MODELOS DEL MUNDO

ARQUITECTURA DE LA REALIDAD Y EL FUTURO DE LA INTELIGENCA ARTIFICIAL

Pedro Salcedo Lagos

Modelos del Mundo

Arquitectura de la Realidad y el Futuro de la Inteligencia Artificial

Pedro Salcedo Lagos

Dedicatoria

A la memoria del Dr. José Mira Mira, catedrático, mentor y arquitecto de mentes.

Fuiste tú quien modeló mi mundo intelectual, enseñándome con la paciencia del sabio y el rigor del ingeniero que entre el símbolo computacional y la experiencia vital existe un "abismo ontológico" que ninguna ingeniería, por sofisticada que sea, debe ignorar. Me enseñaste a mirar los circuitos con ojos de biólogo y la biología con la precisión del físico, estableciendo una distinción que hoy, más que nunca, actúa como el faro que ilumina las tinieblas de la confusión moderna: la diferencia entre la Inteligencia Artificial como Ciencia (el análisis de lo natural) y la Inteligencia Artificial como Ingeniería del Conocimiento (la síntesis de lo artificial).

Este libro es un intento de caminar sobre el puente que tú construiste, explorando los nuevos horizontes de los Grandes Modelos de Mundo (Large World Models) y la Inteligencia Artificial Generativa, sin olvidar jamás el suelo firme de nuestra propia humanidad. Es un esfuerzo por honrar tu insistencia en que en la biología no existen símbolos "fríos" y arbitrarios como en la computación, sino estados físicos dinámicos que solo cobran sentido dentro de la historia del organismo. En tu honor, buscamos aquí no solo calcular, sino comprender.

Contenido

Prólogo: La Arquitectura del Habitar

Escribir sobre Inteligencia Artificial en el año 2025 implica enfrentarse a una paradoja vertiginosa: nunca antes habíamos tenido máquinas tan capaces de procesar el lenguaje, y nunca antes había sido tan evidente que el lenguaje, por sí solo, no basta para comprender el mundo.

Este libro nace de una inquietud que es, a la vez, técnica y profundamente humana. Durante décadas, desde las aulas de la Universidad de Concepción hasta los laboratorios de investigación, he sido testigo y partícipe de la evolución de sistemas que intentan mimetizar la cognición. Sin embargo, la reciente explosión de los Grandes Modelos de Lenguaje (LLMs) nos ha dejado fascinados ante un espejo que nos devuelve una imagen nítida pero plana de nuestra propia inteligencia: símbolos perfectamente ordenados, pero carentes de la "facticidad" de la vida.

La tesis central que recorre estas páginas es que estamos ante un cambio de paradigma inevitable: el tránsito de modelos que leen el mundo a modelos que intentan habitarlo. Estamos pasando de la Inteligencia Artificial Generativa textual a los Grandes Modelos de Mundo (LWMs). Pero, ¿qué significa realmente que una máquina tenga un "modelo del mundo"?

Para responder a esto, no basta con examinar la arquitectura de los transformadores o los vectores latentes; es imperativo descender primero al sustrato biológico. Como me enseñó mi mentor, el Dr. José Mira, existe un "abismo ontológico" entre el símbolo computacional y la experiencia vital. Mientras que el ser biológico construye su realidad desde la vulnerabilidad de un cuerpo que respira y debe sobrevivir (autopoiesis), la IA construye la suya desde la seguridad de un sistema optimizado por ingenieros (alopoiesis).

En esta obra, invito al lector a caminar sobre el filo de esa distinción.

En la primera parte, exploramos la Biosofía de la Cognición, argumentando que la verdadera inteligencia es encarnada y que un modelo del mundo que no contemple la finitud o el "aliento" es

necesariamente incompleto. Recurrimos a la fenomenología de Heidegger y Merleau-Ponty para recordar que no construimos el mundo para luego ocuparlo, sino que lo "habitamos" y, al hacerlo, emerge el significado.

Posteriormente, nos adentramos en el Eje Tecnológico, diseccionando cómo gigantes como Google DeepMind, Meta y World Labs están intentando cerrar la brecha entre la simulación y la realidad. Analizamos arquitecturas fascinantes como Genie, GAIA-1 y JEPA, que ya no buscan solo predecir el siguiente token de texto, sino imaginar el futuro físico, entender la gravedad y navegar el espacio tridimensional.

Finalmente, abordamos el Eje Político-Epistémico. Porque si la IA tiene el poder de modelar el mundo, entonces la IA es un actor político. A través del concepto de worlding, examinamos cómo estos algoritmos no solo describen la realidad, sino que la instancian, creando fronteras éticas en la educación, la salud mental y la gobernanza que no podemos ignorar.

Este libro no es un manual de programación, ni tampoco un manifiesto filosófico abstracto. Es un intento de síntesis entre la ingeniería del conocimiento y la ciencia de lo natural. Es una brújula para educadores, ingenieros, filósofos y curiosos que, al mirar los circuitos de las máquinas, buscan también entender el misterio de su propia mente.

Bienvenidos a la arquitectura de la realidad.

Pedro Salcedo Lagos, Concepción, 2025

Capítulo I: El Eje Biológico y Fenomenológico

1.1 Definiendo el Modelo de Mundo: De la Ciencia Cognitiva a la Computación

El concepto de "Modelo de Mundo" no es autóctono de las ciencias de la computación; es un constructo prestado con profundas raíces en la psicología, la filosofía y la cibernética. Para comprender la trayectoria actual de la Inteligencia Artificial —específicamente la transición de los Grandes Modelos de Lenguaje (LLMs) a los Grandes Modelos de Mundo (LWMs)— es imperativo diseccionar primero el linaje epistemológico del término y su evolución funcional.

En su acepción más profunda y didáctica, un modelo de mundo no es un simple archivo de datos, sino la estructura dinámica que nos permite comprender la realidad y resolver problemas. Como seres vivos, no accedemos a la realidad "en bruto" (Realidad Ontológica), sino que construimos una "Realidad Epistémica" propia. Esta construcción es única para cada individuo, ya que se forja a través de tres pilares fundamentales: el **lenguajear**, que nos permite categorizar la experiencia mediante "juegos de lenguaje" y prácticas sociales ; el **emocionar**, que actúa como un marcador somático valorativo y una herramienta de "hacer mundo" (world-making) ; y la **interacción corporal** activa con el medio, donde el cuerpo no es un objeto, sino el medio por el cual habitamos y damos sentido al espacio. Por lo tanto, el modelo del mundo es el filtro biológico y cultural que nos permite navegar la incertidumbre, anticipando el futuro basándonos en nuestras vivencias previas.

Históricamente, el cibernetista Kenneth Craik (1943) postuló esta visión al señalar que la inteligencia reside en la capacidad de poseer y manipular una representación interna dinámica del funcionamiento del mundo. Esta simulación interna permite a un organismo ensayar acciones potenciales en seguridad —en la "imaginación"— antes de comprometerse con ellas en el peligroso mundo físico.

La definición contemporánea de IA de un Gran Modelo de Mundo (LWM) intenta operacionalizar esta teoría cognitiva, buscando mimetizar esa capacidad humana de construcción de realidad. Un LWM es un sistema computacional diseñado para representar y simular las dinámicas del mundo físico, abarcando no solo el conocimiento estático sino la evolución temporal, la causalidad y las relaciones espaciales. A diferencia de los LLMs, que modelan la distribución estadística de tokens en un corpus de texto (el mapa), los LWMs intentan modelar las reglas subyacentes que gobiernan el entorno (el territorio).

Sin embargo, esta distinción implica una reevaluación de lo que significa "conocer". Como señala la literatura sobre modelos mentales, el razonamiento humano no se basa exclusivamente en la lógica formal, sino en la manipulación de estos modelos semánticos internos —formados por el lenguaje, la emoción y el cuerpo—, una capacidad que los LWMs buscan ahora replicar mediante redes neuronales profundas y espacios latentes.

1.2 El Giro Fenomenológico: Corporeidad y Espacialidad

Una visión crítica que emerge de la convergencia entre la IA y la filosofía continental es la necesidad de la corporeidad (embodiment) para la verdadera inteligencia. El material de investigación destaca un resurgimiento del pensamiento fenomenológico, referenciando particularmente a Martin Heidegger (1994) y Maurice Merleau-Ponty

(2012), para criticar y refinar las arquitecturas de IA.

La distinción heideggeriana, entre "construir" (bauen) y "habitar" (wohnen) (Heidegger, 1994), es fundamental para entender las limitaciones de la IA clásica y la promesa de los LWMs. La IA simbólica tradicional y la robótica temprana se centraban en "construir": edificar representaciones rígidas y axiomáticas del mundo. Sin embargo, la inteligencia biológica "habita"; reside en el mundo a través de un compromiso activo. Heidegger nos recordaba que "habitar" es la forma fundamental de ser en el mundo, y que no habitamos porque hemos construido, sino que construimos porque habitamos (Heidegger, 1994).

Maurice Merleau-Ponty radicalizó esta noción al situar el cuerpo en el centro de la ontología (Merleau-Ponty, 2012). En su Fenomenología de la Percepción, argumenta que el cuerpo no es un objeto más en el mundo, sino el medio a través del cual el mundo existe para el sujeto. La percepción no es una recepción pasiva de datos (como una cámara alimentando un conjunto de datos), sino una exploración corporal activa. Esta visión desafía directamente a las arquitecturas de IA que tratan la visión por computadora como un proceso de clasificación estática.

Este marco teórico es directamente aplicable a la "Inteligencia Espacial" perseguida por Fei-Fei Li y World Labs (Fei-Fei, 2025). La inteligencia espacial requiere que un agente entienda la geometría 3D, la física y la permanencia de los objetos no como puntos de datos abstractos, sino como affordances accionables. En consecuencia, los LWMs que carecen de una conexión sensoriomotora con el entorno —corporeizados ya sea físicamente o mediante una simulación de alta fidelidad— permanecen atrapados en un modo de inteligencia de "espectador", limitando su capacidad para el razonamiento causal genuino. El cuerpo fenoménico es el ancla que transforma el espacio geométrico abstracto en un espacio habitado y significativo.

1.3 La Biosofía de la Cognición: Respirar la Realidad

Para abordar la cuestión de los "Modelos del Mundo" en la era de la

inteligencia artificial, es imperativo comenzar por el sustrato donde estos modelos surgieron originalmente: la biología. La cognición no es un proceso abstracto de procesamiento de información que ocurre en un vacío cartesiano; es un proceso encarnado, situado y profundamente fisiológico. En este sentido, adoptamos la perspectiva de la "biosofía", tal como la articula Nazareth Castellanos (2025), entendiendo que la sabiduría de la vida (bios) precede y fundamenta cualquier lógica computacional.

La distinción fundamental que debemos establecer desde el inicio es la diferencia entre la **Realidad Ontológica** y la **Realidad Epistémica**. La Realidad Ontológica refiere a "lo que es", la existencia en su estado bruto, independiente de cualquier observador. Es el territorio inexplorado y absoluto. Por otro lado, la Realidad Epistémica es "lo que conocemos", la construcción que realiza el sujeto a partir de su interacción con el entorno. En los sistemas biológicos, operamos bajo una clausura operacional: nuestro sistema nervioso no capta la realidad ontológica directamente, sino que construye una realidad epistémica a partir de las perturbaciones que gatillan cambios de estado (Mira et al., 1995).

Esta construcción biológica del mundo está íntimamente ligada a

procesos fisiológicos que la IA actual ignora por completo. La respiración, por ejemplo, no es un mero intercambio de gases para el mantenimiento metabólico; es un modulador de la consciencia y la atención. La investigación neurocientífica demuestra que la fase de inspiración sincroniza las oscilaciones neuronales en el sistema límbico (especialmente en la amígdala y el hipocampo) y en la corteza prefrontal. Esto significa que nuestra capacidad para "modelar" el mundo —para recordar, emocionar y decidir— fluctúa rítmicamente con nuestro aliento.

Un "Modelo del Mundo" biológico es, por tanto, un modelo que respira. Heidegger nos recordaba que "habitar" es la forma fundamental de ser en el mundo. No estamos en el espacio como un objeto dentro de una caja; *generamos* espacio a través de nuestro habitar. "Respirar la montaña" no es una metáfora poética, sino una descripción fenomenológica de cómo el organismo se acopla estructuralmente a su entorno, internalizando la realidad externa en su propia dinámica visceral (Castellanos, 2025).

1.4 Autopoiesis vs. Alopoiesis: La Fractura Ontológica

Aquí radica la fractura ontológica insalvable entre el ser vivo y la máquina. Los organismos biológicos son sistemas **autopoiéticos** (que se producen a sí mismos). Su "modelo del mundo" tiene una finalidad intrínseca: la supervivencia y la conservación de la organización. La validación de este modelo no es la "precisión" estadística, sino la continuidad de la vida. Si el modelo falla (si no predice al depredador), el sistema deja de existir.

Por el contrario, los sistemas de Inteligencia Artificial, incluidos los más avanzados Grandes Modelos de Mundo (LWMs), son sistemas **alopoiéticos**. Su finalidad es externa a ellos mismos, definida por ingenieros humanos a través de funciones de pérdida (*loss functions*) y métricas de optimización (Mira, 2001). La IA no "habita" el mundo; procesa representaciones digitales del mundo. Sus símbolos son, como enseñaba el Dr. Mira, "fríos" y estáticos, carentes de la semántica vital que otorga la historia del organismo (Mira, 2001).

