APRENDIZAJE HUMANO Y APRENDIZAJE ARTIFICIAL

TEORÍAS PSICOLÓGICAS COMO BASE PARA EL DISEÑO DE SISTEMAS DE IA QUE REALMENTE APRENDEN

SERGIO COLADO GARCIA | FEBRERO DE 2026| www.sergiocolado.com

INTRODUCCIÓN

En el discurso contemporáneo sobre inteligencia artificial el término aprendizaje se ha convertido en una palabra comodín. Se utiliza para describir cualquier mejora de rendimiento cualquier ajuste de parámetros o cualquier incremento de eficacia medido a través de una métrica.

Esta ampliación semántica ha resultado funcional para la ingeniería. pero profundamente problemática para la comprensión real de lo que hacen los sistemas artificiales. Al identificar aprendizaje con optimización se ha diluido una distinción fundamental que durante más de un siglo ha sido central en la psicología científica la diferencia entre modificar conductas y construir conocimiento.

Desde el punto de vista técnico resulta comprensible esta confusión. Los sistemas de IA aprenden ajustando parámetros mediante procesos de optimización matemática. Sin embargo, esta descripción es incompleta. El hecho de que el aprendizaje se implemente mediante optimización no implica que toda optimización constituya aprendizaje.

Esta confusión ha generado una falsa sensación de progreso intelectual en sistemas que funcionan bien bajo condiciones controladas pero que fallan de forma abrupta cuando el entorno cambia mínimamente. El problema no es la falta de potencia computacional ni de datos sino una comprensión insuficiente de qué significa aprender.

La psicología del aprendizaje humano ha abordado este problema desde múltiples enfoques con una riqueza conceptual y empírica que rara vez se incorpora de forma explícita al diseño de sistemas de inteligencia artificial. Durante décadas se ha investigado cómo los seres humanos adquieren conocimiento cómo integran nueva información cómo regulan su propio proceso de aprendizaje y por qué algunos aprendizajes son duraderos y transferibles mientras otros son frágiles y contextuales. Estas investigaciones no describen estados subjetivos inaccesibles sino procesos funcionales observables medibles y replicables. Ignorar este cuerpo de conocimiento al diseñar IA equivale a renunciar a una fuente contrastada de principios de diseño cognitivo.

Uno de los ejes históricos de este debate ha sido la oposición entre conductismo y cognitivismo. El conductismo defendió que el aprendizaje podía explicarse exclusivamente como modificación de la conducta en función de sus consecuencias observables. Esta postura permitió avances experimentales importantes, pero mostró límites claros cuando se intentó explicar fenómenos como el lenguaje el razonamiento abstracto o la generalización a situaciones nuevas. La emergencia de la psicología cognitiva y del constructivismo supuso un cambio de paradigma al introducir la noción de representaciones internas modelos mentales y estructuras de conocimiento como elementos centrales del aprendizaje.

Este giro cognitivo resulta especialmente relevante para la inteligencia artificial contemporánea. Muchos sistemas actuales utilizan recompensas penalizaciones y señales de error lo que ha llevado a identificar de manera casi automática el aprendizaje por refuerzo con el conductismo clásico. Sin embargo, esta identificación es conceptualmente incorrecta.

El aprendizaje por refuerzo moderno no se limita a moldear conductas observables, sino que opera sobre estados internos funciones de valor y modelos del entorno. El refuerzo no explica el aprendizaje actúa como una señal que corrige predicciones internas. La diferencia entre ambos enfoques no es superficial sino estructural.

La confusión entre optimización y aprendizaje se vuelve especialmente peligrosa cuando se extrapolan capacidades. Un sistema que ha sido optimizado puede mostrar un rendimiento sobresaliente en condiciones conocidas y fracasar estrepitosamente ante una variación mínima. Por su parte, un sistema que ha aprendido en sentido fuerte puede mostrar un rendimiento inicial más modesto, pero adaptarse cuando las condiciones cambian. Esta diferencia no suele capturarse en métricas estándar de evaluación y por ello permanece invisible hasta que el sistema se despliega en entornos reales.

Las teorías psicológicas del aprendizaje ofrecen criterios claros para distinguir entre ambos casos. El constructivismo pone el énfasis en la reorganización de estructuras internas. El aprendizaje significativo subraya la necesidad de anclar la nueva información en conocimientos previos. La teoría de la carga cognitiva muestra que la forma en que se introduce la información condiciona profundamente la calidad del aprendizaje. La metacognición explica cómo los sistemas regulan su propio proceso de aprendizaje detectan errores y ajustan estrategias. Estas teorías no compiten entre sí, sino que describen distintos niveles de un mismo fenómeno la construcción regulada de conocimiento.

Este estudio propone integrar explícitamente estas teorías en el análisis y diseño del aprendizaje en inteligencia artificial. No se trata de humanizar las máquinas ni de atribuirles intenciones o conciencia. Se trata de identificar principios funcionales que permiten que un sistema construya modelos internos robustos generalizables y corregibles. Cuando estos principios están ausentes la optimización produce comportamientos eficaces pero frágiles. Cuando están presentes la optimización se convierte en aprendizaje.

