Fecha de recepción: mayo 2026
Fecha de aceptación: julio 2026

Herbert Simon entre símbolos y subsímbolos. Diseño, complejidad y legitimidad bajo racionalidad acotada

Luis Everardo Castro Solís [1]

Resumen: En *Las ciencias de lo artificial*, Herbert Simon formula la hipótesis del sistema físico de símbolos como fundamento de la inteligencia, entendida como procesamiento de representaciones internas orientadas a la resolución de problemas. El presente texto reexamina dicha hipótesis a la luz del auge de las arquitecturas subsimbólicas, cuya competencia emerge de configuraciones estadísticas distribuidas sin manipulación simbólica explicita. Desde la noción de racionalidad acotada, se argumenta que, aunque la inteligencia pueda surgir sin símbolos, la legitimidad pública no: la opacidad de los modelos subsimbólicos limita la trazabilidad y la producción de razones auditables. Se propone que la gobernanza de sistemas de alto impacto requiere una capa simbólica exógena capaz de habilitar una justificación proporcional al riesgo. El capítulo identifica la arquitectura neurosimbólica como principio de diseño adecuado para tales fines.

Palabras clave: sistemas simbólicos - conexionismo - inteligencia artificial - racionalidad acotada - gobernanza algorítmica

[Resúmenes en inglés y portugués en la página 153]

[1] Ver CVs en p. 153

Introducción

La discusión sobre las ciencias de lo artificial como sistemas simbólicos, no solo es pertinente sino también altamente relevante para una ética práctica, ya que hoy en día, la síntesis ingenieril de objetos (el mundo de lo artificial) tienen injerencia en un sinnúmero de procesos y sistemas naturales, poniendo de relieve la distinción entre lo descriptivo (las ciencias naturales) versus lo normativo que debe aparecer en la síntesis al ponderar su impacto en la naturaleza.

En el presente texto de carácter reflexivo propositivo, se muestra que la capacidad estructural de los modelos subsimbólicos, desde la racionalidad acotada, exige traducir

principios en reglas ejecutables y trazas auditables: ahí radica la función normativa de lo simbólico. Aunque la inteligencia puede emerger sin símbolos explícitos, su legitimidad pública exige *una capa simbólica capaz de producir razones auditables y controles proporcionales al riesgo*; pudiera denominarse a este criterio *principio de suficiencia justificativa bajo racionalidad acotada*, cuya realización técnica son las arquitecturas neurosimbólicas de gobernanza.

1. El sistema de símbolos físico como fundamento del diseño racional

La conceptualización del sistema de símbolos físicos constituye el núcleo teorético de *Las ciencias de lo artificial*[1] y la clave para entender la visión de Herbert Simon sobre la IA. El tema del procesamiento simbólico es central en el argumento de Simon y de gran importancia para una discusión comparada desde su pensamiento, pertinente para el análisis de la complejidad del diseño e incluso para la psicología cognitiva a través del concepto de símbolo.

1.1. Lo artificial como arquitectura funcional entre objetivos y entorno

Herbert Simon (2006), considera que, especialmente en la fase de diseño, la caracterización de la función, objetivos y adaptación al entorno es la distinción fundamental entre lo artificial y lo natural, en términos tanto imperativos como descriptivos, dando pie posteriormente a su analogía de la mente como un sistema de procesamiento de información.
Al considerar la *función* en la caracterización de lo artificial, Simon pregunta por el propósito buscado por el objeto artificial en relación con su entorno, y encuentra un trinomio de elementos: "el objetivo o propósito, el carácter del artefacto, y el entorno (o ambiente) en que éste actúa" (Simon, 2006, p. 6), en donde el artefacto aparece como una interfase operacional en términos actuales o, aún más, como un amplificador de potencia que divide el espacio observacional en un entorno interno, en donde la materia es organizada en relaciones precisas, vale decir, es diseñada, para adaptarse en alguna medida, a un entorno externo, para cumplir el propósito deseado, configurando una arquitectura de complejidad en el diseño (jerarquías, relaciones).
La racionalidad constituye una restricción fundamental al operar con hipótesis que configuran una comprensión, imperfecta, del ambiente externo, en torno a la toma de decisiones (Simon, 2006). Por lo tanto, la descripción organizativa y funcional de la interfase (el objeto diseñado) es fundamental para el diseño, poniendo las propiedades internas al servicio de objetivos en las propiedades externas. Sin embargo, lo artificial no va a diferenciarse totalmente de lo natural, pues no debe perderse de vista el hecho de que "el sistema interior consiste en una organización de fenómenos naturales capaz de lograr los objetivos deseados dentro de un rango de ambientes" (Simon, 2006, p. 10-13), de forma tal que lo artificial, la organización funcional de lo natural de acuerdo con propósitos predefinidos es la materialización de una teleología funcional,