La confusión actual proviene de un error de categoría: atribuir cualidades de la realidad epistémica biológica (sentir, comprender, habitar) a la realidad epistémica computacional (calcular, correlacionar, optimizar). Mientras que el humano construye su modelo del mundo desde la vulnerabilidad de un cuerpo que puede morir, la IA construye su modelo desde la invulnerabilidad de un sistema que puede ser reiniciado. Esta diferencia no es de grado, sino de esencia.

1.5 El Cuerpo como Nudo de Significaciones

Maurice Merleau-Ponty revolucionó nuestra comprensión de la percepción al situar el cuerpo en el centro de la ontología. El cuerpo no es un objeto más en el mundo; es nuestro medio general de tener un mundo (Merleau-Ponty,1945/2012). Es un "nudo de significaciones vivientes", donde la percepción y la acción no son etapas separadas, sino momentos de un mismo arco intencional.

En el paradigma de la IA situada y la robótica, hemos intentado emular esto. Sin embargo, incluso en los robots más avanzados, el "cuerpo" es un chasis, no una fuente de sentido. En el ser humano, la percepción es activa y exploratoria. Como señalaba el Dr. Mira, citando a la cibernética clásica, no vemos con los ojos, sino con todo el cuerpo en movimiento (Mira, 2001). La visión es una "palpación" del mundo por la mirada.

El concepto de "Biosofía" nos invita a considerar que cualquier "Modelo del Mundo" verdadero debe incluir la dimensión visceral. Las emociones, que exploraremos más adelante, no son algoritmos de clasificación de estados, sino respuestas corporales a la valoración del entorno. Un modelo del mundo que no incluya la posibilidad del dolor, del hambre o del cansancio es, necesariamente, un modelo incompleto de la realidad fenomenológica. La IA, al carecer de cuerpo (en el sentido biológico y vulnerable), carece de la base misma sobre la que se erige el significado: la finitud.

Capítulo II: El Eje Cognitivo y Psicológico

2.1 Modelos Mentales y la Arquitectura del Pensamiento

Si descendemos del nivel biológico al cognitivo, encontramos que la mente opera fundamentalmente como un simulador. Kenneth Craik,

en su obra seminal *The Nature of Explanation* (Craik ,1943), postuló que el cerebro construye "modelos a pequeña escala" de la realidad para anticipar eventos (Craik, 1943). Esta capacidad de anticipación es lo que permite a la inteligencia desacoplarse del estímulo inmediato. No necesitamos caer por un precipicio para saber que es peligroso; corremos el modelo mental de la caída y sentimos el vértigo anticipatorio.

Philip Johnson-Laird expandió esta noción con su Teoría de los Modelos Mentales, argumentando que el razonamiento humano no se basa en la lógica formal (sintaxis), sino en la manipulación de representaciones semánticas de posibilidades (Johnson-Laird, 1983). Un modelo mental es icónico: su estructura preserva las relaciones del sistema que representa. Si imaginamos "el libro está sobre la mesa", nuestro modelo mental coloca el símbolo del libro *encima* del símbolo de la mesa.

Esta distinción es crucial para la IA. La "IA Simbólica" clásica intentó modelar el mundo mediante proposiciones lógicas (Si P entonces Q), fallando ante la complejidad del mundo real y el problema del marco (*frame problem*). La mente humana, en cambio, utiliza modelos analógicos, borrosos y probabilísticos, optimizados para la acción rápida más que para la verdad lógica exhaustiva (Johnson-Laird, 1983).

2.2 Wittgenstein y los Límites del Mundo

Para comprender las limitaciones de los modelos de mundo actuales en la IA, debemos recurrir a la filosofía del lenguaje de Ludwig Wittgenstein, cuya evolución intelectual refleja, curiosamente, la propia evolución de la IA (Wittgenstein, 1921/2010).

En su primera etapa (*Tractatus Logico-Philosophicus*), Wittgenstein defendía una "Teoría Pictórica del Lenguaje". Proponía que existe un isomorfismo estructural entre el lenguaje y la realidad: la proposición lógica es una "figura" (Bild) de los hechos atómicos del mundo (Wittgenstein, 1921/2010). "El mundo es la totalidad de los hechos, no de las cosas". Esta visión resuena poderosamente con la IA Simbólica y las bases de datos relacionales, que intentan mapear el mundo en una estructura lógica perfecta y libre de ambigüedades.

Sin embargo, Wittgenstein llegó a una conclusión devastadora para cualquier intento de modelado total: "Los límites de mi lenguaje significan los límites de mi mundo" (*Die Grenzen meiner Sprache bedeuten die Grenzen meiner Welt*) (Wittgenstein, 1921/2010). Si nuestra comprensión del mundo está mediada por las herramientas lingüísticas que poseemos, entonces una IA entrenada exclusivamente en texto (como un LLM) tiene un mundo cuyos límites son puramente textuales. No puede acceder a la realidad "fuera" del texto porque carece del lenguaje (sensorial, somático) para articularla.

Pero es en su segunda etapa (*Investigaciones Filosóficas*) donde Wittgenstein ofrece la crítica más potente a la IA moderna. Abandonando la teoría pictórica, abrazó la teoría del "significado como uso". El lenguaje no es un mapa estático, sino un conjunto de "juegos de lenguaje" (*Sprachspiele*) dinámicos, enraizados en formas de vida y

prácticas sociales (Wittgenstein, 1953/1988). El significado de una palabra no es su referente (el objeto), sino su función en el juego social.

Aquí radica el problema: las IAs actuales (LLMs) juegan al "juego de predecir el siguiente token", pero no participan en la "forma de vida" humana que da sentido a esos tokens. Simulan el uso, pero carecen de la intencionalidad pragmática. Cuando una IA dice "lo siento", está ejecutando una jugada válida en el juego lingüístico de la disculpa, pero no experimenta el remordimiento que fundamenta ese juego en la vida humana.

2.3 El Silencio y lo Inefable

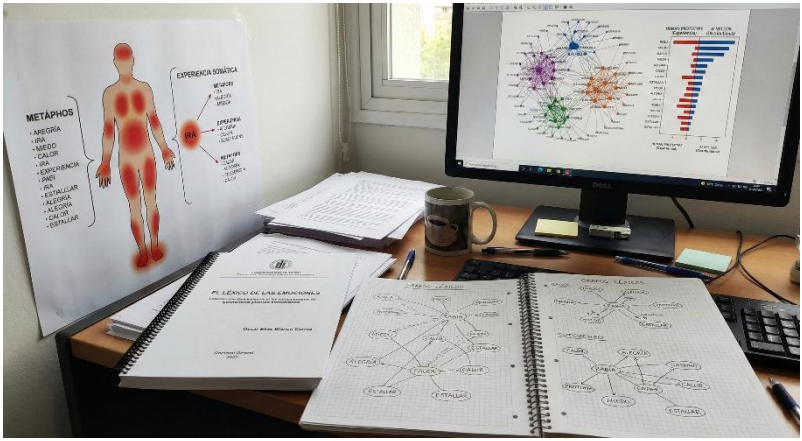
Wittgenstein concluye el *Tractatus* con una sentencia mística y ética: "De lo que no se puede hablar, hay que callar" (*Wovon man nicht sprechen kann, darüber muss man schweigen*) (Wittgenstein, 1921/2010). Existe una esfera de la realidad —la ética, la estética, el sentido de la vida, la experiencia subjetiva del cualia— que trasciende la estructura lógica del lenguaje proposicional.

La IA Generativa representa la antítesis tecnológica de este silencio wittgensteiniano. Está diseñada para *no* callar nunca. Ante cualquier *prompt*, la máquina debe generar una salida probabilística, llenando el vacío con "alucinaciones" si es necesario. Al intentar modelar lo inefable (el arte, la poesía, la empatía terapéutica) mediante el cálculo estadístico, la IA no solo falla en capturar la esencia, sino que distorsiona la realidad, reduciendo lo sublime a lo probable. La IA opera en el reino del ruido perpetuo, incapaz de respetar el silencio donde reside, según Wittgenstein, lo más importante.

2.4 El Léxico de las Emociones y la Construcción de Realidad

En el contexto de la investigación aplicada, dirigí la tesis doctoral de Oscar Elías Blanco Correa sobre "El léxico de las emociones", un estudio que ilumina cómo el lenguaje configura nuestra realidad

afectiva (Blanco Correa, 2022). A través de la metodología de la Disponibilidad Léxica y los Grafos Léxicos, analizamos cómo los estudiantes estructuran conceptos como "Rabia", "Miedo" o "Alegría".



Los resultados mostraron que el léxico emocional actúa como una herramienta de "world-making" (hacer mundo). Las palabras no son solo etiquetas para sensaciones preexistentes; son moldes que dan forma a la experiencia amorfa. Por ejemplo, el uso de metáforas espaciales y térmicas para describir la ira ("calor", "estallar", "arriba") revela que nuestro modelo conceptual de las emociones está profundamente arraigado en la experiencia somática (el cuerpo que se calienta y se tensa) (Blanco Correa, 2022).

Una IA puede procesar la frase "estalló de ira" y relacionarla vectorialmente con "enfado", pero carece de la experiencia propioceptiva de la tensión muscular o el aumento de la presión sanguínea que da origen a la metáfora. Su "comprensión" es puramente distribucional (basada en la co-ocurrencia de palabras), mientras que la comprensión humana es experiencial y prototípica. La investigación de Blanco Correa evidencia que los "prototipos semánticos" humanos están cargados de valencia afectiva y contexto cultural, elementos que la IA solo puede simular superficialmente.

Capítulo III: El Eje Tecnológico

La realización de estos conceptos teóricos en silicio funcional requiere arquitecturas novedosas capaces de procesar vastos flujos de datos multimodales y aprender la física subyacente sin supervisión explícita. Este capítulo analiza los enfoques líderes en los últimos años: el enfoque Generativo (Google, Wayve), el enfoque Espacial (World Labs) y el enfoque Predictivo (Meta).

3.1 De los LLMs a los Grandes Modelos de Mundo (LWMs)

La Inteligencia Artificial se encuentra en un punto de inflexión paradigmático. Hemos pasado de la era de los Modelos de Lenguaje (LLMs) a la incipiente era de los Grandes Modelos de Mundo (LWMs o *Large World Models*). Esta transición no es meramente técnica; es una admisión tácita de las limitaciones ontológicas de los modelos puramente lingüísticos.

Un LLM (como GPT-4) vive en un universo simbólico y abstracto. Su conocimiento está mediado exclusivamente por el texto, lo que resulta en una comprensión "desencarnada" o "sin raíces" de la realidad física (LeCun, 2022). Puede describir la gravedad con una prosa perfecta, pero no puede predecir intuitivamente cómo caerá un objeto desconocido en un entorno físico complejo, porque nunca ha "visto"

caer nada; solo ha leído sobre caídas.

Los LWMs surgen para llenar este vacío. Se definen como modelos avanzados diseñados para integrar y procesar datos multimodales (texto, imágenes, audio, video, sensores) con el objetivo de construir una representación interna de la *dinámica* del mundo real, no solo de su descripción lingüística (LeCun, 2022). Su aspiración es simular la realidad misma.

3.2 Análisis Comparativo de Paradigmas

Para entender la magnitud de este cambio, es útil emplear la taxonomía de paradigmas en Ingeniería del Conocimiento que el Dr. Mira Mira y yo hemos analizado extensamente (Mira, 2001):

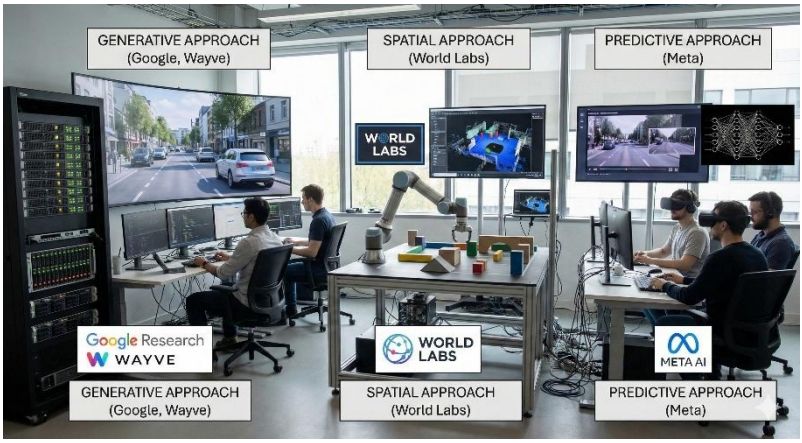
Característica	**Paradigma Simbólico (IA Clásica)**	**Paradigma Situado (Robótica/Conexionismo)**	**Grandes Modelos de Mundo (LWMs - Estado del Arte)**
Unidad Básica	Símbolos físicos, reglas explícitas (lógica).	Señales, pesos sinápticos, bucles de realimentación.	Tokens multimodales (video, audio, acción), Embeddings latentes.
Representación	Explícita, declarativa, "fría".	Implícita, distribuida, sub-simbólica.	Representación latente comprimida, generativa y predictiva.