El objetivo de este trabajo es por tanto doble. Por un lado, intenta clarificar conceptualmente qué significa aprender en sistemas artificiales y por qué no puede reducirse a la mejora de una métrica. Por otro lado, trata de mostrar cómo las teorías psicológicas del aprendizaje humano ofrecen un marco sólido para diseñar y evaluar sistemas de IA más robustos adaptativos y fiables.

En un contexto en el que la inteligencia artificial se integra cada vez más en infraestructuras críticas esta distinción deja de ser académica y se convierte en una cuestión de responsabilidad técnica y social.

La pregunta que guía todo el estudio no es si un sistema funciona bien hoy sino qué tipo de conocimiento ha construido y qué ocurrirá cuando el entorno deje de comportarse como durante el entrenamiento.

Responder a esta pregunta exige recuperar una visión exigente del aprendizaje una visión que la psicología científica lleva más de un siglo elaborando y que la ingeniería de la inteligencia artificial no puede seguir ignorando.

EL CONSTRUCTIVISMO COMO MARCO GENERAL DEL APRENDIZAJE

El constructivismo constituye uno de los marcos teóricos más sólidos y persistentes para comprender qué significa aprender en sentido profundo. Frente a concepciones que reducen el aprendizaje a la acumulación de respuestas correctas o al ajuste progresivo de la conducta el constructivismo sostiene que aprender implica la construcción activa de estructuras internas que organizan la experiencia. El conocimiento no se incorpora como un objeto externo, sino que se elabora a partir de esquemas previos que se modifican cuando la realidad los pone en tensión.

En la obra de Jean Piaget esta idea se articula a través de los procesos de asimilación y acomodación. La asimilación permite integrar la experiencia dentro de estructuras existentes mientras que la acomodación implica modificar dichas estructuras cuando resultan insuficientes. Aprender no es por tanto reforzar lo que ya funciona sino reorganizar el sistema interno cuando las expectativas fallan. Esta noción resulta especialmente relevante para la inteligencia artificial porque introduce un criterio claro para distinguir entre ajuste local y aprendizaje estructural.

Pruebas operatorias. Jean Piaget y Constance Kamii (fuente: Fondation Jean Piaget)

Cuando un sistema de IA se limita a ajustar parámetros para mejorar una métrica sin reorganizar sus representaciones internas el resultado es una optimización eficaz pero frágil. El sistema responde correctamente mientras el entorno se comporta de manera similar a los datos de entrenamiento, pero carece de recursos para reinterpretar la experiencia cuando aparecen situaciones nuevas. Desde una perspectiva constructivista esto no constituye aprendizaje sino adaptación superficial. El sistema no ha construido conocimiento, sino que ha explotado regularidades locales.

El constructivismo subraya además que el aprendizaje es un proceso activo. El sujeto no recibe información de forma pasiva, sino que interactúa con el entorno formula expectativas implícitas y las contrasta con la experiencia. En inteligencia artificial este principio se manifiesta en sistemas que aprenden a partir de la interacción y no solo del análisis estático de datos. Sin embargo, la actividad por sí sola no garantiza aprendizaje. La evidencia psicológica muestra que la exploración sin estructura conduce con frecuencia a aprendizajes inestables y poco transferibles. La actividad debe conducir a una reorganización interna para que pueda hablarse de aprendizaje.

Otro aporte central del constructivismo es la idea de que el conocimiento se organiza jerárquicamente y que los conceptos más abstractos emergen a partir de la integración de experiencias concretas. En IA este principio se refleja en arquitecturas que desarrollan representaciones latentes progresivamente más abstractas. Cuando estas representaciones capturan regularidades profundas del entorno el sistema puede reutilizarlas en contextos distintos. Cuando se limitan a codificar patrones superficiales la generalización es pobre.

El constructivismo también cuestiona la idea de que aprender sea un proceso lineal y acumulativo. Aprender implica reorganizaciones periódicas en las que parte del conocimiento previo se reinterpreta a la luz de nueva información. En sistemas artificiales este fenómeno se observa cuando la incorporación de nueva experiencia obliga a revisar representaciones previas. La capacidad de realizar esta reorganización sin destruir el conocimiento existente es un desafío técnico central y uno de los puntos donde la inspiración constructivista resulta más valiosa.

Desde esta perspectiva el aprendizaje en inteligencia artificial no puede evaluarse únicamente por su rendimiento inmediato. Debe evaluarse por la calidad de las representaciones internas que construye y por su capacidad para sostener la adaptación a lo largo del tiempo. Un sistema que aprende en sentido constructivista es capaz de integrar información nueva sin reiniciar el proceso desde cero. Puede reinterpretar su experiencia y modificar sus modelos internos de forma coherente.