De tal forma, Simon (2006) señala que los objetivos determinan la conexión entre el interior y el exterior:

> Sí el sistema interno está correctamente diseñado, se adaptará al entorno externo, de forma que su comportamiento quedará determinado en gran parte por el comportamiento de este último, exactamente como ocurría en el caso del hombre económico. Para predecir cómo se comportará, lo único que tenemos que preguntar es ¿Cómo funcionaría en estas circunstancias un sistema diseñado racionalmente? El funcionamiento adopta la forma del entorno en que se mueve (p. 13).

En este sentido, desde la cibernética de sistemas teleológicos, el que el sistema sea reflexivamente un "buen" modelo de su entorno parece ser un requisito funcional de este tipo de los sistemas artificial con propósito (Conant y Ashby, 1970).
La comprensión de cómo configurar los medios para cumplir los fines da lugar a la síntesis de lo artificial, vale decir, el diseño en ingeniería, como una tecnología que hace posible la configuración de medios materiales necesarios para el control de la naturaleza en una interfase de fenómenos naturales, interiores y exteriores, con un comportamiento observable adecuado a fines preconcebidos, dentro de sus márgenes de desempeño, equivalentes a las irracionalidades visibles derivadas de la cognición limitada.
La confiabilidad de lo artificial depende de la organización de los elementos y sus interconexiones de manera específica y su complejidad obedece en gran medida "a la complejidad del entorno al que el programa intenta adaptar su comportamiento" (Simon, 2006, p. 24). En este sentido, un sistema mal adaptado estaría fallando por déficit de complejidad frente a su entorno.
Entonces, para Simon, la relación entre interior y exterior exige una arquitectura representacional justificable, lo que anticipa la relevancia contemporánea del problema simbólico en la gobernanza algorítmica. Si el diseño es, como sostiene Simon, una actividad de razonamiento simbólico orientado a fines, entonces la gobernanza de sistemas que ya no razonan simbólicamente requiere reintroducir condiciones simbólicas de control, para adaptarse a la complejidad del entorno. En los siguientes subapartados se tratará de desglosar las implicaciones de ese requerimiento.

1.2. Racionalidad acotada y la organización del diseño en sistemas complejos

La noción de racionalidad acotada, o la toma de decisiones imperfecta bajo condiciones de información incompleta aparece como un tópico altamente relevante en torno a los procesos de cognición en sistemas complejos adaptativos. Simon argumenta que el diseño (la síntesis de lo artificial) es una actividad intelectual que implica que los sistemas manipulen símbolos y usen representaciones simbólicas para computar cursos alternativos de acción y encontrar soluciones (heurística), sujetas a condiciones de información y cognición limitadas, seleccionado subóptimamente las soluciones:

> *Lo que una persona no puede hacer no lo va a hacer*, por más que quiera. Ante la complejidad de la realidad [se] prefieren procedimientos que hallen respuestas suficientemente buenas a cuestiones cuyas respuestas óptimas son desconocidas. Dado que, en el mundo real, la optimización, con o sin ordenadores, es imposible, el actor económico real es, de hecho, un *satisfictor*, una persona que acepta alternativas bastante buenas, no porque se conformo con menos, sino porque no tiene alternativa (Simon, 2006, p, 33).

Esto introduce una base material de realidad muy potente en el pensamiento de Simon, alejada de la entelequia del *homo oeconomicus* de la economía neoclásica, al reconocer los límites de la decisión y acción posible que hacen posible la modelización evolutiva del diseño de lo artificial, pues "una representación verosímil, [debe] incorporar los límites de procesamiento de la información que sus propios sistemas internos imponen. La imagen también debe incluir la racionalidad consciente de los decisores" (Simon, 2006, p. 58). Por lo tanto, implica una decisión y selección económicamente acotada por parte del diseñador.

