos el cableado? Eventualmente la variable esencial –la temperatura– va a salirse de su rango de variabilidad. Y esto a su vez detonará un cambio aleatorio en uno de los parámetros del sistema. En este punto, dos cosas pueden pasar: o bien este cambio en la organización del comportamiento del sistema lleva a una mejora, y en este caso la variable esencial regresa a sus límites; o la situación continúa siendo insostenible, y en dado caso habrá otro cambio en la organización del sistema. En efecto, el sistema continuará cambiando la forma de su comportamiento hasta encontrar una nueva organización estable.

## De la ultraestabilidad a la autopoiesis

Hemos visto cómo los cibernéticos han tratado de explicar el comportamiento intencionado –dirigido a metas intrínsecas– y la autoadaptación interna en términos de diferentes tipos de retroalimentación. Pero aún existe una diferencia crucial entre estos sistemas artificiales y los seres vivos, que está relacionada con el origen de su identidad, la cual define el tipo de sistemas que son. Aunque algunos artefactos pueden autorregularse y autoadaptarse, la noción de su "sí mismo" se mantiene elusiva. En el análisis final, éstos siguen recibiendo su identidad y existencia por parte de los ingenieros que los construyen y diseñan, y su frontera es arbitraria.

Un organismo, por otro lado, está obligado a mantener su identidad física mediante la regulación de sus procesos metabólicos y adaptativos, que se remontan a una forma de autoproducción de su identidad como un sistema. Fue otra gran contribución latinoamericana el resaltar esta importante diferencia entre artefactos cibernéticos y sistemas vivos, y darle una nueva etiqueta: autopoiesis (Maturana y Varela, 1973). Para Maturana (Figura 3) y Varela –quien era estudiante del primero–, la autopoiesis era la fundación sistémica de la autonomía biológica. *Autoproducción* significa que la identidad y el comportamiento de un ser vivo están intrínsecamente determinados por ese organismo en sí mismo. En el caso de los seres vivos, mas no de los mecanismos artificiales, ser y hacer son interdependientes (Froese y Ziemke, 2009).

Maturana y Varela tenían la intención de que su teoría de la autopoiesis fuera solamente sobre la organización de los seres vivos, sin comprometerse a cualquier especulación sobre sus potenciales consecuencias subjetivas. Sin embargo, durante la era de la cibernética había también un movimiento en la filosofía continental que sostenía que la perspectiva del sujeto (su fenomenología) podía estar establecida en la constitución corpórea de lo vivo. Desde la perspectiva de la nueva ciencia cognitiva, el filósofo alemán Hans Jonas proveyó de algunas de las más importantes contribuciones al respecto (Jonas, 2000). Argumentó que experimentamos un mundo con significado y nuestra existencia nos concierne porque el cuerpo biológico es *precario*. Esencialmente, porque somos individuos

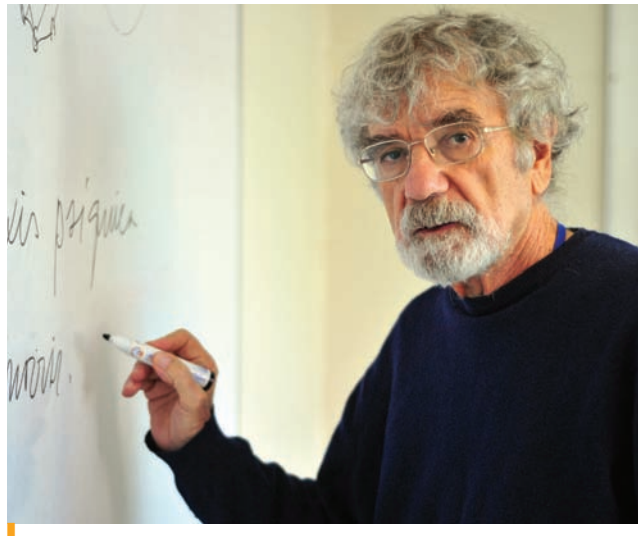


**Figura 3.** Humberto Maturana (tomada de www.mcafestival.cl/portal/images/stories/humberto%20%20maturana.jpg).

finitos constantemente amenazados por la posibilidad de no existencia, existencia que nos importa en primer lugar. El significado que experimentamos depende de la mortalidad.