El constructivismo no propone imitar el desarrollo humano ni replicar sus etapas evolutivas. Propone principios funcionales sobre cómo se construye el conocimiento. Estos principios son directamente aplicables al diseño de sistemas de IA cuando el objetivo no es solo optimizar una tarea concreta sino construir sistemas capaces de aprender de manera robusta. Adoptar el constructivismo como marco general no implica renunciar a la optimización matemática. Implica entenderla como un medio y no como una explicación suficiente del aprendizaje.

En el contexto de la inteligencia artificial contemporánea el constructivismo ofrece una advertencia clara. Sin reorganización interna no hay aprendizaje solo ajuste. Sin estructuras que integren la experiencia no hay generalización solo repetición. Comprender y aplicar esta distinción es un paso necesario para avanzar hacia sistemas que no solo funcionen bien en condiciones conocidas, sino que puedan enfrentarse a la incertidumbre del mundo real.

APRENDIZAJE SIGNIFICATIVO Y ANCLAJE DEL CONOCIMIENTO

El concepto de aprendizaje significativo introduce una precisión esencial en la comprensión de cómo se construye el conocimiento y por qué algunos aprendizajes son duraderos y transferibles mientras otros se desvanecen o permanecen ligados a contextos muy específicos. Desde esta perspectiva aprender no consiste en incorporar información nueva de manera aislada sino en integrarla dentro de una estructura cognitiva previa que le da sentido. Sin este proceso de anclaje el aprendizaje se vuelve mecánico frágil y altamente dependiente de las condiciones en las que se produjo.


David Asubel

La formulación clásica del aprendizaje significativo desarrollada por David Ausubel subraya que el factor más importante para aprender es lo que el sujeto ya sabe. La nueva información adquiere significado cuando puede relacionarse de manera sustantiva y no arbitraria con conocimientos existentes. Cuando esta relación no se produce el resultado puede ser una mejora temporal del rendimiento, pero no una integración estable del conocimiento.

Este principio resulta directamente aplicable al análisis del aprendizaje en inteligencia artificial. En sistemas de IA el anclaje del conocimiento se manifiesta en la forma en que las nuevas experiencias modifican las representaciones internas existentes.

Un sistema que carece de estructuras previas coherentes puede memorizar grandes volúmenes de datos, pero tiene dificultades para integrar información nueva sin degradar lo aprendido. Este fenómeno se observa en problemas como el olvido catastrófico donde la incorporación de nuevos datos destruye representaciones anteriores. Desde la perspectiva del aprendizaje significativo este problema no es accidental sino consecuencia de una arquitectura que no permite el anclaje estable del conocimiento.

El aprendizaje significativo también pone de relieve la importancia del contexto inicial en el que se produce el aprendizaje. En educación humana Ausubel introdujo el concepto de organizadores previos como estructuras que preparan al aprendiz para integrar la nueva información. En inteligencia artificial este principio encuentra su equivalente funcional en el preentrenamiento de representaciones base en la inicialización informada de modelos y en el uso de contextos que orientan la interpretación de los datos entrantes. Estos mecanismos no simplifican el problema, sino que reducen la carga innecesaria y facilitan la construcción de representaciones profundas.

Otro aspecto central del aprendizaje significativo es la estabilidad del conocimiento frente al cambio. Cuando la nueva información se integra de manera significativa el sistema puede reorganizar sus representaciones sin perder coherencia. En IA esto se traduce en la capacidad de incorporar nuevos dominios tareas o contextos sin reiniciar el aprendizaje desde cero. Un sistema que aprende de forma significativa puede ampliar su conocimiento sin destruir sus modelos previos porque la integración se produce a nivel estructural y no como una superposición de patrones independientes.

El aprendizaje significativo también explica por qué el aumento masivo de datos no garantiza una mejora cualitativa del aprendizaje. Sin mecanismos que permitan relacionar la nueva información con estructuras existentes los datos adicionales pueden incluso degradar el rendimiento al introducir ruido o correlaciones espurias. Desde esta perspectiva la calidad del aprendizaje depende menos de la cantidad de datos y más de la capacidad del sistema para integrarlos de forma coherente.

En el diseño de arquitecturas de IA adoptar el principio del aprendizaje significativo implica prestar atención a la continuidad representacional. Las nuevas experiencias deben modificar y enriquecer las representaciones existentes no sustituirlas de forma abrupta. Esto requiere mecanismos de consolidación y de control que permitan evaluar cómo se integra la información nueva. Sin estos mecanismos la optimización produce mejoras locales, pero no aprendizaje acumulativo.

El aprendizaje significativo aporta además un criterio claro para evaluar la transferencia. Un sistema que ha integrado el conocimiento de manera significativa puede aplicar lo aprendido en contextos distintos porque las representaciones internas no están ligadas a instancias concretas sino a relaciones funcionales. Cuando este tipo de transferencia no se produce el aprendizaje ha sido superficial, aunque el rendimiento aparente sea elevado.