1.3. El sistema de símbolos como modelo operativo de la cognición

Reflexionando sobre la naturaleza del ordenador, Simon enuncia "una familia importante de artefactos denominados […] sistemas físicos de símbolos [cuya] adaptabilidad al entorno constituye toda su razón de ser" (Simon, 2006, p. 25), en estos sistemas, y agrega que

> Las estructuras de símbolos pueden servir […] como representaciones internas de los ambientes a los que el sistema de símbolos intenta adaptarse. Permiten modelar aquel ambiente con mayor o menor verosimilitud y con mayor o menor detalle y, en consecuencia, razonar sobre el mismo (p. 26).

La hipótesis empírica sobre el sistema de símbolos define a la inteligencia (de cualquier tipo) como computación: "La inteligencia es producto del trabajo de sistemas de símbolos […] un sistema físico de símbolos [que] tiene los medios *necesarios y suficientes* para la acción inteligente en general" (Simon, 2006, p. 27).
La condición de suficiencia queda evidenciada a partir del diseño de sistemas de símbolos capaces de acción inteligente mientras que la condición de necesareidad, para Simon, se podría comprobar empíricamente al constatar que "todos los sistemas inteligentes conocidos (cerebros y ordenadores) son sistemas de símbolos" (Simon, 2006, p. 27).
Aplicada a la psicología humana, la hipótesis del sistema de símbolos físicos es apoyada por los experimentos de pensamiento en voz alta (protocolos verbales) que muestran que las personas resuelven problemas manipulando símbolos paso a paso, evidenciando empíricamente, para Simon, que la mente humana es un sistema de símbolos físicos.
A partir de la arquitectura conceptual que se ha perfilado, Simon buscó unificar la conceptualización de la inteligencia artificial, de base biológica o de base silicio, y durante algunas décadas caracterizó el paradigma dominante en la ciencia cognitiva y en la inteligencia artificial, sin embargo, han aparecido fisuras en ese entramado teórico, pues si bien, Simon

ubica el origen de la capacidad de síntesis en la inteligencia artificial en el procesamiento de sistemas simbólicos explícitamente formulados, en sus diferentes contextos aplicados, sin embargo, los grandes modelos de lenguaje natural hoy en uso, se basan mayormente en la lógica de redes neuronales en donde no existe una representación simbólica explícita sino más bien matrices de funciones de conectividad y pesos distribuidos en la red, que gestionan su espacio de estado y producen la salida alcanzada a partir de su capa sensorial, enfoque *conexionista* que contrasta con el enfoque *simbólico* de la hipótesis de Simon sobre la inteligencia artificial.

2. Arquitecturas subsimbólicas y su lógica de operación estadística

La conceptualización de los modelos de redes neuronales (Amari, 1998) emerge a partir de analogías de la complejidad del sistema nervioso desde la biología matemática (McCulloch y Pitts, 1943) e hizo posible la síntesis de los grandes modelos de lenguaje (LLM, por sus siglas en inglés) utilizados actualmente. Un problema fundamental en este tipo de sistemas es la operacionalización de una gran cantidad de componentes discretos interrelacionados, problema análogo al de la teoría cinética de los gases, y que ha sido posible analizar a partir de la base teórica de la termodinámica y la mecánica estadística para el cálculo del macroestado de la masa de gas mediante la distribución estocástica de microestados (Bar-Yam, 2018, pp. 58-74; Boltzman, 1973).

La gobernanza de sistemas subsimbólicos puede comprenderse mejor mediante una analogía orientada al control y la ingeniería institucional, pues cuando se construyen sistemas complejos con millones de parámetros entrenados estadísticamente, lo que emerge no es una representación del entorno, sino un patrón de correlaciones internas que permite una competencia funcional sin una comprensión explícita de las reglas que rigen dicho entorno. Desde la perspectiva de la ingeniería, esto equivale a operar con un mecanismo cuyo diagrama interno es inaccesible: se conoce el comportamiento observable, pero no las rutinas internas que generan la salida de la "caja negra".

Esta constricción que pudiéramos denominar, *densificación estadística del comportamiento*: en vez de símbolos articulados y manipulables, el sistema produce decisiones basadas en regularidades de entrenamiento, comprimidas en matrices de pesos. Desde la racionalidad acotada y la justificación pública, dichas regularidades no son equivalente a razones. Son respuestas útiles, pero no explicables.