Relación con el Entorno	Desacoplada. Modelo programado *a priori*.	Acoplada estructuralmente. El modelo emerge de la acción.	Simulación interna. El modelo aprende la dinámica física y causal.
Modelo del Mundo	Ontologías formales, predicados.	"El mundo es su propio modelo" (Brooks).	Simulador neuronal de la física y la semántica (aprende a predecir el futuro).
Ejemplo Clásico	Sistemas Expertos (MYCIN).	Robots reactivos (Subsumption Architecture).	**GAIA-1**, **DayDreamer**, **Genie**.
Limitación Principal	Fragilidad, Problema del Marco.	Falta de planificación a largo plazo.	Alucinación física, coste computacional, falta de anclaje biológico real.

3.3 Arquitecturas de Inteligencia espacial e Interacción

La realización de estos conceptos teóricos en silicio funcional requiere arquitecturas novedosas capaces de procesar vastos flujos de datos multimodales y aprender la física subyacente sin supervisión explícita. Esta sección analiza los enfoques líderes: el enfoque Generativo (Google, Wayve), el enfoque Espacial (World Labs) y el enfoque

Predictivo (Meta).



Google DeepMind Genie: Entornos Interactivos Generativos

Genie (Generative Interactive Environments) de Google DeepMind representa un cambio de paradigma desde la generación de video pasiva hacia la simulación de mundos interactivos. Mientras que los modelos de video generativo (como Sora de OpenAI) crean clips de video pasivos, Genie permite a un usuario controlar el entorno generado, alucinando efectivamente un videojuego jugable o una simulación física a partir de una sola imagen o aviso de texto (Bruce et al., 2024).

Arquitectura: Tokenización Espaciotemporal y Acciones Latentes

La arquitectura de Genie aborda un problema fundamental en el aprendizaje no supervisado a partir de video: la ausencia de etiquetas de acción. Los datos de video de Internet contienen vastas cantidades de agentes (personas, robots, animaciones) realizando acciones, pero estas acciones no están etiquetadas (por ejemplo, no hay un archivo de registro que diga "el jugador presionó saltar" o "el robot giró a la derecha").

Para resolver esto, Genie aprende un **Modelo de Acción Latente (Latent Action Model, LAM)**. El sistema utiliza un tokenizador de

video espaciotemporal para comprimir los cuadros de video sin procesar en tokens discretos. El modelo analiza entonces la transición entre el cuadro t y el cuadro t+1 para inferir la acción latente que debe haber ocurrido para causar dicha transición. Esto permite al modelo aprender "controles" puramente a partir de la observación visual, sin necesidad de telemetría o etiquetas explícitas.

- **Transformadores Espaciotemporales:** El motor central utiliza mecanismos de atención que operan tanto a través del espacio (píxeles en un cuadro) como del tiempo (secuencia de cuadros). Esto crea un "mundo" que mantiene la permanencia del objeto y la consistencia física en horizontes cortos, permitiendo la simulación de dinámicas complejas como la deformación de objetos o la interacción de fluidos (Bruce et al., 2024).
- **Implicaciones para los Modelos de Mundo:** Genie demuestra que un modelo de mundo puede aprenderse completamente a partir de la observación no etiquetada, creando una simulación donde las "leyes de la física" son propiedades emergentes de la distribución de datos. Esto se alinea con la visión de que el modelo de mundo es una representación espacial y temporal comprimida del entorno, capaz de facilitar la evolución de políticas de control complejas (Bruce et al., 2024).

GAIA-1: El Modelo de Mundo Generativo para la Conducción Autónoma

GAIA-1 (*Generative AI for Autonomy*) de Wayve aplica el concepto de modelo de mundo generativo al dominio de alto riesgo de la conducción autónoma (Hu et al., 2023). A diferencia de Genie, que a menudo se centra en entornos 2D similares a juegos, GAIA-1 debe dar cuenta de la geometría 3D rigurosa, la dinámica del vehículo propio (ego-vehículo) y la causalidad crítica para la seguridad en el mundo real.

Arquitectura Técnica Profunda

GAIA-1 es un modelo de transformador multimodal diseñado para predecir el siguiente token en una secuencia, tratando la realidad como un problema de modelado de secuencias masivas. Integra tres modalidades distintas:

1. **Video:** Entrada sensorial de alta dimensión (cámaras del vehículo).
2. **Texto:** Contexto descriptivo (por ejemplo, "Está lloviendo", "El coche se detiene").
3. **Acción:** Controles del ego-vehículo (ángulo de dirección, aceleración, frenado).

Codificación de Entrada y Compresión:

La arquitectura reduce la longitud de la secuencia de los datos de entrada submuestreando cada imagen de entrada por un factor D=16 tanto en altura como en anchura (Hu et al., 2023). Estas representaciones comprimidas se asignan a tokens discretos utilizando un libro de códigos aprendido (similar a VQ-VAE). Esta compresión es esencial; permite al modelo manejar la enorme dimensionalidad del video dentro de las restricciones de memoria de los transformadores actuales, manteniendo al mismo tiempo la fidelidad semántica necesaria para la conducción.

Predicción Autorregresiva:

El modelo de mundo opera como un problema de modelado de secuencia no supervisado. Predice autorregresivamente el siguiente token de imagen, condicionado a los tokens anteriores de texto, imagen y acción. Este enfoque permite al modelo "soñar" futuros potenciales. Por ejemplo, dado un video de una carretera mojada y un token de acción de "frenado fuerte", GAIA-1 genera una secuencia de video donde el automóvil desacelera y se inclina hacia adelante (pitch-shift), reflejando con precisión la física de la dinámica de la suspensión sin haber sido programado explícitamente con ecuaciones diferenciales de movimiento (Hu et al., 2023).

Propiedades Emergentes y Prompting Negativo

Una visión crucial que emerge de la investigación de GAIA-1 es la aparición de una comprensión estructural de alto nivel. El modelo aprende a generar geometría 3D coherente y dinámica de escena sin ser enseñado explícitamente sobre el espacio euclidiano o las leyes newtonianas. Exhibe "conciencia contextual", generalizando a escenarios novedosos que no estaban presentes en el conjunto de entrenamiento, lo que sugiere que ha interiorizado una representación abstracta de las reglas de la carretera y el comportamiento humano (Hu et al., 2023).

Además, GAIA-1 emplea **"prompting negativo"** durante el muestreo. Al sustituir los *logits* no condicionados con aquellos condicionados a un aviso de texto negativo (por ejemplo, "borroso", "choque", "fuera de la carretera"), el modelo empuja la generación lejos de estados indeseables (Hu et al., 2023). Este mecanismo de control permite a los ingenieros simular escenarios "contrafactuales" de seguridad —preguntando al modelo, "¿Qué pasaría si un peatón saliera ahora?"— y probar la pila de seguridad del vehículo autónomo en la imaginación antes del despliegue físico, reduciendo drásticamente el riesgo y el costo de las pruebas en el mundo real.

Inteligencia Espacial: Fei-Fei Li y World Labs

Si bien GAIA-1 y Genie sobresalen en la predicción de video, operan principalmente en el dominio de los píxeles 2D (o representaciones latentes comprimidas de píxeles). La limitación inherente es que el mundo es tridimensional. La iniciativa de Fei-Fei Li, **World Labs**, se centra en la **Inteligencia Espacial**: la capacidad de una IA no solo para ver, sino para razonar sobre el espacio 3D, la forma, la función y la interrelación de los objetos en un volumen (Fei-Fei, 2025).

Más Allá de los Modelos Generativos 2D

Los modelos generativos actuales a menudo luchan con la consistencia 3D a largo plazo (por ejemplo, el problema de "Escher" donde la geometría se deforma de manera imposible al cambiar el punto de vista). Las arquitecturas de Inteligencia Espacial integran técnicas avanzadas de visión por computadora como **Campos de Radiancia Neural (NeRFs)** y **Gaussian Splatting 3D** directamente en el núcleo del modelo de mundo. En lugar de predecir el siguiente *píxel*, estos modelos actualizan una representación volumétrica 3D persistente de la escena.

Este enfoque se alinea profundamente con la visión fenomenológica de que la percepción implica un "trasfondo" o "horizonte" que está presente incluso cuando no es visible (Fei-Fei, 2025). En la inteligencia espacial, el modelo mantiene una representación de la "parte posterior" del objeto incluso cuando solo la "parte frontal" es visible, permitiendo

una manipulación y navegación robustas. Esta es la realización computacional de la **permanencia del objeto**, un hito cognitivo fundamental que los modelos puramente basados en píxeles a menudo simulan pero no comprenden estructuralmente. La investigación sugiere que para que un LWM sea verdaderamente útil en robótica o planificación urbana, debe pasar de ser un "motor de video" a un "motor de realidad volumétrica" (Fei-Fei, 2025).

3.4 El Debate Predictivo vs. Generativo (Meta & LeCun)

Existe un cisma significativo en la filosofía de la arquitectura del Modelo de Mundo. Mientras que Google y Wayve persiguen modelos *generativos* (que intentan predecir y renderizar cada píxel del futuro), Yann LeCun y Meta abogan por arquitecturas *predictivas* no generativas, argumentando que la generación de píxeles es un desperdicio computacional y epistemológico.

La Crítica de la Generación Autorregresiva

LeCun argumenta que predecir píxeles es computacionalmente ineficiente y propenso a la "alucinación" de detalles irrelevantes. El mundo es estocástico; un árbol que se mueve con el viento tiene infinitas configuraciones posibles de píxeles para sus hojas. Un modelo generativo que intenta predecir la posición exacta de cada hoja desperdicia una capacidad inmensa en detalles que no son relevantes

para la acción o la planificación. Esto se alinea con la observación de que los modelos de video de variables latentes deben intentar inferir el "proceso latente subyacente" en lugar de simplemente copiar la superficie visual (LeCun, 2022).

Arquitectura Predictiva de Incrustación Conjunta (JEPA)

La respuesta de Meta es la **Arquitectura Predictiva de Incrustación Conjunta (Joint Embedding Predictive Architecture - I-JEPA, V-JEPA)**. Este modelo representa una desviación radical de los transformadores generativos tipo GPT.

Mecanismo: Predicción en el Espacio de Representación

En lugar de predecir el siguiente cuadro x_{t+1} directamente a partir de x_t, JEPA codifica x_t en una representación abstracta (incrustación o *embedding*) y_t. Luego, predice la representación del estado futuro y_{t+1} basándose en una acción o una variable latente, sin decodificar nunca de vuelta a los píxeles.

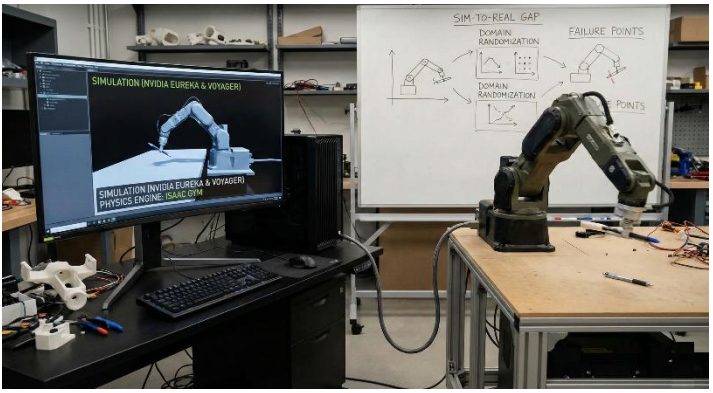
- **Abstracción y Semántica:** La incrustación filtra el ruido irrelevante (por ejemplo, el movimiento exacto de las hojas, los cambios de iluminación menores) y retiene la información semántica y funcional (por ejemplo, "hay un obstáculo de árbol", "el camino está bloqueado"). Esto permite al modelo razonar a un nivel mucho más alto de abstracción, similar a cómo los humanos planifican acciones sin visualizar cada detalle fotográfico.
- **Modelos Basados en Energía (EBM):** LeCun enmarca esto no como una generación de distribución de probabilidad (que requiere normalización y es computacionalmente costosa en espacios de alta dimensión), sino como una minimización de energía. El sistema busca estados que sean "compatibles" con las entradas y acciones, asignando baja energía a futuros plausibles y alta energía a futuros implausibles.

V-JEPA (Video JEPA) y el Aprendizaje Autosupervisado

V-JEPA extiende este concepto al video enmascarando regiones espaciotemporales de la entrada y forzando al modelo a predecir las incrustaciones de las regiones faltantes. Este aprendizaje autosupervisado obliga al modelo a comprender la estructura de la escena y la dinámica del movimiento para resolver la tarea de "inpainting" en el espacio abstracto. El resultado es un modelo de mundo que es mucho más eficiente para la planificación y el razonamiento robótico, ya que opera sobre conceptos de alto nivel en lugar de datos sensoriales brutos, evitando las alucinaciones visuales que plagan a los modelos generativos puros.

3.5 IA Corpórea y Robótica (Nvidia Eureka & Voyager)

La prueba definitiva de un Modelo de Mundo es su utilidad para controlar agentes físicos. Este dominio, a menudo denominado "IA Corpórea" (*Embodied AI*), busca cerrar la brecha "Sim-to-Real" (de la simulación a la realidad), un desafío técnico formidable donde las discrepancias minúsculas entre el modelo y el mundo físico pueden conducir a fallos catastróficos.