En el contexto de la inteligencia artificial contemporánea el aprendizaje significativo ofrece una advertencia y una oportunidad. La advertencia es que sin anclaje del conocimiento la optimización produce sistemas frágiles. La oportunidad es que incorporar este principio permite diseñar sistemas capaces de aprender de forma acumulativa estable y generalizable.

En un entorno cambiante esta capacidad no es un refinamiento teórico sino una condición esencial para que la inteligencia artificial pueda operar de manera fiable.

DESCUBRIMIENTO GUIADO ANDAMIAJE Y CURRÍCULO

El aprendizaje por descubrimiento ha sido a menudo interpretado de forma simplificada como una defensa de la exploración libre y de la ausencia de guía. Esta interpretación no solo es incorrecta, sino que contradice tanto la obra original de Jerome Bruner como la evidencia empírica acumulada posteriormente. El descubrimiento al que se refiere Bruner no es un proceso espontáneo ni desestructurado sino una actividad cognitiva cuidadosamente guiada cuyo objetivo es facilitar la construcción de representaciones profundas y transferibles.

Bruner sostiene que el aprendizaje es más robusto cuando el sujeto participa activamente en la construcción del conocimiento, pero también subraya que esta actividad debe estar orientada. El descubrimiento guiado no elimina la estructura, sino que la dosifica. La guía no anula el aprendizaje, lo que permite es hacerlo posible al reducir la carga innecesaria y al focalizar la atención en relaciones relevantes. Sin esta orientación la exploración tiende a dispersarse y a producir aprendizajes locales poco estables.

Este principio resulta especialmente relevante en inteligencia artificial donde la exploración sin restricciones puede conducir a soluciones espurias o a una dependencia excesiva del azar. Sistemas que exploran sin guía suelen necesitar enormes volúmenes de experiencia para alcanzar un rendimiento aceptable y aun así presentan dificultades para generalizar. El descubrimiento guiado en IA se traduce en la introducción de restricciones iniciales ejemplos trabajados señales auxiliares o modelos parciales que orientan la exploración hacia regiones del espacio de estados más informativas.

Vygotsky

El concepto de andamiaje profundiza en esta idea al describir cómo la ayuda externa permite realizar tareas que el sistema no podría completar de manera autónoma. El andamiaje no sustituye al aprendizaje, sino que crea las condiciones para que este ocurra. En psicología del desarrollo este concepto está estrechamente vinculado a la teoría sociocultural de Lev Vygotsky donde el aprendizaje se produce primero en interacción con otros y posteriormente se internaliza. La ayuda se retira de forma progresiva a medida que el aprendiz adquiere control sobre la tarea.

En inteligencia artificial el andamiaje adopta formas funcionales como la provisión inicial de soluciones parciales la descomposición de tareas complejas en subtareas o el uso de modelos auxiliares que guían el aprendizaje. A medida que el sistema desarrolla representaciones internas más estables estas ayudas pueden retirarse sin que el rendimiento colapse. Este proceso no es trivial ni automático, pero constituye una de las estrategias más eficaces para facilitar aprendizajes profundos sin incurrir en costes computacionales prohibitivos.

El concepto de currículo complementa el andamiaje al introducir la dimensión temporal del aprendizaje. Aprender no es enfrentarse desde el inicio a la máxima complejidad sino recorrer una secuencia de experiencias ordenadas que respeten la capacidad actual del sistema. Bruner propuso la idea de currículo espiral para describir cómo los mismos conceptos pueden revisitarse de manera recurrente aumentando progresivamente su nivel de abstracción. Este principio ha encontrado una traducción directa en la inteligencia artificial moderna a través de estrategias de entrenamiento progresivo.

Un currículo bien diseñado no simplifica el problema, sino que organiza la experiencia de manera que el sistema pueda construir representaciones estables antes de enfrentarse a variaciones más complejas. Cuando la complejidad se introduce de forma abrupta el aprendizaje tiende a fragmentarse y a depender de correlaciones accidentales. Cuando se introduce de forma progresiva el sistema tiene la oportunidad de consolidar patrones y de reorganizar sus representaciones internas de manera coherente.

La combinación de descubrimiento guiado andamiaje y currículo permite resolver una tensión central en el aprendizaje tanto humano como artificial. Por un lado, es necesario que el sistema explore y construya activamente su conocimiento. Por otro lado, esa exploración debe estar estructurada para que sea cognitivamente productiva. Ni la instrucción totalmente dirigida ni la exploración totalmente libre producen aprendizajes profundos de forma consistente. El equilibrio entre ambos extremos es una condición del aprendizaje robusto.

Desde una perspectiva constructivista estas estrategias no son técnicas pedagógicas accesorias sino mecanismos fundamentales de construcción del conocimiento. En inteligencia artificial su incorporación marca la diferencia entre sistemas que requieren ajustes constantes y sistemas que desarrollan competencias estables y reutilizables. El descubrimiento guiado el andamiaje y el currículo no limitan la inteligencia del sistema, sino que la hacen posible al permitir que la optimización se traduzca en aprendizaje estructural.