Dado que es técnicamente inviable rastrear cada parámetro, el reto es reconocer que la estructura misma del sistema subsimbólico impide reconstruir un proceso justificativo comparable al del razonamiento simbólico, de ahí que, para sistemas de alto impacto, se requiere una capa simbólica exógena, con funciones normativas y de diseño, que imponga límites, trazas y condiciones de transparencia, en donde el sistema híbrido asume la forma de un motor subsimbólico heurístico y un regulador simbólico inductivo, que garantice reglas públicas verificables.

Admitiendo que los sistemas subsimbólicos presenten un comportamiento realmente inteligente, y no la mera apariencia mimética de ello, y pese a que esto desafía la condición

de necesareidad de la hipótesis de Simon como un debate abierto en ciencias computacionales, el problema central de su empleo para la complejidad de sistemas es: si la inteligencia puede emerger sin símbolos, la legitimidad no.

3. Desafíos contemporáneos: subsimbolismo, opacidad y límites justificativos

Mientras la hipótesis del sistema físico de símbolos en *Las ciencias de lo artificial* confiere a la inteligencia un carácter explícito y justificable, ya que pensar equivale a manipular símbolos que pueden auditarse y comprenderse (Simon, 2006), haciendo inteligible la cognición y por tanto responsabilizable, la emergencia de arquitecturas subsimbólicas (como las redes neuronales profundas y los LLM) tensionan su alcance, pues producen conductas inteligentes a partir de configuraciones estadísticas distribuidas (Bar-Yam, 2018), que hacen necesario explorar las implicaciones éticas de dicho contraste.

3.1. Implicaciones epistémicas de la opacidad simbólica en la trazabilidad

En primer lugar, cabe preguntarse desde la reflexión epistémica, si acaso no el procesamiento estadístico de los LLM, vale decir, su capacidad para predecir la siguiente palabra con base en patrones probabilísticos equivale genuinamente a razonar. Como ha sido ampliamente señalado, estos modelos aprenden correlaciones en los datos lingüísticos, no necesariamente reglas causales o inferencias lógicas. Su comportamiento inteligente puede ser, en este sentido, una forma sofisticada de imitación sin comprensión. Esta distinción remite a debates clásicos en filosofía de la mente, por ejemplo, el conocido argumento de la "habitación china" de Searle (1980) y sugiere que la inteligencia subsimbólica, al menos en su forma actual, podría carecer de aspectos centrales de la cognición humana, como la capacidad de formular y evaluar argumentos con base en principios normativos.
Ello conlleva (i) un déficit de inteligibilidad normativa porque los modelos explican menos de lo que deciden; (ii) dificultades de imputación ya que, si la autoría efectiva de la decisión queda disuelta en un proceso de correlación estadística, ¿quién responde por el daño posible?; y (iii) asimetría epistémica, dado que, quienes diseñan, ajustan u operan los modelos detentan un saber técnico que no se traduce en razones públicas comprensibles.
En términos éticos, si la inteligencia deja de ser simbólica, también lo hace su justificación moral pues se pasa de decisiones con razones a resultados con métricas, mientras que el criterio de "razón pública" en gobernanza exige pasos justificables (no solo *outputs*). Sin negar la competencia funcional de las LLM, es importante relocalizar el criterio de legitimidad en razones auditables.
Esto compromete su gobernanza en el plano ético, en un sistema simbólico, la cadena de decisión puede, al menos en principio, reconstruirse: reglas, premisas, estados intermedios, conclusiones. Esa trazabilidad semántica facilita la justificación pública de los resultados, la detección de sesgos y la asignación de responsabilidades. Por el contrario,

en los sistemas subsimbólicos los "razonamientos" no están representados en unidades discretas auditables, sino embebidos en pesos y activaciones distribuidas (Burrell, 2016; Mittelstadt *et al.*, 2016).

3.2. Opacidad subsimbólica y gobernanza algorítmica

Una constricción fundamental presente es el hecho de que la opacidad en modelos de aprendizaje profundo no es meramente accidental, sino obedece a limitaciones estructurales que los distancian de la cognición humana: el funcionamiento interno no se deja reescribir en un metalenguaje de reglas sin pérdida sustantiva de información. Las técnicas *post hoc* (mapas de activación) ofrecen indicios, pero no explicaciones operativas (Lipton, 2016; Doshi-Velez y Kim, 2017), son indicios del *output,* pero no razones del proceso y la cuestión es que el desempeño no es equivalente a la legitimidad.