Una de las contribuciones más importantes de Varela (Figura 4) hacia el final de su vida fue integrar estas comprensiones fenomenológicas con la teoría sistémica de la autopoiesis, con el argumento de que, de esta forma, la ciencia cognitiva será capaz de dar cuenta de la existencia humana en su totalidad, incluyendo los aspectos tanto objetivos como subjetivos (Froese, 2011). Esta propuesta se ha vuelto la fundamentación teórica de la nueva ciencia cognitiva enactiva.

Otro desarrollo reciente ha sido regresar a las ideas de la era de la cibernética en cuanto a la autorregulación y la autoadaptación. La autopoiesis como concepto categorial no nos dice nada sobre los estados posibles mientras un sistema se mantiene existiendo tras su autoproducción. Por ejemplo, podemos estar más o menos sanos, más o menos adaptados a las circunstancias. Para entender cómo puede ocurrir esta variabilidad, necesitamos combinar la autopoiesis con otro principio, uno que permita a un sistema medir grados de posibilidades. El científico argentino Ezequiel Di Paolo ha argumentado que esto se puede lograr si ponemos en claro que los seres vivos son agentes que tienden a regular sus interacciones sensoriomotoras en relación con sus variables esenciales (Di Paolo, 2015). Esta idea de la *agencialidad* –es decir, ser agente– está ilustrada en forma de diagrama en la Figura 5.

El sistema autopoiético es dibujado como una flecha que se cierra a **sí** misma, formando así un círculo. Esto representa la autoproducción (*AP*). Al igual que en la Figura 2 de la ultraestabilidad, el sistema puede afectar al ambiente (línea ondulada) mediante el movimiento (*M*), y a cambio ser afectado por el ambiente mediante sensaciones (*S*). Y una vez más, el sistema puede regular (*R*) la forma particular de su acoplamiento sensoriomotor.

Hay, por lo tanto, muchas semejanzas entre la teoría de ultraestabilidad de Ashby y la teoría de la autopoiesis de Maturana y Varela, pero también existen diferencias importantes (Froese y Stewart, 2010). Ashby estaba interesado en entender la capacidad de un sistema para regular sus interacciones, pero esto sólo se mantuvo implícito en el trabajo de Maturana y Varela, hasta que Di Paolo de nuevo enfatizó su importancia. Por otra parte, la pregunta sobre la identidad de un sistema se mantuvo implícita en los trabajos de cibernéticos como Wiener y Ashby, hasta que fue destacada por Maturana y Varela. Finalmente, mientras Maturana y Varela inicialmente rechazaron cualquier interpretación subjetiva de la organización biológica, al final Varela se convenció de que la autoproducción y su precariedad son la clave para entender que somos individuos que llevan a cabo vidas significativas.

**Figura 4.** Francisco Varela (tomada de http://en.wikipedia.org/wiki/Francisco_Varela).

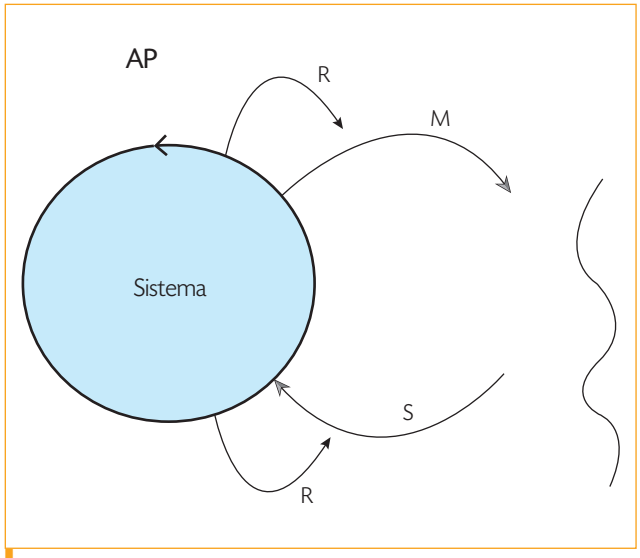
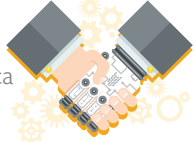