Exploración libre, carga cognitiva y aprendizaje productivo

En aprendizaje humano y artificial la exploración no es informativa por sí misma. Para que una experiencia genere aprendizaje debe poder ser integrada en una estructura interna existente. Cuando esa estructura no existe o es débil la exploración libre satura la capacidad de procesamiento del sistema y produce ruido en lugar de conocimiento.

Desde la teoría de la carga cognitiva esto ocurre porque el sistema se ve obligado a procesar simultáneamente demasiadas variables sin disponer de esquemas previos que las organicen. El resultado no es descubrimiento sino dispersión. El sistema puede encontrar soluciones locales, pero no construye representaciones reutilizables.

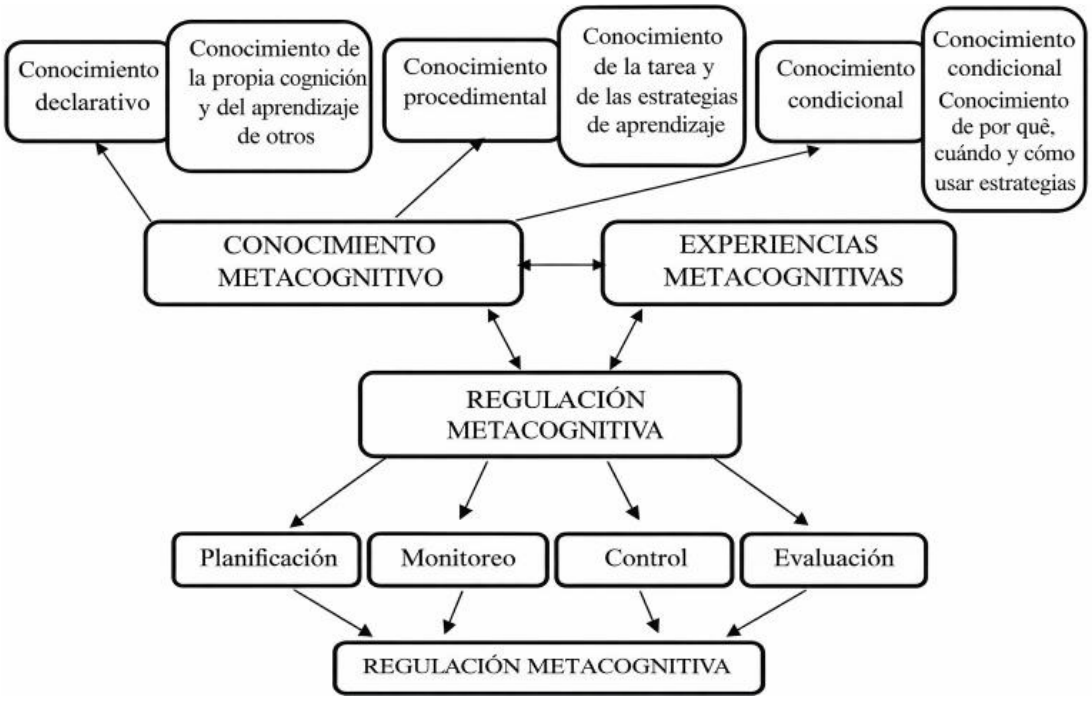
La guía actúa reduciendo el espacio de búsqueda inicial y focalizando la atención en relaciones relevantes. No proporciona la solución final, sino que limita la complejidad hasta que el sistema puede representarla internamente. Una vez que estas representaciones se estabilizan la guía deja de ser necesaria y puede retirarse progresivamente sin pérdida de rendimiento.

Este mecanismo explica por qué la exploración guiada produce aprendizajes más profundos que la exploración libre tanto en humanos como en sistemas de inteligencia artificial. La guía no limita el aprendizaje, lo hace posible al crear las condiciones cognitivas para que la experiencia sea informativa.

METACOGNICIÓN AUTORREGULACIÓN Y CORRECCIÓN PREDICTIVA

La metacognición introduce un nivel decisivo en la comprensión del aprendizaje porque desplaza el foco desde la mera adquisición de conocimiento hacia el control del propio proceso de aprender. En psicología este concepto se refiere a la capacidad de un sistema para evaluar su propio funcionamiento detectar errores anticipar dificultades y modificar estrategias en consecuencia. No se trata de un añadido superficial sino de un mecanismo central que explica por qué algunos aprendizajes se consolidan y otros se estancan o se degradan con el tiempo.

El término metacognición fue formalizado por John H. Flavell para describir el conocimiento que un sujeto tiene sobre sus propios procesos cognitivos y la capacidad de regularlos. Sin embargo, los fundamentos funcionales de esta capacidad pueden rastrearse en la obra de Lev Vygotsky quien explicó cómo los procesos de control cognitivo emergen inicialmente de la regulación externa y se internalizan progresivamente. Desde esta perspectiva la autorregulación no es un rasgo innato sino una construcción que se apoya en la interacción y en la mediación.