La opacidad afecta a la gobernanza en al menos tres dimensiones, a saber: (i) transparencia y auditabilidad, esto es, los estándares clásicos nacidos para algoritmos simbólicos producen, en este contexto, una falsa sensación de explicabilidad; (ii) regulación y cumplimiento, en el sentido de que las exigencias de reportabilidad o transparencia presuponen unidades semánticas auditables que no existen como tales en modelos distribuidos; y (iii) control democrático, vale decir, sin explicaciones transferibles al lenguaje ordinario, la deliberación pública se empobrece y la autoridad algorítmica se vuelve ilegible (Selbst y Barocas, 2018; Floridi y Cowls, 2019).

Entonces, la opacidad subsimbólica reconfigura el vínculo entre el desempeño que se maximiza y la legitimidad que se ve comprometida, pues sus representaciones no están ancladas en la experiencia en el mundo, sino en relaciones estadísticas entre palabras, a diferencia de la cognición humana, arraigada en la percepción, la acción y la interacción con un entorno compartido.

Ese déficit de realidad parece contradecir la afirmación de Simon acerca de la representación del entorno como fundamento de la adaptabilidad. Incluso podría explicar la perniciosa propensión de los LLM a "alucinar", vale decir, a producir enunciados falsos con plena confianza. No obstante, cabe conjeturarse que su contenido de realidad es aportado por grandes cantidades de datos lingüísticos humanos que los entrenan, los cuales portan implícitamente reglas, conceptos y formas de razonamiento culturalmente mediados. Esto sugiere que la inteligencia emergente de estos modelos no es independiente de los sistemas simbólicos, sino que los presupone en su entrenamiento.

3.3. Racionalidad acotada como premisa ética y de diseño.

En la arquitectura conceptual de Simon, las limitaciones que se vienen señalando adquieren una dimensión particular cuando se consideran a la luz de su concepto de racionalidad acotada. Los agentes deciden bajo límites razonables dadas sus restricciones, pues para Simon, los seres humanos operan bajo límites cognitivos insuperables: información incompleta, tiempo limitado, capacidad de procesamiento finita, inherentes a la cognición natural.

Aceptando esta premisa, análogamente, para la IA contemporánea, con sus ventanas de contexto limitadas y su dependencia de datos muestrales y sus restricciones computacionales, la gobernanza no debe exigir transparencia y explicabilidad totales, aspiración quimérica en sistemas subsimbólicos, sino mecanismos proporcionalmente "acotados" que permitan decisiones suficientemente justificables para el contexto de riesgo.
Ello implica: (i) explicabilidad pragmática, es decir, en vez de reconstrucciones exhaustivas del proceso interno, proveer razones operativas mínimas (propósito, datos críticos, límites y garantías) acordes al impacto de la decisión; (ii) trazabilidad por capas, por ejemplo, combinar bitácoras de uso, metadatos de entrenamiento y controles de entrada/salida puede ofrecer evidencia suficiente sin "abrir" la caja negra; (iii) umbrales de gobernanza dependientes de riesgo, vale decir, a mayor impacto sobre derechos, mayor exigencia de controles *ex ante* (pruebas de sesgo, simulaciones) y de razones *ex post* comprensibles; y (iv) *satisficing* normativo, suficiencia justificativa proporcional al riesgo, vale decir, ajustar el nivel de razón pública al potencial de daño y la afectación de derechos, pues el objetivo no es una justificabilidad perfecta, sino razonable y verificable dadas las constricciones (Floridi y Cowls, 2019), llevando a una definición operativa: llamo *suficiencia justificativa proporcional al riesgo* al criterio según el cual *la carga de razón pública debe ajustarse al potencial de daño y a la afectación de derechos*.
En este sentido, la racionalidad acotada no debilita las exigencias éticas, sino que las hace realizables. Donde la opacidad subsimbólica impide explicaciones completas, el diseño institucional puede compensar la trazabilidad, controles proporcionales y una contabilidad de riesgos que habilite responsabilidades identificables, aun sin símbolos internos (Simon, 2006).

3.4. ¿Puede el diseño simbólico habilitar control, interpretabilidad y responsabilidad?