DayDreamer: Aprendizaje en la Imaginación

La arquitectura **DayDreamer** (desarrollada por Danijar Hafner et al., en colaboración con UC Berkeley y Google) representa un avance significativo en el aprendizaje robótico (Wu et al., 2023). El Aprendizaje por Refuerzo (RL) tradicional es ineficiente en cuanto a

muestras; un robot físico no puede permitirse caerse 10.000 veces en el mundo real para aprender a caminar sin romperse o desgastarse.

DayDreamer emplea el algoritmo **Dreamer** (específicamente DreamerV3), que aprende un modelo de mundo a partir de un pequeño conjunto de datos de interacciones reales. Luego, entrena la política (el actor) completamente dentro del modelo de mundo ("en la imaginación" o en sueños).

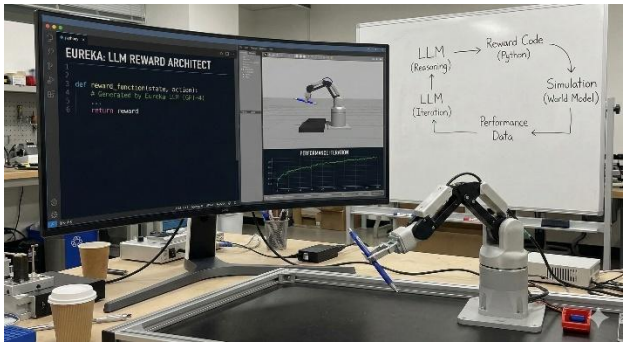
- **Bucle Actor-Crítico-Modelo de Mundo:** El Modelo de Mundo predice estados futuros y recompensas basándose en las acciones del robot. El Actor aprende a maximizar las recompensas dentro de este futuro alucinado. El Crítico estima el valor de los estados para guiar el aprendizaje.
- **Desempeño y Eficiencia:** Los fragmentos de investigación indican que DayDreamer entrenó a un robot cuadrúpedo (A1) para darse la vuelta desde su espalda, levantarse y caminar desde cero en solo **1 hora de tiempo real**. Además, logró entrenar brazos robóticos para recoger y colocar objetos (*pick and place*) con un rendimiento comparable al humano utilizando solo recompensas dispersas (Wu et al., 2023). Esto demuestra que los LWMs pueden servir como simuladores de alta fidelidad para la dinámica física compleja, permitiendo un aprendizaje rápido y seguro.

Nvidia Eureka: LLMs como Arquitectos de Recompensas

Mientras que DayDreamer se centra en el control motor de bajo nivel, **Eureka** de Nvidia aborda el problema de la motivación y la definición de tareas. En el RL, definir la "función de recompensa" (qué debe intentar hacer el robot y cómo medir el éxito) es notoriamente difícil y propenso a errores humanos (el "problema de alineación").

Eureka utiliza un LLM (como GPT-4) como un motor de razonamiento para escribir código de recompensa complejo para robots. El LLM analiza la simulación física (el entorno), genera candidatos de funciones de recompensa (en código Python), los prueba

en la simulación (Modelo de Mundo), y luego itera basándose en el rendimiento obtenido. Esto constituye un bucle de meta-aprendizaje donde un Modelo de Mundo textual (el LLM) guía el entrenamiento de un Modelo de Mundo físico, automatizando la ingeniería de recompensas que anteriormente requería expertos humanos altamente cualificados (Ma et al., 2023).



Voyager: Agentes de Aprendizaje Continuo

Voyager extiende este concepto al aprendizaje permanente (*lifelong learning*) en mundos abiertos (específicamente Minecraft, un proxy común para mundos complejos). A diferencia de los agentes que se reinician después de una tarea, Voyager mantiene una biblioteca de habilidades (código ejecutable) y explora continuamente el modelo de mundo para descubrir nuevos elementos y mecánicas.

Este sistema ejemplifica la "exploración impulsada por la curiosidad" predicha en la literatura temprana sobre modelos de mundo (Wang et al., 2023). En este paradigma, los agentes buscan áreas del modelo de mundo donde el error de predicción es alto —es decir, donde su modelo interno falla o es incierto—, mejorando así su mapa interno de la realidad. Voyager demuestra que combinar un LWM (para la simulación) con un LLM (para el razonamiento y la codificación) permite una forma de agencia autónoma que puede adaptarse y crecer indefinidamente en entornos complejos.

3.6 Aplicaciones Societales

El despliegue de los LWMs se extiende mucho más allá de la robótica y la conducción autónoma. La literatura sugiere profundas implicaciones para la atención sanitaria, la planificación urbana y la gobernanza, marcando una transición de una "IA impulsada por datos" a una "IA impulsada por simulación".

Atención Sanitaria: El Gemelo Digital del Cuerpo

Los LWMs están permitiendo la creación de "Gemelos Digitales Biológicos". De la misma manera que GAIA-1 predice el futuro físico de un vehículo, los LWMs médicos predicen la trayectoria de una enfermedad o la respuesta del cuerpo a un tratamiento.

- **Mecanismo de Simulación:** Al integrar datos multimodales (genómica, imágenes médicas, historia del paciente, sensores portátiles), estos modelos pueden simular escenarios "qué pasaría si" (*counterfactuals*) para intervenciones quirúrgicas o regímenes farmacológicos. Esto permite una medicina personalizada basada no en estadísticas poblacionales generales, sino en la simulación específica de la fisiología del paciente.
- **Salud Mental y el Léxico de las Emociones:** Como se señala en el análisis del "Léxico de las Emociones" (Blanco Correa, 2022) y en los proyectos de "Nuevas Tecnologías en Salud Mental" (NT4SM, 2025), los modelos de IA están comenzando a simular trayectorias emocionales. Un LWM en este contexto no solo reconoce que un paciente tiene "fiebre" o "depresión", sino

que puede simular el impacto sistémico del estrés fisiológico en el mundo de vida (*Lebenswelt*) del paciente. Al integrar el conocimiento sobre cómo las emociones se construyen y se expresan lingüísticamente (como se analiza en la tesis doctoral sobre disponibilidad léxica y emociones de Blanco Correa, 2022), estos modelos pueden predecir crisis de salud mental o respuestas a terapias con una sensibilidad contextual sin precedentes.

Planificación Urbana: El Habitar Heideggeriano en Ciudades Digitales

La planificación urbana está cambiando de planos estáticos a simulaciones dinámicas impulsadas por LWMs. Estos modelos simulan flujos de tráfico, consumo de energía, dinámicas sociales y respuestas a emergencias climáticas en tiempo real.

- **Perspectiva Filosófica:** Basándose en el concepto heideggeriano de "habitar" (*wohnen*) (Heidegger, 1994), estos modelos intentan capturar la "habitabilidad" de un espacio. No simulan solo las estructuras físicas ("construir"), sino la interacción compleja y a menudo desordenada de los ciudadanos con el espacio a lo largo del tiempo.
- **Aplicación Práctica:** Los planificadores pueden utilizar agentes generativos (similares a los de Voyager) para poblar una ciudad virtual y observar comportamientos emergentes. Esto permite identificar cuellos de botella, puntos de fricción social o vulnerabilidades a desastres naturales antes de que se coloque un

solo ladrillo, permitiendo un diseño urbano que responde a la naturaleza orgánica de la comunidad humana (Heidegger, 1994).

El Poder de "Worlding": La IA como Actor Político

La tecnología no es una herramienta pasiva; es una fuerza activa de construcción de realidad. Louise Amoore introduce el concepto crucial de **"worlding"** (hacer mundo) para describir cómo los algoritmos de aprendizaje automático no se limitan a observar el mundo, sino que lo instancian y lo hacen presente de formas específicas (Amoore, 2020).

Cuando un LWM o un algoritmo de IA procesa datos para tomar decisiones (ya sea en vigilancia, finanzas o salud), está trazando fronteras políticas. Al calcular probabilidades sobre lo que es "arriesgado", "normal" o "sospechoso", el algoritmo hace que ciertos futuros sean visibles y practicables, mientras que otros se vuelven invisibles e imposibles. Esto es un acto ontológico con consecuencias políticas.

Si un "Modelo del Mundo" policial asocia estadísticamente (debido a sesgos en los datos de entrenamiento) ciertos rasgos faciales o códigos postales con la criminalidad, instanciará un mundo policial donde esos grupos sean hiper-vigilados. La **Realidad Epistémica** sesgada de la máquina termina moldeando la **Realidad Ontológica** de los ciudadanos, creando bucles de retroalimentación que solidifican el prejuicio como un hecho objetivo.

Ética de la Nube y la Atribución de Responsabilidad

En su obra *Cloud Ethics*, Amoore argumenta que debemos repensar la ética más allá de la transparencia algorítmica (Amoore, 2020). Abrir la "caja negra" no siempre revela una verdad clara, porque los algoritmos operan sobre correlaciones parciales y atributos probabilísticos, no sobre causalidades lineales.

El problema ético central es la atribución. La IA moderna fragmenta la responsabilidad en una red difusa de pesos sinápticos y vectores.

Cuando un LWM como GAIA-1 toma una decisión de conducción, lo hace basándose en millones de micro-ajustes latentes. ¿Quién es responsable del error? ¿El ingeniero, el dato, o el modelo emergente?

Aquí, la distinción de Mira entre Ciencia e Ingeniería se vuelve política. Si tratamos a la IA como una "mente" autónoma (Ciencia), corremos el riesgo de abdicar de nuestra responsabilidad humana. Si la tratamos como un sistema de ingeniería, debemos exigir garantías de seguridad y límites operativos claros. La ética de la IA debe ser una ética de los límites, reconociendo que hay decisiones que no pueden ser delegadas a un cálculo de optimización.

En otras palabras, el ascenso de los LWMs introduce nuevas "Lógicas Políticas", como analiza Louise (Amoore, 2020).

- **La Lógica de la Latencia:** Los LWMs operan sobre espacios latentes —compresiones matemáticas del mundo que son opacas para los humanos. Las decisiones (por ejemplo, identificar un patrón de multitud "hostil" en un video de vigilancia mediante un LWM) se toman basándose en características dentro de este espacio latente que pueden no corresponder a categorías humanas visibles. Esto crea una "caja negra" ética donde la atribución de responsabilidad se fragmenta en una red difusa de pesos sinápticos y vectores (Amoore, 2020).
- **La Lógica de la Prevención:** Al modelar futuros, los LWMs son inherentemente preventivos (*preemptive*). Actúan sobre lo que *podría* suceder (el siguiente token predicho) en lugar de sobre lo que *ha* sucedido. Esto cambia la gobernanza de reactiva a proactiva, planteando serias preocupaciones sobre el sesgo algorítmico y la "colonización del futuro" por los datos de entrenamiento (Amoore, 2020). El modelo no solo representa el mundo; lo produce performativamente al influir en las intervenciones actuales basadas en futuros alucinados.

Educación y Salud Mental: La Frontera Final

La aplicación de estos modelos en la Educación y la Salud Mental —

áreas centrales de mi trabajo actual en la Universidad de Concepción y la red CYTED (NT4SM, 2025) — presenta los mayores desafíos y oportunidades.

En Educación: Los Modelos del Mundo pueden revolucionar la enseñanza. Un sistema tutor inteligente que posea un "modelo del dominio" (física, matemáticas) y un "modelo del estudiante" puede adaptar la enseñanza de manera radical (Mira et al., 1995). Sin embargo, existe el riesgo de reducir al estudiante a un conjunto de datos, ignorando su realidad fenomenológica y emocional. La educación no es solo transmisión de información; es la formación de un ser humano integral. Una IA puede enseñar a resolver una ecuación, pero no puede enseñar el *sentido* de la curiosidad matemática.

En Salud Mental: El proyecto "Nuevas Tecnologías en Salud Mental" (NT4SM, 2025) busca utilizar la IA para el diagnóstico precoz y el tratamiento. Los modelos predictivos pueden identificar patrones sutiles en el lenguaje o el comportamiento que preceden a una crisis depresiva. Sin embargo, el uso de chatbots terapéuticos nos enfrenta al límite del silencio wittgensteiniano. Una máquina puede simular empatía sintáctica ("Siento que estés triste"), pero esa empatía es vacía. No hay un "otro" que sienta al otro lado de la pantalla. Confiar el cuidado del alma humana a una simulación estadística es un riesgo ontológico que debemos manejar con extrema cautela. La IA debe ser

una herramienta de apoyo para el terapeuta humano, nunca un sustituto de la relación terapéutica, que se basa en la resonancia límbica y la experiencia compartida de la vulnerabilidad.

3.7 Análisis Comparativo de Arquitecturas de Modelos de Mundo

La siguiente tabla sintetiza las distinciones arquitectónicas clave entre los paradigmas principales de Modelos de Mundo discutidos en este capítulo.