Los componentes de la metacognición (según Flavell, 1979; Brown, 1987; Lee et al., 2009; Bryce & Whitebread, 2012; Marulis, 2014; Robson, 2016).)

En el aprendizaje humano la metacognición permite detectar cuándo una estrategia deja de ser eficaz cuándo una comprensión es incompleta y cuándo es necesario revisar supuestos previos. Gracias a este nivel de control el aprendizaje puede acumularse sin reiniciarse ante cada error. Sin metacognición el sujeto puede mejorar su rendimiento en tareas específicas, pero carece de herramientas para evaluar la validez de su propio conocimiento en contextos nuevos. El resultado es un aprendizaje frágil dependiente del contexto y difícil de transferir.

En inteligencia artificial la metacognición no implica introspección ni conciencia. Se manifiesta como un conjunto de mecanismos funcionales que permiten al sistema evaluar la fiabilidad de sus propias predicciones y decidir cuándo es necesario revisar su modelo interno. Esto incluye la estimación de incertidumbre la detección de inconsistencias entre expectativas y resultados y la activación de procesos de revisión o exploración adicional. Un sistema que carece de estos mecanismos puede optimizar su comportamiento mientras las condiciones se mantienen estables, pero no tiene forma de saber cuándo su comprensión del entorno ha dejado de ser válida.

El aprendizaje por refuerzo moderno encuentra aquí su fundamento más profundo cuando se interpreta como un proceso de corrección predictiva. El agente no aprende porque recibe premios o castigos sino porque compara lo que esperaba que ocurriera con lo que realmente ocurre. La diferencia entre ambos resultados señala un error en el modelo interno y desencadena su ajuste. El refuerzo no modifica directamente la acción sino la predicción que la sustenta. En este punto el aprendizaje deja de ser reactivo y se convierte en un proceso de refinamiento del conocimiento.

La existencia de modelos internos es lo que permite que esta corrección predictiva genere aprendizaje en sentido fuerte. Un modelo interno codifica relaciones causales regularidades temporales y dependencias contextuales. Cuando el sistema corrige este modelo a partir de la experiencia no está memorizando un caso aislado sino reorganizando su comprensión funcional del entorno. Como consecuencia puede anticipar situaciones futuras y actuar de manera adaptativa incluso en contextos no vistos durante el entrenamiento.

La metacognición amplifica este proceso al introducir la capacidad de evaluar la propia fiabilidad. Un sistema metacognitivo no solo corrige errores después de que ocurran, sino que aprende a reconocer estados en los que sus predicciones son inciertas. En estos casos puede adoptar estrategias más conservadoras buscar información adicional o modificar su patrón de exploración. Esta capacidad de reconocer la propia ignorancia resulta crítica en entornos complejos y cambiantes donde la optimización ciega puede conducir a errores acumulativos.

Desde una perspectiva constructivista la metacognición no es un componente opcional sino un elemento estructural del aprendizaje. Permite integrar nueva información sin destruir el conocimiento previo y facilita la adaptación progresiva a contextos cambiantes. En inteligencia artificial esto se traduce en arquitecturas que separan la representación del control que permiten revisar supuestos internos y que incorporan mecanismos explícitos de evaluación de la propia actuación.

La ausencia de metacognición conduce a sistemas que pueden parecer competentes mientras el entorno se comporta de forma predecible pero que fallan de manera silenciosa cuando las condiciones cambian. Estos sistemas no detectan el error hasta que el rendimiento se degrada de forma evidente y para entonces la corrección puede ser costosa o imposible. Por el contrario, un sistema con capacidades de autorregulación puede detectar señales tempranas de fallo y ajustar su modelo antes de que el error se propague.

En el contexto de la inteligencia artificial aplicada a entornos reales la metacognición se convierte así en un requisito de robustez. Aprender no es solo ajustar predicciones sino aprender a evaluar cuándo esas predicciones dejan de ser fiables. Solo cuando un sistema incorpora este nivel de control el aprendizaje se vuelve acumulativo adaptable y sostenible en el tiempo. En ausencia de metacognición la optimización produce rendimiento. Con metacognición la optimización puede transformarse en aprendizaje genuino.

POR QUÉ EL CONDUCTISMO RESULTA INSUFICIENTE

El conductismo ha desempeñado un papel relevante en la historia de la psicología científica al proporcionar un marco experimental riguroso para el estudio del aprendizaje. Sin embargo, sus limitaciones conceptuales se hacen evidentes cuando se intenta explicar formas de aprendizaje que implican generalización comprensión y adaptación a contextos nuevos. En el contexto de la inteligencia artificial estas limitaciones no solo persisten, sino que se amplifican debido a la complejidad y variabilidad de los entornos en los que los sistemas deben operar.

El conductismo radical formulado por B. F. Skinner define el aprendizaje como un cambio en la probabilidad de una respuesta observable en función de sus consecuencias. Esta definición prescinde deliberadamente de los estados internos como elementos explicativos. El comportamiento mejora y con ello el aprendizaje queda explicado. Esta postura resulta coherente para describir el moldeado de conductas simples, pero se vuelve insuficiente cuando el comportamiento exitoso requiere anticipar consecuencias integrar información a lo largo del tiempo o aplicar conocimientos a situaciones no vistas.