Como sabemos, los sistemas simbólicos exhiben interpretabilidad intrínseca de sus estructuras lógicas, ontologías y reglas que permiten rastrear la causalidad de la decisión e identificar la norma aplicada y asignar responsabilidades (Doshi-Velez y Kim, 2017; Floridi y Cowls, 2019).
Esta propiedad es normativamente atractiva ya que admite (i) una justificación pública, pues las decisiones pueden acompañarse de razones legibles para las partes afectadas; (ii) una responsabilidad atribuible, ya que es posible localizar dónde falló la inferencia (regla, dato, parámetro) y quién la diseñó o validó; y (iii) un control normativo *ex ante*, es decir, incorporar en su diseño principios éticos, tales como la no discriminación, la proporcionalidad, el debido proceso, el principio *pro persona* o el principio precautorio, entre otros, que pueden codificarse como restricciones o pruebas lógicas previas a la acción. Sin embargo, como ya se ha mencionado, la evidencia empírica contemporánea sugiere que ningún enfoque por sí solo es suficiente, pues los modelos simbólicos explican y justifican, pero no escalan bien en tareas perceptuales, en tanto que los subsimbólicos rinden a gran escala, pero no explican, no obstante, la gobernanza algorítmica requiere articular el núcleo subsimbólico con un marco simbólico normativo capaz de producir razones públicas.

Esto lleva a considerar la relevancia práctica de la IA híbrida o *neurosimbólica*, donde el componente subsimbólico provee una capacidad inductiva y adaptación mientras el componente simbólico actúa como una capa de gobernanza: impone restricciones normativas, habilita trazas auditables y favorece explicaciones operativas (d'Avila Garcez *et al.*, 2009; Besold *et al.*, 2017; 2019), en donde el símbolo recupera su papel no como base única de la inteligencia, sino como infraestructura ética de la decisión.

Conclusiones: arquitectura neurosimbólica para la gobernanza algorítmica

La presente reflexión teórica ha contrastado dos aproximaciones fundamentales a la inteligencia artificial: el enfoque simbólico de Herbert Simon y el enfoque subsimbólico o conexionista, caracterizado por el procesamiento distribuido y estadístico de la información. A lo largo del texto se ha mostrado que, si bien la propuesta de Simon constituyó el paradigma dominante durante décadas y sigue siendo teóricamente poderosa, el desarrollo de modelos conexionistas -particularmente los grandes modelos de lenguaje (LLM)- plantea desafíos significativos a su pretensión de necesidad, pero, sobre todo por su opacidad y subsimbolismo, plantean problemas éticos para su empleo en la gobernanza algorítmica.
La analogía termodinámica explica por qué los sistemas con miles de componentes (como las redes neuronales) pueden exhibir comportamiento inteligente sin recurrir a representaciones simbólicas explícitas, no obstante, para responder al reto ético que implica un cálculo "ciego" a las reglas, la hipótesis de Simon puede matizarse: quizá los símbolos no sean condición necesaria de la inteligencia, pero sí componen una condición operativa para su gobernanza bajo racionalidad acotada. En tal sentido, los símbolos serían menos el sustrato cognitivo universal y más el andamiaje justificativo que hace posible la responsabilidad moral y el control institucional.
Si bien, el sistema de símbolos físicos de Simon, sigue siendo un marco potente para pensar el diseño de lo artificial, el desarrollo del conexionismo obliga a repensar su alcance: Su potencial actualmente puede residir en la posibilidad de habilitar diseños híbridos donde lo simbólico protege la gobernanza de lo subsimbólico.
La racionalidad acotada, que Simon describió como un rasgo inescapable de la cognición, se convierte así en un argumento a favor de la transparencia artificial: si no podemos eliminar los límites, al menos podemos diseñar sistemas cuyos límites sean visibles, auditables y, en última instancia, sujetos a control democrático, lo anterior fundamenta la necesidad de proponer la implementación de un principio de suficiencia justificativa proporcional al riesgo: gobernar lo subsimbólico exige articular el núcleo inductivo con una capa simbólica de control normativo capaz de producir razones públicas acordes al potencial de daño e impacto en derechos.
Esto lleva a proponer la regla de diseño explícito: en sistemas subsimbólicos de alto impacto implementar una capa simbólica que (i) codifique principios en restricciones simbólicas ex ante, (ii) genere trazas auditables, (iii) produzca razones públicas proporcionales al riesgo, y (iv) proponga responsabilidad atribuible (quién, dónde y por qué).

Notas

1. Para el presente texto se cita la obra *The Sciences of Artificial* de Herbert Simon editada por The MIT Press en 1996, editada por COMARES (2006).