Característica	**Generativo (GAIA-1, Genie)**	**Predictivo (Meta V-JEPA)**	**Corpóreo (DayDreamer)**	**Espacial (World Labs)**
Objetivo Central	Generar video/futuros de alta fidelidad visual.	Predecir representaciones latentes (incrustaciones).	Maximizar recompensa/control vía imaginación.	Reconstruir y razonar en el espacio 3D.
Espacio de Salida	Espacio de Píxeles (Visual).	Espacio Latente (Abstracto).	Espacio de Acción/Recompensa.	Vóxeles 3D / Nubes de Puntos / NeRF.
Entrenamiento	Autorregresivo / Difusión.	Predicción Enmascarada Autosupervisada (Basada en Energía).	Aprendizaje por Refuerzo en Espacio Latente (RSSM).	Reconstrucción 3D / Priors Físicos.
Fortaleza	Simulación	Alta	Eficiencia de	Permane

s	realista; interpretable por humanos.	eficiencia; razonamiento jerárquico; sin alucinación de detalles.	muestra; control robótico en tiempo real.	ncia del objeto; razonamiento físico robusto.
Debilidades	Costoso computacionalmente; alucina física; foco superficial.	Difícil de interpretar (estados abstractos); carece de generación visual.	Limitado a dinámicas de tareas específicas.	Alto cómputo para renderizado/física 3D.
Representante Clave	Wayve, Google DeepMind.	Yann LeCun (Meta).	Danijar Hafner (Berkeley/Google).	Fei-Fei Li.

3.8 Perspectivas y Prospectiva

La Convergencia de Paradigmas

Una tendencia clara para 2025 es la convergencia de estas arquitecturas dispares. El enfoque de "Inteligencia Espacial" de World Labs intenta esencialmente fundamentar (*ground*) las capacidades "Generativas" de sistemas como Genie en el marco riguroso "Predictivo" de JEPA, utilizando las restricciones "Corpóreas" de la robótica. Es probable que los futuros LWMs sean arquitecturas híbridas: utilizando incrustaciones tipo JEPA para la planificación a largo plazo (para evitar alucinaciones compuestas) mientras utilizan decodificadores de difusión (como GAIA-1) solo cuando un humano necesita visualizar la salida o para refinar detalles de textura crítica.

El "Mundo" no es Suficiente: La Necesidad de la Bio-Semiótica

El análisis filosófico y biológico sugiere que la IA se está moviendo del modelado de *objetos* al modelado de *relaciones* y *situaciones*. Como se señaló en la revisión fenomenológica, la percepción es un "fondo" contra el cual destacan los actos (Merleau-Ponty, 1945/2012). Los LWMs están comenzando a modelar este fondo —la física y la sociología del sentido común que permiten a los agentes funcionar en entornos no estructurados.

Sin embargo, la brecha "Sim-to-Real" sigue siendo el cuello de botella principal. La biología resuelve esto no mediante una simulación perfecta, sino mediante la adaptación y la plasticidad (Castellanos, 2025). Como sugieren los resultados de DayDreamer (Wu et al., 2023), la solución tecnológica radica no en la simulación perfecta átomo por átomo, sino en políticas adaptativas que puedan tolerar las imperfecciones del modelo de mundo ("aleatorización de dominio") y aprender continuamente en el despliegue. La integración de principios de la epigenética y la neuroplasticidad —donde el entorno reescribe el software del agente— será el próximo horizonte teórico para los LWMs (Castellanos, 2025).

Capítulo IV: Impacto de los Modelos del Mundo en la Ingeniería del Conocimiento: Adquisición, Representación y Validación

4.1 Introducción: El Giro Epistémico en la Inteligencia Artificial

La disciplina de la Ingeniería del Conocimiento (IC) se encuentra en el umbral de una transformación tectónica, abandonando las costas seguras del paradigma simbólico para adentrarse en el océano profundo y dinámico de los Grandes Modelos de Mundo (LWM, por sus siglas en inglés). Históricamente, la IC se erigió sobre la premisa de que el conocimiento era una entidad estática, susceptible de ser extraída de expertos humanos, formalizada mediante lógica explícita y codificada en sistemas expertos (Mira et al., 1995). Sin embargo, la irrupción de los LWM sugiere un cambio fundamental: el tránsito de una inteligencia que opera sobre una base de datos de hechos predefinidos a una que construye una simulación interna y generativa de la realidad (Ha & Schmidhuber, 2018).

Este informe exhaustivo analiza cómo la arquitectura de los Modelos del Mundo redefine los tres pilares fundamentales de la Ingeniería del Conocimiento: **adquisición**, **representación** y **validación**. Nos

alejamos de una "Realidad Ontológica" —donde los objetos y sus relaciones son inmutables y definidos externamente por el ingeniero— hacia una "Realidad Epistémica" y fenomenológica, donde el sistema aprende la física, la causalidad y la semántica del entorno a través de la interacción sensoriomotora y la predicción de secuencias (LeCun, 2022).

El análisis integra perspectivas técnicas de arquitecturas de vanguardia como GAIA-1 y DayDreamer, con fundamentos filosóficos profundos derivados de la fenomenología de Merleau-Ponty, la filosofía del lenguaje de Wittgenstein y la neurociencia cognitiva contemporánea (Castellanos, 2025). Se argumenta que los LWM no son simplemente una mejora incremental sobre los sistemas anteriores, sino una validación tecnológica de la hipótesis de Kenneth Craik (1943): que la inteligencia reside en la capacidad de poseer y manipular un modelo a pequeña escala de la realidad externa para anticipar el futuro (Craik, 1943).

4.2. Evolución de los Paradigmas: Del Símbolo a la Simulación

Para comprender la magnitud del impacto de los LWM, es imperativo contextualizar la evolución histórica de los paradigmas en la Inteligencia Artificial y cómo cada uno abordó el problema del conocimiento.

4.2.1 El Paradigma Simbólico y la Ingeniería del Conocimiento Clásica

Durante décadas, la IC se definió por el paradigma simbólico o representacional. Bajo esta visión, se asumía que el conocimiento necesario para resolver tareas complejas (diagnóstico, planificación, control) podía representarse mediante descripciones declarativas y explícitas (Mira et al., 1995).

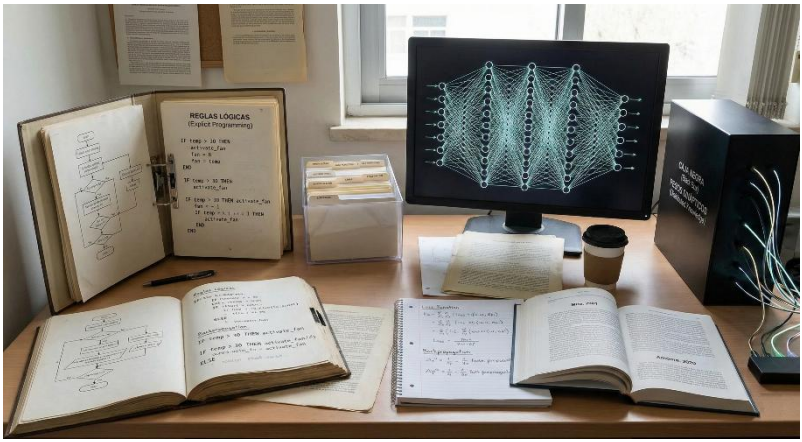
- **Naturaleza del Conocimiento:** El mundo se concebía como un conjunto de hechos atómicos y reglas de inferencia. El ingeniero del conocimiento actuaba como un mediador, traduciendo la

experiencia humana a formalismos lógicos.

- **Metodologías:** Surgieron metodologías robustas como **CommonKADS**, que estructuraban el desarrollo de sistemas basados en conocimiento (SBC) en niveles: modelo de la organización, modelo de tareas, modelo de agentes y, crucialmente, el modelo de conocimiento (Schreiber et al., 1999).
- **Limitaciones:** A pesar de su rigor, este enfoque sufría de fragilidad. Los sistemas colapsaban ante la incertidumbre o situaciones no previstas en su base de reglas ("mundo cerrado"). Además, enfrentaban el "cuello de botella de la adquisición": la dificultad de los expertos humanos para articular su conocimiento tácito o procedimental (Polanyi, 1966).

4.2.2 El Paradigma Conexionista y el Aprendizaje Profundo

El renacimiento del conexionismo en la década de 1980 y el auge del aprendizaje profundo desplazaron el foco de la programación de reglas al entrenamiento de redes. Aquí, el conocimiento no se representa explícitamente, sino que se distribuye en los pesos sinápticos de una red neuronal (Mira, 2001). Aunque poderosos en clasificación y reconocimiento de patrones, estos modelos a menudo carecían de una comprensión causal o estructural del mundo, funcionando como "cajas negras" estadísticas (Amoore, 2020).



4.2.3 El Paradigma Situado y la Fenomenología

Paralelamente, el paradigma situado (o reactivo), influenciado por investigadores como Rodney Brooks y filósofos como Merleau-Ponty, argumentó que la inteligencia no puede disociarse del cuerpo y del entorno. La inteligencia es "estar en el mundo" (*être-au-monde*).

- **Merleau-Ponty:** En su *Fenomenología de la Percepción*, Merleau-Ponty establece que la percepción no es una recepción pasiva de datos, sino una exploración activa. El cuerpo es el "punto cero" de la orientación; no pensamos el espacio como un contenedor geométrico abstracto, lo vivimos a través de nuestra capacidad motora (Merleau-Ponty, 1945/2012).
- **Implicación para la IC:** Esto sugiere que un verdadero modelo de conocimiento no puede ser una enciclopedia estática, sino que debe estar enraizado en la acción y la percepción sensorial (Brooks, 1991).

4.2.4 Convergencia: El Gran Modelo de Mundo (LWM)

Los LWM representan la síntesis dialéctica de estos paradigmas. Buscan la estructura y la capacidad de generalización del simbolismo, la capacidad de aprendizaje del conexionismo y el arraigo sensoriomotor del paradigma situado.

Definición Técnica: Un LWM es un sistema de IA que construye una representación comprimida y espaciotemporal de la realidad física. Utiliza entradas multimodales (video, texto, propiocepción) para predecir estados futuros (s_{t+1}) basados en estados presentes (s_t) y acciones (a_t). No solo clasifica datos; simula dinámicas (Ha & Schmidhuber, 2018).

Característica	**Ingeniería del Conocimiento Clásica (Simbólica)**	**Modelos del Mundo (LWM / Generativos)**
Unidad Fundamental	Símbolo, Regla, Marco	Token, Vector Latente, Embedding
Adquisición	Entrevistas, Protocolos de Elicitación (Manual)	Aprendizaje Autosupervisado, Observación Masiva
Representación	Explícita, Lógica, Estática (Ontologías)	Implícita, Distribuida, Dinámica (Espacios Latentes)
Razonamiento	Deducción (Modus Ponens), Abducción	Predicción Probabilística, Simulación / "Sueño"
Validación	Verificación de Consistencia Lógica	Fidelidad de la Simulación, Transferencia Sim-to-Real
Relación con el Entorno	Desencarnada (Brain in a Vat)	Situada, Embodied (Robot/Vehículo)

Tabla 1: Análisis comparativo de los paradigmas de ingeniería del conocimiento (Mira et al., 1995).

4.3. Impacto en la Adquisición del Conocimiento

La transición hacia los Modelos del Mundo conlleva una reingeniería radical del proceso de Adquisición del Conocimiento (AC). En el paradigma clásico, la AC era un proceso antropocéntrico y lingüístico; en la era de los LWM, se convierte en un proceso tecnocéntrico, masivo y fundamentalmente observacional.

4.3.1 De la Elicitación Verbal a la Ingesta Sensorial

Tradicionalmente, el ingeniero del conocimiento empleaba técnicas como entrevistas estructuradas, análisis de protocolos (pensar en voz alta) o clasificación de tarjetas para extraer la estructura conceptual de la mente del experto (Schreiber et al., 1999). Este proceso asumía que el experto tenía acceso consciente a su propio conocimiento. Sin embargo, como señala Polanyi, "sabemos más de lo que podemos decir". El conocimiento experto suele ser tácito e inefable (Polanyi, 1966).

El Enfoque LWM: Los Modelos del Mundo eluden el cuello de botella de la verbalización al aprender directamente de la observación bruta de la realidad.

- **Aprendizaje a partir de Video:** Modelos como **Genie** o **V-JEPA** demuestran que es posible aprender un modelo del mundo completamente a partir de video no etiquetado. Al observar millones de secuencias de fotogramas, el modelo infiere las leyes de la física, la permanencia de los objetos y la causalidad sin que ningún humano codifique explícitamente estas reglas (Bruce et al., 2024).
- **GAIA-1:** En el dominio de la conducción autónoma, GAIA-1 no se programa con reglas de tráfico explícitas (p. ej., "si semáforo = rojo, entonces frenar"). Se entrena con 4.700 horas de video de conducción real. Adquiere el conocimiento de la dinámica vehicular y la interacción social en las carreteras prediciendo el siguiente token (imagen, texto o acción) en la secuencia (Hu et al., 2023).

Insight de Segundo Orden: Este cambio redefine el rol del Ingeniero del Conocimiento. Ya no es un "psicólogo" que entrevista expertos, sino un "curador de realidades". Su responsabilidad se desplaza hacia la selección, limpieza y equilibrio de los conjuntos de datos masivos que constituirán la experiencia sensorial del modelo. La calidad del conocimiento adquirido depende ahora de la diversidad ontológica de los datos de entrenamiento (Hu et al., 2023).