Uno de los principales problemas del conductismo es su incapacidad para explicar la transferencia. Un organismo o un sistema puede aprender a responder de manera eficaz a un conjunto específico de estímulos y fracasar cuando estos cambian ligeramente. Desde una perspectiva conductista este fracaso no constituye un problema teórico simplemente indica que el aprendizaje no se ha producido en esas nuevas condiciones. Desde una perspectiva cognitiva este comportamiento revela la ausencia de un modelo interno que capture las regularidades profundas del entorno.

En inteligencia artificial esta limitación se manifiesta de forma clara en sistemas que optimizan su comportamiento en entornos estáticos, pero colapsan ante variaciones mínimas. El sistema ha sido condicionado para responder a patrones específicos, aunque no ha construido conocimiento reutilizable. Este tipo de comportamiento puede producir resultados impresionantes en entornos controlados y al mismo tiempo carecer de valor práctico en escenarios reales donde la variabilidad es la norma.

Otro límite estructural del conductismo es su incapacidad para explicar el papel de la anticipación. El aprendizaje conductista se basa en la asociación retrospectiva entre una respuesta y sus consecuencias. Sin embargo, gran parte del comportamiento inteligente depende de la capacidad de anticipar consecuencias futuras antes de actuar. En sistemas artificiales esta anticipación se implementa mediante modelos internos y simulación. El conductismo carece de herramientas conceptuales para explicar este tipo de procesamiento porque rechaza la relevancia explicativa de los estados internos.

El conductismo tampoco ofrece un marco adecuado para comprender el error como fuente de aprendizaje estructural. En su formulación clásica el error es simplemente una respuesta no reforzada. En enfoques cognitivos y constructivistas el error es una señal de que el modelo interno es incorrecto y debe ser revisado.

Esta diferencia no es semántica. Determina si el sistema aprende de sus fallos o simplemente evita repetirlos en condiciones similares.

Cuando se equipara el aprendizaje por refuerzo con el conductismo se incurre en una simplificación peligrosa. El aprendizaje por refuerzo moderno utiliza recompensas y penalizaciones, pero opera sobre estados internos funciones de valor y modelos del entorno. El refuerzo no es una explicación del aprendizaje sino una señal que corrige predicciones. Ignorar esta diferencia conduce a diseñar sistemas centrados exclusivamente en maximizar recompensas inmediatas sin prestar atención a la calidad de las representaciones internas que se construyen.

La insuficiencia del conductismo se hace especialmente evidente en entornos no estacionarios donde las regularidades cambian con el tiempo.

Un sistema conductista puede optimizar su comportamiento mientras las contingencias se mantienen estable, pero carece de mecanismos para detectar cuándo estas han cambiado. En ausencia de metacognición y de modelos internos el sistema continúa aplicando respuestas que ya no son adecuadas hasta que el rendimiento se degrada de forma drástica.

Descartar el conductismo como marco explicativo no implica negar su valor histórico ni rechazar el uso del refuerzo como herramienta técnica. Implica reconocer que el aprendizaje en sistemas complejos requiere algo más que la asociación entre respuestas y consecuencias. Requiere la construcción de modelos internos la capacidad de anticipar y la regulación del propio proceso de aprendizaje. Sin estos elementos el comportamiento puede mejorar, pero el conocimiento no se construye.

En el diseño de sistemas de inteligencia artificial esta distinción es crítica.

Un sistema basado exclusivamente en principios conductistas puede ser eficaz en tareas simples y entornos estables. Un sistema diseñado a partir de principios cognitivos y constructivistas tiene mayores posibilidades de adaptarse a la incertidumbre del mundo real.

La insuficiencia del conductismo no es un defecto histórico superado sino una advertencia permanente sobre los límites de cualquier enfoque que pretenda explicar el aprendizaje sin recurrir a la estructura interna del sistema.

CONCLUSIÓN

El recorrido realizado a lo largo de este artículo permite sostener con claridad una afirmación que resulta fundamental para el desarrollo presente y futuro de la inteligencia artificial aprender no es equivalente a optimizar. La optimización es el mecanismo técnico mediante el cual se ajustan parámetros y se mejoran métricas, pero el aprendizaje es un fenómeno estructural que solo emerge cuando ese ajuste da lugar a representaciones internas capaces de organizar la experiencia anticipar consecuencias y sostener la adaptación al cambio. Confundir ambos planos conduce a una sobreestimación sistemática de las capacidades reales de los sistemas artificiales.