Referencias

Amari, S. (1998). Natural gradient Works efficiently in learning. *Neural computation*, 10(2), 251-276.

Bar-Yam, Y. (2018). *Dynamics of complex systems*. Routledge.

Besold, T. R., d'Avila Garcez, A. S., Bader, S., Bowman, H., Domingos, P., Hitzler, P., Kühnberger, K., y otros (2017). Neural-symbolic learning and reasoning: A survey and interpretation. *Dagstuhl Reports*, 6(7), 1–30.

Besold, T. R., d'Avila Garcez, A. S., Bader, S., Bowman, H., Domingos, P., Hitzler, P., Kühnberger, K., y otros (2019). *Neuro-symbolic AI: The state of the art. In AAAI 2019 Spring Symposium Series*. AAAI Press.

Boltzmann, L. (1973). *The Boltzmann equation: Theory and applications*. Cohen, E. y Thirring, W. Springer-Verlag Wien.

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning. *Big Data & Society,* 3(1), 1–12.

Conant, R y Ashby, W. R. (1970). Every Good regulator of a system must be a model of that system. *Int. Jour. Sys. Sc*., 1(2), 89-87.

d'Avila Garcez, A. S., Lamb, L. C., y Gabbay, D. (2009). *Neural-symbolic cognitive reasoning*. Springer.

Doshi-Velez, F., Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608.*

Floridi, L., Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1).

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.08328.*

McCulloch, W y Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*. 5(4). 115-133.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., y Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1–21. https://doi.org/10.1177/2053951716679679

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424.

Selbst, A. D., y Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review,* 87(3), 1085–1139.

Simon, H. A. (2006). *Las ciencias de lo artificial*. Comares.

Abstract: In *The Sciences of the Artificial*, Herbert Simon formulates the hypothesis of a physical system of symbols as the foundation of intelligence, understood as the processing of internal representations oriented towards problem-solving. This text re-examines this hypothesis considering the rise of sub-symbolic architectures, whose competence emerges from distributed statistical configurations without explicit symbolic manipulation. From the notion of bounded rationality, it is argued that, although intelligence can arise without symbols, public legitimacy cannot: the opacity of sub-symbolic models limits traceability and the production of auditable justifications. It is proposed that the governance of high-impact systems requires an exogenous symbolic layer capable of enabling justification proportional to risk. The chapter identifies neuro-symbolic architecture as a suitable design principle for such purposes.

Keywords: symbolic systems - connectionism - artificial intelligence - bounded rationality - algorithmic governance

Resumo: Em *As Ciências do Artificial*, Herbert Simon formula a hipótese de um sistema físico de símbolos como fundamento da inteligência, entendida como o processamento de representações internas orientadas para a resolução de problemas. Este texto reexamina essa hipótese à luz da ascensão das arquiteturas subsimbólicas, cuja competência emerge de configurações estatísticas distribuídas sem manipulação simbólica explícita. A partir da noção de racionalidade limitada, argumenta-se que, embora a inteligência possa surgir sem símbolos, a legitimidade pública não pode: a opacidade dos modelos subsimbólicos limita a rastreabilidade e a produção de justificativas auditáveis. Propõe-se que a governança de sistemas de alto impacto requer uma camada simbólica exógena capaz de permitir a justificação proporcional ao risco. O capítulo identifica a arquitetura neurosimbólica como um princípio de projeto adequado para tais fins.

Palavras-chave: sistemas simbólicos - conexionismo - inteligência artificial - racionalidade limitada - governança algorítmica

[Las traducciones de los abstracts fueron supervisadas por el autor de cada artículo.]

Luis E. Castro Solís es profesor investigador adscrito la Facultad de Ingeniería US de la Universidad Autónoma de Coahuila. Ingeniero civil (UAdeC, 1990), Maestro en Ingeniería Ambiental (ITESM, 2000) y Doctor en Arquitectura, diseño y urbanismo (UAEM, 2015). Es coordinador académico de la Maestría en Investigación Social de la Facultad de Psicología de la UAdeC. Sus líneas de investigación son: ecología política, sistemas socioecológicos y paisaje cultural; intersecciones entre ciencia, tecnología y humanismo. El doctor Castro es Investigador Nacional nivel 1 de la SECIHTI. Correo electrónico: lucastros@uadec.edu.mx