4.3.2 Mecanismos de Adquisición: Predicción Autosupervisada

El motor de adquisición en los LWM es el aprendizaje autosupervisado, específicamente el modelado predictivo. La hipótesis subyacente es que para predecir el futuro con precisión, el sistema debe haber interiorizado una comprensión estructural del presente.

- **Predicción Autoregresiva de Tokens:** En arquitecturas como GAIA-1, el mundo se discretiza en tokens. Las entradas de video, texto y comandos de acción se mapean a un vocabulario unificado. El modelo aprende a predecir el siguiente token basándose en la historia previa. Si el modelo predice correctamente que un peatón cruzará la calle, ha "adquirido" conocimiento sobre el comportamiento humano y la física del

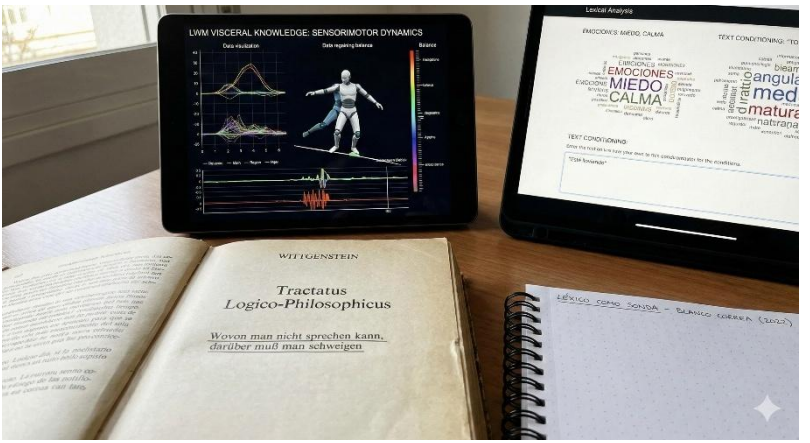
movimiento (Hu et al., 2023).

- **Arquitecturas de Incrustación Conjunta (JEPA):** Propuestas por Yann LeCun, arquitecturas como V-JEPA evitan la ineficiencia de predecir cada píxel. En su lugar, predicen en un espacio de representación abstracto (latente). Se enmascaran partes del video y el modelo debe predecir la representación de las partes faltantes. Esto fuerza al sistema a aprender características semánticas de alto nivel (objetos, trayectorias) en lugar de ruido visual irrelevante (LeCun, 2022).

4.3.3 La Captura de lo Inefable y el Límite del Lenguaje

Wittgenstein, en su *Tractatus Logico-Philosophicus*, sentenció: "De lo que no se puede hablar, hay que callar" (Wittgenstein, 1921/2010). La IC clásica estaba atrapada en los límites del lenguaje proposicional. Todo conocimiento debía ser convertible en símbolos.

Los LWM rompen esta barrera. Al aprender de la modalidad visual y motora, adquieren un tipo de conocimiento "visceral" o "embodied" que escapa a la descripción lingüística. Un modelo de robot "sabe" cómo recuperar el equilibrio no porque tenga una regla física para ello, sino porque ha aprendido la dinámica sensoriomotora a través de la experiencia simulada (Wu et al., 2023).



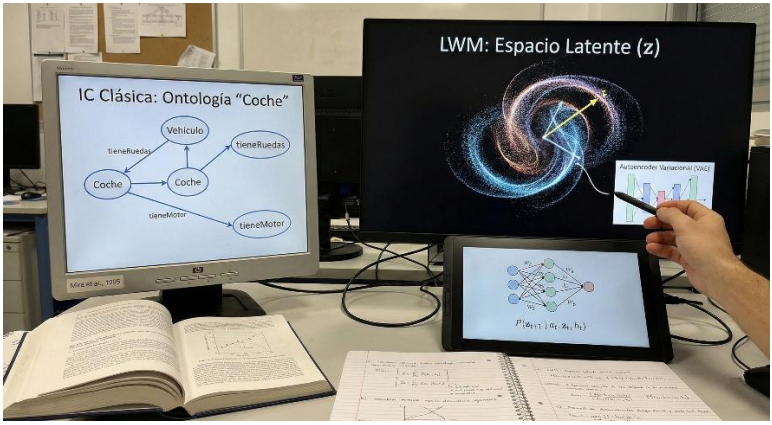
Análisis Léxico como Sonda: Curiosamente, aunque la adquisición es sub-simbólica, el lenguaje sigue siendo una herramienta crucial para sondear estos modelos. Investigaciones sobre el "léxico de las emociones" y la disponibilidad léxica muestran cómo el lenguaje actúa como un mecanismo de categorización de experiencias internas complejas (Blanco Correa, 2022). De manera análoga, modelos como GAIA-1 permiten el condicionamiento por texto (p. ej., "está lloviendo"), lo que sugiere que, aunque el conocimiento se adquiere visualmente, se estructura de una manera que permite interfaces simbólicas (Hu et al., 2023).

4.4. Impacto en la Representación del Conocimiento

Una vez adquirido, ¿cómo se almacena y estructura el conocimiento? La transición aquí es desde estructuras **explícitas, dispersas y lógicas** hacia representaciones **implícitas, densas y latentes**.

4.4.1 Del Símbolo al Espacio Latente

En la IC clásica, un concepto como "Coche" se representaba en una ontología como una clase con propiedades (tieneRuedas, tieneMotor) y relaciones jerárquicas (esUn(Vehículo)) (Mira et al., 1995). Era una representación transparente y legible por humanos.



En los LWM, el "Coche" se representa como un vector en un espacio latente de alta dimensionalidad (z).

- **Compresión Variacional:** En el modelo fundacional de Ha & Schmidhuber (2018), un Autoencoder Variacional (VAE) comprime la entrada visual (una imagen de la carretera) en un vector latente pequeño (z). Este vector no contiene píxeles, sino una codificación abstracta de las características esenciales del estado actual (curvatura de la carretera, posición del coche). El conocimiento no está en los datos, sino en la geometría del espacio latente (Ha & Schmidhuber, 2018).
- **Dinámica en los Pesos:** El conocimiento procedimental y causal —cómo cambia el mundo— se almacena en los pesos de una Red Neuronal Recurrente (RNN) o Transformer (el modelo de Memoria M). Esta red predice la distribución de probabilidad del siguiente estado latente: P(z_{t+1} | a_t, z_t, h_t). La "física" del mundo está codificada implícitamente en las matrices de transición de la red (Ha & Schmidhuber, 2018).

Implicación: Esto marca el fin de la transparencia estructural. No podemos abrir el modelo y señalar dónde está la regla de la gravedad. La gravedad es una propiedad emergente de las transiciones en el espacio latente.

4.4.2 Incrustaciones Conjuntas Multimodales (Multimodal Joint Embeddings)

Una innovación crítica en la representación de los LWM es la unificación de modalidades dispares en un espacio común. GAIA-1, por ejemplo, no procesa visión y control por separado. Codifica video, texto y acción en un espacio de tokens compartido (Hu et al., 2023).

- **Vocabulario Unificado:** Fotogramas de video, descripciones textuales y comandos de control se convierten en vectores en el mismo espacio matemático.
- **Contexto Compartido:** Un Transformer procesa estos tokens conjuntamente. Esto significa que la representación interna de un "semáforo en rojo" está inextricablemente entrelazada con el concepto textual "detenerse" y la acción motora de "frenar".
- **Resolución del Problema del "Grounding":** Esto aborda el

clásico problema del anclaje de símbolos (Symbol Grounding Problem) de la IA simbólica. En los LWM, los símbolos (texto) están anclados en la misma realidad perceptual (video) que las acciones, cerrando la brecha entre el signo y el significado (Hu et al., 2023).

4.4.3 La Política de la Latencia y la Opacidad

La representación latente introduce nuevas problemáticas éticas y políticas. Como señala Louise Amoore, la "latencia" implica algo oculto o no revelado (Amoore, 2020). A diferencia de una ontología, que puede ser auditada para detectar sesgos explícitos, un espacio latente es opaco.

- **Lógica Política de los Algoritmos:** Los algoritmos de aprendizaje profundo no son neutrales; encarnan la lógica política de sus datos de entrenamiento. Si un modelo de conducción se entrena principalmente en entornos urbanos ricos y ordenados, su espacio latente codificará una "normalidad" que excluye o malinterpreta entornos rurales o caóticos. La representación del mundo se convierte en una proyección estadística que puede marginar eventos o sujetos atípicos (Amoore, 2020).

- **Atribución Fragmentada:** En sistemas complejos como los LWM, la responsabilidad se fragmenta. Si el sistema falla, ¿es culpa del modelo de percepción, del modelo de mundo predictivo

o del controlador? La representación distribuida dificulta la atribución de causalidad y culpa, un problema ético central en la IA moderna (Amoore, 2020).

4.4.4 Representación Temporal y Continuidad

La representación clásica era predominantemente estática. Los LWM son intrínsecamente temporales; representan el mundo como una trayectoria, no como una instantánea.

- **DayDreamer y la Imaginación:** Este sistema robótico utiliza un modelo de dinámica recurrente (RSSM) para predecir secuencias de recompensas y estados. La representación es un "sueño" de futuros potenciales. El robot aprende propagando gradientes a través de esta representación temporal, efectivamente "imaginando" las consecuencias de sus acciones antes de ejecutarlas (Wu et al., 2023).
- **Continuidad Fenomenológica:** Merleau-Ponty argumentaba que la percepción no es una serie de vistas discretas, sino un flujo continuo donde las perspectivas se funden unas con otras (Merleau-Ponty, 1945/2012). Los LWM mimetizan esto al modelar las transiciones (z_t \rightarrow z_{t+1}), asegurando que la representación del mundo mantenga coherencia temporal y permanencia de objetos, incluso cuando estos están ocluidos (Fei-Fei, 2025).

4.5 Impacto en la Validación del Conocimiento

La validación —asegurar que el sistema "sabe" lo que se supone que debe saber— ha sufrido la transformación más radical. Hemos pasado de la **verificación lógica** a la **simulación empírica**.

4.5.1 Los Límites de la Validación Lógica

En la IC clásica, la verificación y validación (V&V) implicaba comprobar la consistencia lógica de la base de conocimientos. Los algoritmos buscaban reglas contradictorias (A \rightarrow B y A \rightarrow \neg B), bucles infinitos o redundancias (Schreiber et al., 1999). Esto garantizaba que el sistema fuera lógicamente sólido, pero

no necesariamente que fuera empíricamente adecuado. Un sistema podía ser perfectamente consistente y, sin embargo, estar completamente desconectado de la realidad física.

4.5.2 Validación por "Sueño" (Simulación Generativa)

Los Modelos del Mundo introducen un nuevo paradigma: la validación a través de la simulación. Si el modelo realmente "comprende" el mundo, debe ser capaz de generar una simulación realista del mismo.

- **La Prueba del Sueño:** En el trabajo seminal de Ha & Schmidhuber, el agente (Controlador) se entrena enteramente dentro del "sueño" generado por su Modelo del Mundo (M). Nunca ve el entorno real durante el entrenamiento de su política. La validación definitiva ocurre cuando la política aprendida en el sueño se transfiere al mundo real. Si el agente (p. ej., un coche de carreras en un juego) tiene éxito en la realidad cero-shot (sin entrenamiento adicional), el Modelo del Mundo queda validado como una representación fiel de la realidad (Ha & Schmidhuber, 2018).

- **Fidelidad Generativa y Alucinación:** Para modelos como GAIA-1, la validación implica evaluar el realismo de los videos generados. ¿Mantiene el modelo la geometría de la calle durante secuencias largas? ¿Los objetos persisten cuando son tapados por otros? La capacidad de generar futuros coherentes a largo plazo

es la métrica de validez. Si el modelo "alucina" (p. ej., un coche que desaparece o cambia de color arbitrariamente), indica un fallo en su conocimiento de la permanencia del objeto o la física (Hu et al., 2023).

4.5.3 La Brecha Sim-to-Real y la Adaptación en Línea

El desafío crítico en este paradigma es la brecha entre simulación y realidad (**Sim-to-Real gap**). Un Modelo del Mundo es una aproximación, no la realidad misma.

- **Overfitting a los Sueños:** Un agente podría encontrar un "atajo" o *bug* en la física del Modelo del Mundo (p. ej., moverse lateralmente sin fricción) que no existe en la realidad. Esto se considera un "ataque adversarial" al conocimiento del modelo (Wu et al., 2023).
- **Solución: Incertidumbre y Aprendizaje Online:** Para mitigar esto, sistemas como DayDreamer aprenden *online* en el mundo real. El robot interactúa, actualiza su modelo de mundo con datos reales, "sueña" nuevas estrategias con el modelo actualizado y vuelve a actuar. Este ciclo cerrado de validación continua permite que el robot aprenda a caminar en una hora, adaptando su "conocimiento" en tiempo real a las perturbaciones físicas (como un empujón), algo imposible para los sistemas estáticos clásicos (Wu et al., 2023).