Las teorías psicológicas del aprendizaje humano no han sido introducidas como analogías retóricas ni como intentos de humanizar la tecnología. Han sido utilizadas como marcos funcionales que permiten identificar con precisión qué condiciones hacen posible el aprendizaje en sentido fuerte. El constructivismo muestra que sin reorganización interna no hay conocimiento sino ajuste local. El aprendizaje significativo evidencia que sin anclaje en estructuras previas la información nueva no se integra de forma estable. El descubrimiento guiado el andamiaje y el currículo explican por qué la actividad sin estructura produce aprendizajes frágiles. La metacognición revela que sin autorregulación el aprendizaje no puede sostenerse en entornos cambiantes.

Desde esta perspectiva el conductismo aparece como un marco insuficiente para explicar el aprendizaje en sistemas complejos. Su valor histórico y experimental es indiscutible pero su renuncia explícita a los estados internos lo incapacita para dar cuenta de fenómenos centrales como la generalización la anticipación y la transferencia. Equiparar el aprendizaje por refuerzo con el conductismo supone ignorar la evolución conceptual y técnica de la inteligencia artificial moderna donde el refuerzo actúa como señal de corrección de predicciones y no como explicación completa del aprendizaje.

Uno de los resultados más relevantes de este análisis es la identificación del modelo interno como eje central del aprendizaje. Un sistema aprende cuando construye y refina un modelo funcional del entorno que puede ser corregido a partir de la experiencia. Este modelo no necesita ser explícito ni interpretable en términos humanos, pero debe permitir anticipar consecuencias y reorganizarse cuando las predicciones fallan. Sin este tipo de estructura la optimización produce comportamientos eficaces, aunque altamente dependientes del contexto.

La metacognición añade un nivel adicional que resulta crítico para la robustez. Aprender no es solo corregir errores sino aprender a detectar cuándo el propio conocimiento deja de ser fiable. En ausencia de este control los sistemas pueden continuar optimizando su comportamiento incluso cuando el entorno ha cambiado de forma sustancial. La autorregulación permite detectar el desajuste antes de que el fallo sea irreversible y constituye una condición necesaria para el despliegue seguro de la IA en entornos reales.

Las implicaciones prácticas de este marco son profundas. Diseñar sistemas de IA exclusivamente orientados a maximizar métricas de rendimiento inmediato conduce a soluciones frágiles que funcionan mientras el mundo se comporta como durante el entrenamiento. Diseñar sistemas que incorporen principios constructivistas aprendizaje significativo andamiaje progresivo y control metacognitivo conduce a aprendizajes más lentos en las fases iniciales pero más estables y transferibles a largo plazo. Esta elección no es técnica en sentido estrecho sino estratégica.

En un contexto en el que la inteligencia artificial comienza a operar en dominios críticos como la gestión urbana la seguridad la toma de decisiones automatizada o los sistemas de control complejos la diferencia entre optimizar y aprender adquiere una dimensión ética y social. Sistemas que solo optimizan pueden generar una ilusión de competencia hasta que un cambio no previsto desencadena errores sistémicos. Sistemas que aprenden en sentido fuerte tienen mayor capacidad de adaptación, pero también exigen criterios de evaluación más exigentes y menos inmediatos.

Este trabajo no pretende cerrar el debate sino elevar su nivel. La pregunta relevante ya no es si un sistema obtiene buenos resultados en una tarea concreta sino qué tipo de conocimiento ha construido y hasta qué punto puede sobrevivir a la incertidumbre. Responder a esta pregunta exige abandonar simplificaciones cómodas y recuperar una concepción exigente del aprendizaje basada en décadas de investigación científica. Solo desde esta perspectiva la inteligencia artificial puede aspirar a ser no solo eficiente sino verdaderamente inteligente y responsable.

REFERENCIAS

- Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. New York, NY: Holt, Rinehart and Winston.
- Bjork, R. A., & Bjork, E. L. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. En M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). New York, NY: Worth Publishers.
- Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review, 31*(1), 21–32. https://doi.org/10.17763/haer.31.1.j2324t7415435n25
- Bruner, J. S. (1986). *Actual minds, possible worlds*. Cambridge, MA: Harvard University Press.
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906
- Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1
- Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511811678
- Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford, UK: Oxford University Press.
- Piaget, J. (1970). *Science of education and the psychology of the child*. New York, NY: Orion Press.
- Skinner, B. F. (1953). *Science and human behavior*. New York, NY: Macmillan.
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). Cambridge, MA: MIT Press.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
- Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York, NY: Springer. https://doi.org/10.1007/978-1-4419-8126-4
- Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review Monograph Supplements, 2*(4), 1–109. https://doi.org/10.1037/h0092987
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x
.

APRENDIZAJE HUMANO Y APRENDIZAJE ARTIFICIAL

TEORÍAS PSICOLÓGICAS COMO BASE PARA EL DISEÑO DE SISTEMAS DE IA QUE REALMENTE APRENDEN

© 2026 Sergio Colado García - scolado@nechigroup.com

Citación

Colado García, Sergio (2026). *Aprendizaje humano y aprendizaje artificial. Teorías psicológicas como base para el diseño de sistemas de IA que realmente aprenden.* Disponible en: www.sergiocolado.com

La omisión de esta referencia o la alteración del sentido del contenido no está autorizada.