**Figura 5.** Esquema de un sistema caracterizado por agencialidad.



## Conclusiones

Los últimos desarrollos en la ciencia cognitiva se establecen en una larga tradición de teorías revolucionarias que comenzaron en la primera mitad del siglo XX con lo que Wiener llamó cibernética (Froese, 2010). Como hemos visto, ha habido una aportación sustancial de Latinoamérica a esta historia, en particular por Rosenblueth, Maturana y Varela, y más recientemente por Di Paolo. Ellos han contribuido a avanzar en la búsqueda de una teoría de la subjetividad humana que logre hacer justicia a nuestra existencia tanto vivida (subjetiva) como viviente (objetiva). Al final, resulta que la mortalidad no es una consecuencia secundaria de nuestra existencia biológica que pudiera ser potencialmente evitada (como en el caso de los artefactos cibernéticos), sino precisamente aquello que en última instancia nos permite ser individuos interesados por el mundo y a quienes les importan los otros.

Esto abre un espacio para consideraciones éticas. La nueva ciencia cognitiva toma con seriedad el hecho de que participamos directamente en las vidas de otros, por ejemplo, por medio de la corregulación de nuestras interacciones sociales (De Jaegher, 2014). Paso a paso, la nueva ciencia cognitiva está poniéndose de acuerdo con toda la complejidad de la existencia humana.

**Tom Froese** es investigador en el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas de la Universidad Nacional Autónoma de México, donde de 2012 a 2014 fue becario posdoctoral. También cuenta con estancias de posdoctorado en la Universidad de Tokio, Japón, de 2010 a 2012, y en la Universidad de Sussex, Inglaterra, en 2010. Obtuvo la maestría de Ingeniería en Ciencia de la Computación y en Cibernética en la Universidad de Reading, Inglaterra, en 2004, y el doctorado en Ciencia Cognitiva en la Universidad de Sussex, Inglaterra, en 2010. En su tesis de doctorado aplicó el método de la robótica evolutiva para resolver problemas teóricos en las ciencias de la mente. Actualmente su investigación está enfocada en entender mejor los aspectos distintivos de la mente humana relacionados con la cultura y la conciencia.

t.froese@unam.mx

El autor agradece el apoyo parcial de la Dirección General de Asuntos del Personal Académico de la Universidad Nacional Autónoma de México (DGAPA-UNAM) y del Consejo Nacional de Ciencia y Tecnología (Conacyt) para poder realizar esta reseña; y a Héctor Gómez por su ayuda con la traducción al español.

### Lecturas recomendadas

Ashby, W. R. (1960), *Design for a brain: the origin of adaptive behaviour*, Londres, Chapman & Hall.

De Jaegher, H. (2014), "Enacción y autonomía: cómo el mundo social cobra sentido mediante la participación", en A. Casado da Rocha (ed.), *Autonomía con otros: ensayos sobre bioética*, Madrid, Plaza y Valdés, pp. 111-131.

Di Paolo, E. A. (2015), "El enactivismo y la naturalización de la mente", en D. Pérez Chico y M. G. Bedia (eds.), *Nueva ciencia cognitiva: hacia una teoría integral de la mente*, Zaragoza, PUZ.

Froese, T. y T. Ziemke (2009), "Enactive artificial intelligence: investigating the systemic organization of life and mind", *Artificial Intelligence*, 173:366-500.

Froese, T. (2010), "From cybernetics to second-order cybernetics: a comparative analysis of their central ideas", *Constructivist Foundations*, 5:75-85.

Froese, T. y J. Stewart (2010), "Life after Ashby: ultrastability and the autopoietic foundations of biological individuality", *Cybernetics & Human Knowing*, 17:83-106.

Froese, T. (2011), "From second-order cybernetics to enactive cognitive science: Varela's turn from epistemology to phenomenology", *Systems Research and Behavioral Science*, 28:631-645.

Jonas, H. (2000), *El principio Vida: hacia una biología filosófica*, Valladolid, Trotta.

Maturana, H. R. y F. J. Varela (1973), *De máquinas y seres vivos: una teoría sobre la organización biológica*, Santiago de Chile, Editorial Universitaria.

Rosenblueth, A. N., N. Wiener y J. Bigelow (1943), "Behavior, purpose and teleology", *Philosophy of Science*, 10:18-24.

Varela, F. J., E. Thompson y E. Rosch (1992), *De cuerpo presente: las ciencias cognitivas y la experiencia humana*, Barcelona, Gedisa.