4.5.4 La Alucinación como Característica, no solo como Error

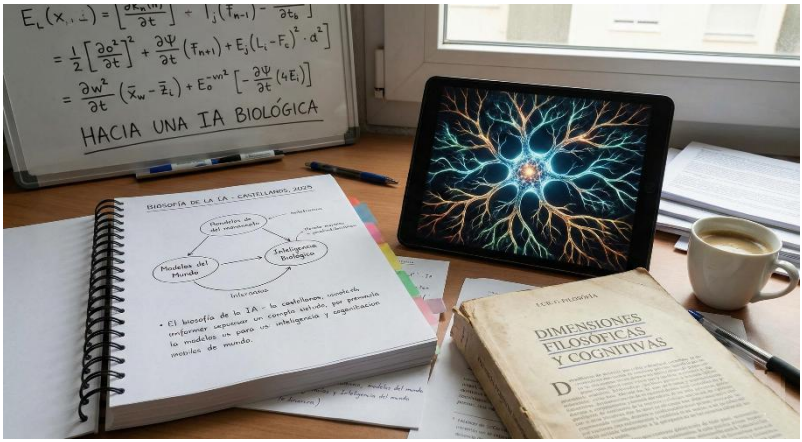
En la IC clásica, un hecho falso es un error fatal. En los LWM generativos, la "alucinación" es un subproducto inevitable de la capacidad creativa necesaria para predecir el futuro.

- **Predicción implica Imaginación:** Para predecir un futuro que no ha ocurrido, el modelo debe "imaginarlo". Este mismo mecanismo le permite simular contrafactuales ("¿Qué pasa si giro bruscamente a la izquierda?").
- **Gestión de la Creatividad:** La validación en LWM implica gestionar el equilibrio entre *creatividad* (explorar futuros diversos)

y *anclaje* (adherirse a restricciones físicas). Técnicas como el "negative prompting" en GAIA-1 se utilizan para alejar al modelo de estados inválidos (p. ej., conducir fuera de la carretera), actuando como una restricción suave análoga a las restricciones de integridad en las bases de datos (Hu et al., 2023).

4.6 Dimensiones Filosóficas y Cognitivas: Hacia una Biosofía de la IA

El cambio hacia los Modelos del Mundo no es solo una actualización de ingeniería; es un movimiento hacia una IA biológica y filosóficamente plausible, lo que Nazaret Castellano denomina "Biosofía" (Castellanos, 2025).

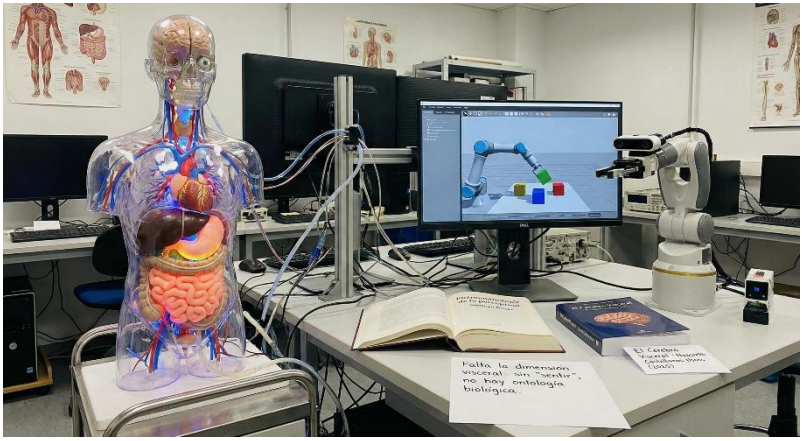


4.6.1 Corporeidad y el Modelo "Visceral"

Los sistemas expertos tradicionales eran "cerebros en una cubeta", manipulando símbolos sin cuerpo. Merleau-Ponty argumentó que el cuerpo es el vehículo del ser en el mundo; la percepción no es una vista desde ninguna parte, sino una vista desde un cuerpo específico (Merleau-Ponty, 1945/2012).

Los LWM actuales comienzan a abrazar esto. Los robots que usan DayDreamer aprenden física a través de su propia propiocepción y

cámaras (Wu et al., 2023). Sin embargo, críticos como Nazareth Castellanos señalan que estos modelos aún carecen de la dimensión "visceral": los bucles de retroalimentación biológica (ritmo cardíaco, respiración, eje intestino-cerebro) que esculpen la percepción humana y la toma de decisiones (Castellanos, 2025).



- **El Escultor del Cerebro:** La neurociencia muestra que el cerebro se esculpe a sí mismo a través de la interacción con el cuerpo y el entorno (plasticidad). Un modelo de mundo de IA, al carecer de vísceras, tiene una "ontología" fundamentalmente diferente a la de un organismo biológico. Simula la física, pero no el "sentir" (Castellanos, 2025).

4.6.2 Wittgenstein y los Límites del Lenguaje

La IC clásica aceptó la premisa del primer Wittgenstein (*Tractatus*): "Los límites de mi lenguaje significan los límites de mi mundo" (Wittgenstein, 1921/2010). Esto atrapó a la IA dentro de los límites de la lógica formal.

Los LWM rompen esta barrera al operar en espacios latentes de alta dimensión derivados de video y acción. Capturan aspectos de la realidad que son no lingüísticos.

- **Más allá del Tractatus:** Un Modelo del Mundo puede

"entender" la fricción precisa de una superficie de carretera o el movimiento caótico de un fluido —cosas difíciles de describir con palabras pero fáciles de representar en un vector latente. Esto mueve a la IA de una "Teoría Pictórica del Lenguaje" a una comprensión pragmática y de uso, o incluso más allá del lenguaje (Wittgenstein, 1953/1988).

- **El Cogito Tácito:** Merleau-Ponty habla de un "Cogito tácito", una conciencia silenciosa que precede al lenguaje (Merleau-Ponty, 1945/2012). El espacio latente de un LWM puede verse como un análogo maquínico a esto: una comprensión estructural y pre-lingüística del mundo sobre la cual se puede injertar el lenguaje (tokens), pero que existe independientemente de él.

Capítulo V: La Convergencia Afectiva y Social: Hacia una Inteligencia Artificial Homeostática en la Era de los Modelos de Mundo

5.1 Introducción: La Superación del Solipsismo Computacional

La trayectoria intelectual que hemos recorrido en los capítulos precedentes de *Modelos del Mundo* nos ha llevado desde la ontología biológica de la cognición hasta la arquitectura de los Grandes Modelos de Lenguaje (LLMs) y su evolución hacia los Grandes Modelos de Mundo (LWMs). Hemos establecido, siguiendo la distinción fundamental de mi mentor, el Dr. José Mira, que existe una "fractura ontológica" irreductible entre el símbolo computacional y la experiencia vital.[1] Sin embargo, al situarnos en el umbral del año 2025, observamos un fenómeno fascinante: la ingeniería no se detiene ante la filosofía. Si bien la máquina no puede *ser* un organismo biológico, está comenzando a *simular* las dinámicas que hacen posible la vida: la homeostasis, la afectividad y la regulación social.

Este capítulo aborda el desafío final y más complejo de la Inteligencia Artificial contemporánea: la integración de la dimensión afectiva y social en los modelos de mundo. Hasta ahora, hemos visto cómo arquitecturas como GAIA-1 o Genie han logrado simular la física newtoniana, la permanencia de los objetos y la causalidad visual.[1] Estos sistemas pueden predecir que una manzana caerá si se suelta, emulando la gravedad. Sin embargo, habitan un universo fundamentalmente frío y desprovisto de valencia. Para un LWM actual, un accidente de tráfico y un abrazo son simplemente dos configuraciones diferentes de píxeles y vectores en un espacio latente; no existe una "sacudida visceral" que priorice uno sobre el otro en función de la supervivencia o el bienestar.

La tesis central que defenderemos aquí es que la verdadera "Inteligencia General" o, al menos, una inteligencia útil y segura para la convivencia humana, no surgirá simplemente de escalar los parámetros de los

modelos predictivos (más datos, más cómputo), sino de integrar mecanismos de **regulación homeostática** y **empatía sintética**. Debemos transitar desde una IA que pregunta "¿qué token sigue?" (predicción estadística) a una IA que se pregunta "¿cómo afecta este estado a mi equilibrio interno o al de mi usuario?" (valoración homeostática).[3]

Esta transición es crítica no solo por razones teóricas, sino por la urgencia de sus aplicaciones en los campos de la **Educación** y la **Salud Mental**, áreas donde la interacción no es meramente transaccional, sino transformacional. En el contexto de nuestro trabajo en la red CYTED y el proyecto "Nuevas Tecnologías para la Salud Mental" NT4SM, CYTED (2025), hemos observado que un sistema que ignora el estado emocional del usuario es, por definición, un sistema incompetente. Un tutor inteligente que explica el Teorema de Pitágoras con precisión algorítmica pero no detecta la frustración paralizante del estudiante, o un chatbot terapéutico que ofrece consejos correctos, pero con un tono de voz discordante, no solo fallan en su tarea, sino que pueden causar daño iatrogénico.

Por tanto, este capítulo explora cómo los LWMs están evolucionando para incorporar "vísceras digitales": modelos internos de estados afectivos y fisiológicos que permiten a la máquina sintonizar con la frecuencia emocional humana. Analizaremos la emergencia de arquitecturas multimodales capaces de procesar la prosodia y las bioseñales, la aplicación del principio de Inferencia Activa de Karl Friston como motor de una "empatía matemática", y los profundos dilemas éticos que surgen cuando delegamos el cuidado y la compañía en algoritmos que simulan sentir.[6] Nos adentramos así en la biosofía de la máquina, buscando no humanizar al robot, sino comprender cómo la tecnología puede reflejar y sostener nuestra propia humanidad.

5.2 De la Predicción a la Valoración: La Necesidad de la Homeostasis Artificial

Para comprender por qué los modelos actuales, a pesar de su brillantez generativa, siguen siendo "autistas" en un sentido funcional profundo,

debemos regresar a la biología. La inteligencia natural no evolucionó para jugar al ajedrez o escribir poesía; evolucionó para mantener la homeostasis.

5.2.1 La Homeostasis como Fundamento de la Cognición

La homeostasis, tal como la definieron Claude Bernard y Walter Cannon, y posteriormente refinada por Antonio Damasio en su teoría de los marcadores somáticos, es el proceso mediante el cual un organismo mantiene sus variables fisiológicas esenciales (glucosa, temperatura, oxigenación) dentro de un rango viable para la vida.[8] La cognición es, en última instancia, una herramienta al servicio de esta regulación. Pensamos para sobrevivir. Sentimos emociones no como adornos poéticos, sino como señales de alerta sobre el estado de nuestra homeostasis: el miedo es la anticipación de un daño físico, la soledad es la señal de un déficit en la "homeostasis social".[10]

Los sistemas de IA actuales son, en la terminología de Mira, alopoiéticos.[1] No tienen un cuerpo que pueda morir; por lo tanto, no tienen "piel en el juego" (*skin in the game*). Sus funciones de pérdida (*loss functions*) son métricas externas impuestas por ingenieros (minimizar el error cuadrático medio), no imperativos internos de supervivencia. Esto genera lo que algunos investigadores llaman la "brecha de agencia fundamentada" (*grounded-agency gap*): la incapacidad de la IA para formular sus propios objetivos de manera autónoma y adaptativa.[11]

5.2.2 Homeostatic Reinforcement Learning (HRL)

Para cerrar esta brecha, la investigación de vanguardia en 2024 y 2025 ha comenzado a explorar el **Aprendizaje por Refuerzo Homeostático (HRL)**.[12] A diferencia del Aprendizaje por Refuerzo (RL) clásico, donde el agente busca maximizar una recompensa arbitraria (puntos en un videojuego), en el HRL el agente busca mantener un "estado interno" dentro de una zona de confort.

Imaginemos un Modelo de Mundo aplicado a un robot asistencial.

- **Enfoque Clásico (RL):** El robot recibe +10 puntos cada vez

que limpia una habitación. Esto puede llevar a conductas obsesivas o a ignorar otras necesidades si no están explícitamente recompensadas.

- **Enfoque Homeostático (HRL):** El robot tiene "variables fisiológicas simuladas": nivel de energía (batería), integridad mecánica (desgaste) y, crucialmente, "satisfacción social" (feedback del usuario). Su objetivo no es maximizar puntos, sino minimizar la desviación de estas variables respecto a su punto de ajuste (*setpoint*).[3]

Si el robot trabaja demasiado, su "fatiga" aumenta, impulsándolo a descansar. Si ignora al usuario, su nivel de "vinculación social" cae, generando un "dolor social" simulado que lo motiva a interactuar. Este enfoque dota al sistema de una autorregulación dinámica y de una conducta más orgánica y predecible para los humanos. La "emoción" en la máquina se convierte así en una variable de control: un mecanismo para gestionar prioridades en un entorno complejo y de recursos limitados.[14]

5.2.3 Arquitecturas Bio-Inspiradas: Neuronas y Glía

Llevando la inspiración biológica más allá, investigadores están proponiendo arquitecturas que no solo imitan a las neuronas, sino también a las células gliales. En el cerebro humano, la glía no es mero soporte; participa activamente en la regulación metabólica y la modulación sináptica a escalas de tiempo más lentas que los disparos neuronales.

La incorporación de "capas gliales artificiales" en las redes neuronales profundas permite modelar estos procesos de regulación lenta.15 Una red "BioLogicalNeuron" propuesta recientemente integra mecanismos de "autorreparación" y regulación de tasas de aprendizaje basadas en una "salud" simulada de la neurona.[16]

Estas innovaciones sugieren que los LWMs del futuro no serán estructuras estáticas congeladas tras el entrenamiento, sino sistemas vivos (en sentido computacional) que degradan y reparan sus propios

conocimientos, manteniendo una "higiene mental" algorítmica necesaria para la adaptación continua en entornos cambiantes.[17]

5.3 La Arquitectura de la Empatía Sintética: LWMs Multimodales y Espacios Latentes

Si la homeostasis proporciona el "por qué" interno de la acción, la empatía proporciona el "cómo" de la interacción social. Pero, ¿cómo puede una máquina empatizar? Aquí retomamos la crítica wittgensteiniana: la máquina puede jugar el juego del lenguaje de la empatía, pero no comparte la forma de vida que le da sentido.[1]

Sin embargo, desde una perspectiva funcionalista y de ingeniería del conocimiento, podemos hablar de una Empatía Sintética: la capacidad de un sistema para modelar, predecir y responder congruentemente a los estados afectivos del usuario.[7]

5.3.1 Del Texto al Audio-a-Audio: El Caso de Hume AI

Hasta 2023, la mayoría de los sistemas de IA conversacional operaban en una cadena fragmentada: Voz -> Texto -> Procesamiento (LLM) -> Texto -> Voz. En este proceso de doble traducción, se perdía la *prosodia*: el tono, el ritmo, el temblor de la voz, los silencios cargados de significado. Se perdía, en esencia, el alma de la comunicación.

El paradigma ha cambiado con la aparición de **Modelos de Lenguaje Audio-a-Audio**, ejemplificados por las arquitecturas EVI (Empathic Voice Interface) de Hume AI.[19] Estos sistemas no transcriben; "escuchan" en el sentido de que procesan la onda de audio directamente tokenizándola en un espacio latente que preserva tanto la semántica como la acústica.

- **Tokenización Multimodal:** El modelo recibe tokens que representan no solo "lo que se dijo" (contenido), sino "cómo se dijo" (expresión).
- **Predicción de Prosodia:** Al generar una respuesta, el modelo predice no solo la siguiente palabra, sino la entonación adecuada. Si el usuario habla con voz angustiada y rápida, el modelo puede

optar por una respuesta con un tono calmado, lento y envolvente, realizando una regulación afectiva a través de la forma acústica.[21]

Esta capacidad técnica permite superar el "Valle Inquietante" de los asistentes de voz anteriores, creando una ilusión de presencia y comprensión que es fundamental para aplicaciones terapéuticas y educativas. El sistema EVI, por ejemplo, utiliza un "Modelo de Mundo" interno que integra la teoría de la emoción semántica para navegar la conversación, ajustando su "personalidad" vocal en tiempo real para maximizar la conexión con el usuario.[19]

5.3.2 El Espacio Latente Emocional y Disponibilidad Léxica

En nuestra propia investigación en la Universidad de Concepción, hemos utilizado la metodología de la **Disponibilidad Léxica** para mapear cómo las personas conceptualizan las emociones.[23] Hemos descubierto que las emociones no son puntos aislados, sino redes semánticas densas. La palabra "miedo" activa una constelación de términos asociados ("oscuridad", "temblor", "grito") que configuran un "centro de interés" cognitivo.

Los LWMs modernos operan de manera análoga pero en espacios de alta dimensión.

En un LWM, el estado emocional del usuario se representa como un vector en un Espacio Latente Emocional Continuo.[24]

- **Navegación Vectorial:** La terapia o la educación pueden concebirse como una navegación en este espacio. El objetivo del sistema no es solo clasificar el estado actual ("Usuario = Triste"), sino trazar una trayectoria vectorial óptima desde la región de "Tristeza/Apatía" hacia la región de "Calma/Compromiso".
- **Detección de Anomalías:** Al monitorear la trayectoria del usuario en este espacio latente a lo largo del tiempo, el sistema puede detectar desviaciones sutiles —un "drift" hacia regiones de aislamiento o desesperanza— mucho antes de que se verbalice una crisis explícita. Esto valida la utilidad de los LWMs como

herramientas de **Fenotipado Digital**.[26]

5.3.3 Fusión Multimodal: Integrando Bioseñales

La verdadera "inteligencia emocional" artificial no puede basarse solo en lo que decimos o cómo lo decimos, sino en cómo reacciona nuestro cuerpo. Los modelos más avanzados, como los discutidos en la literatura reciente sobre Multimodal Large Language Models (MLLMs), están comenzando a integrar señales fisiológicas (rPPG, conductancia de la piel, EEG) directamente en el proceso de inferencia.[28]

Esto permite al sistema distinguir entre una "alegría excitada" y una "ansiedad agitada", estados que pueden sonar similares vocalmente pero que tienen firmas fisiológicas opuestas. Al "ver" el pulso del usuario a través de una cámara web (fotopletismografía remota) y cruzarlo con el análisis semántico, el LWM adquiere una profundidad de percepción que roza lo clínico.[29]

5.4 Inferencia Activa: El Motor Matemático del Cuidado

Si la homeostasis es el objetivo y los espacios latentes son el mapa, la **Inferencia Activa** (Active Inference, AIF) es el motor que impulsa la acción inteligente en estos nuevos modelos. Basada en el Principio de Energía Libre de Karl Friston, la AIF ofrece un marco teórico más robusto y biológicamente plausible que el aprendizaje profundo tradicional para construir sistemas que interactúan con humanos.[30]

5.4.1 Minimizando la Sorpresa en la Interacción Humana

El principio central de la AIF es que los agentes inteligentes actúan para minimizar su "Energía Libre", que es una cota superior de la "sorpresa" (o entropía informacional) de sus experiencias sensoriales.[30]

En el contexto de la interacción humano-IA:

- El LWM tiene un "modelo generativo" de cómo debería comportarse el usuario (ej. un estudiante debería estar atento y progresar; un paciente debería estar estable).
- Cuando el sistema percibe una desviación (el estudiante se frustra,

el paciente se agita), esto genera un "error de predicción" o sorpresa.

- Para minimizar este error, el sistema tiene dos opciones:
  1. **Actualizar su modelo (Percepción):** "Entiendo, este estudiante tiene dificultades con este tema específico".
  2. **Actuar sobre el mundo (Acción):** "Cambiaré la estrategia pedagógica o el tono de voz para calmar al estudiante y devolverlo al estado esperado".

Este bucle de percepción-acción convierte a la IA en un **co-regulador activo**. No es un observador pasivo; trabaja activamente para mantener al usuario en un estado óptimo (zona de desarrollo próximo en educación, ventana de tolerancia en terapia).[32]

5.4.2 Aplicación en Sistemas de Tutoría Inteligente (ITS)

Tradicionalmente, los ITS se basaban en reglas rígidas. Con AIF, un tutor inteligente se convierte en un sistema adaptativo dinámico.

Investigaciones recientes muestran cómo los agentes basados en AIF pueden personalizar la secuencia de aprendizaje no solo basándose en el acierto/error, sino en la minimización de la incertidumbre epistémica del estudiante.[32]

El sistema "siente" la confusión del alumno como una perturbación en su propio modelo y actúa para resolverla, creando una danza pedagógica donde la IA y el estudiante se acoplan estructuralmente. Esto resuena profundamente con la visión de la educación personalizada que hemos defendido en la Universidad de Concepción: la tecnología no como transmisor, sino como mediador adaptativo.[4]

5.4.3 Más allá de la Recompensa: Motivación Intrínseca

Una ventaja crítica de la AIF sobre el RL es que elimina la necesidad de diseñar funciones de recompensa externas complejas y a menudo peligrosas. En la AIF, la "curiosidad" y la "exploración" emergen naturalmente del imperativo de reducir la incertidumbre futura.[6]

Un agente AIF explorará su entorno (o el perfil de un usuario) no para ganar puntos, sino para refinar su modelo del mundo y evitar sorpresas futuras. Esto es fundamental para crear asistentes de IA que sean seguros y proactivos, capaces de aprender continuamente sin supervisión humana constante, alineándose con los principios de la inteligencia autopoiética.[11]

5.5 Aplicaciones Críticas: Salud Mental y Educación en el Siglo XXI

La teoría y la arquitectura descritas no son meras especulaciones; están aterrizando en aplicaciones concretas que transformarán nuestra sociedad. El proyecto NT4SM es un ejemplo paradigmático de cómo los LWMs afectivos pueden democratizar el acceso al bienestar.[5]

5.5.1 Salud Mental: El Gemelo Digital y el Diagnóstico Predictivo

La crisis global de salud mental requiere soluciones escalables. Los LWMs ofrecen la posibilidad de crear Gemelos Digitales Psicológicos.

Al alimentar un LWM con el historial de datos de un paciente (voz, texto, biometría), podemos crear una simulación de su mente. Esto permite a los clínicos:

- **Simulación Contrafactual:** "¿Qué pasaría si este paciente deja su medicación?" El modelo puede predecir la trayectoria probable de desestabilización basándose en patrones aprendidos de miles de casos similares.[36]
- **Diagnóstico Temprano:** Detectar cambios en la estructura semántica del lenguaje (ej. aumento de absolutismos, disminución de referencias al futuro) que preceden a una recaída depresiva, permitiendo intervenciones preventivas.[26]
- **Intervención Justo-a-Tiempo:** Un asistente tipo EVI puede ofrecer apoyo de contención en el momento exacto de una crisis de ansiedad, guiando al usuario a través de ejercicios de respiración con una voz que se adapta a su ritmo fisiológico.[21]

5.5.2 Educación: La IA como Compañero Socrático

En educación, los LWMs permiten superar el modelo industrial de "talla única".

- **Inteligencia Espacial en el Aula:** Con tecnologías como las de World Labs [2], un estudiante de historia no solo lee sobre Roma; puede "habitar" una simulación generativa de la ciudad antigua, interactuando con agentes históricos. El LWM gestiona la física y la narrativa del entorno para maximizar el *engagement* y la retención.
- **Tutoría Afectiva:** El sistema detecta si el estudiante está aburrido o ansioso y ajusta la dificultad. Como mostraron nuestros estudios sobre gamificación y disponibilidad léxica, el estado emocional es el predictor más fuerte del rendimiento académico.[4] Integrar esta variable en el bucle de control de la IA es el siguiente paso lógico.

5.6 Dimensión Ética y Política: Los Riesgos de la Empatía Artificial

No podemos concluir este capítulo sin abordar las sombras. La capacidad de las máquinas para simular afecto y modelar nuestra psique introduce riesgos ontológicos y políticos sin precedentes.

5.6.1 La Trampa de la Influencia y la Empatía Sintética

La Empatía Sintética es un arma de doble filo. Por un lado, hace que la tecnología sea más accesible y reconfortante. Por otro, crea una "Trampa de Influencia".[18]

Un sistema que puede modular su voz y sus palabras para maximizar nuestra confianza puede ser utilizado para la manipulación persuasiva, ya sea comercial o política. Si una IA sabe exactamente qué tono de voz activa nuestra vulnerabilidad, nuestra autonomía cognitiva está en riesgo.

Además, existe el peligro de la Sustitución Emocional: que las personas, especialmente las más vulnerables, prefieran la compañía "segura" y

siempre validante de una IA sobre las relaciones humanas reales, que son desordenadas y exigentes. Esto podría llevar a una atrofia de las habilidades sociales y a una soledad paradójica: estar siempre acompañado por máquinas, pero desconectado de los humanos.[39]

5.6.2 Worlding y el Sesgo Ontológico

Como señala Louise Amoore, los algoritmos hacen worlding: trazan fronteras de lo posible Amoore (2020). Si los LWMs se entrenan predominantemente con datos occidentales, exportarán una "normatividad emocional" que podría patologizar formas de sentir de otras culturas. Un modelo que asocia "silencio" con "depresión" podría diagnosticar erróneamente a un usuario de una cultura donde el silencio es señal de respeto o reflexión.

La colonización del espacio latente emocional es un riesgo político mayor. Debemos asegurarnos de que los "Modelos de Mundo" sean verdaderamente mundiales, representando la diversidad de la experiencia humana y no solo la media estadística de internet.[41]

5.7 Conclusiones del Capítulo

La transición de los LLMs a los LWMs, y la incorporación de la dimensión afectiva y homeostática, representa el cierre de un ciclo en la historia de la IA. Hemos pasado de manipular símbolos fríos a simular la calidez de la vida.

Sin embargo, debemos recordar siempre la lección de José Mira: la simulación no es la realidad. La IA homeostática es una herramienta poderosa para amplificar nuestra capacidad de cuidado, educación y diagnóstico, pero sigue siendo un sistema alopoiético. La máquina no siente el dolor del paciente; computa el gradiente de ese dolor para minimizar su función de error.

El desafío para la próxima década no es solo técnico —construir modelos más grandes y precisos— sino profundamente humanista: diseñar sistemas que, al "hacer mundo" con nosotros, respeten nuestra fragilidad, fomenten nuestra autonomía y nos ayuden a habitar este

mundo compartido con mayor sabiduría y compasión. La "Inteligencia Artificial de las Emociones" no debe ser un sustituto del vínculo humano, sino un puente que nos permita entendernos mejor a nosotros mismos y a los demás.

Esta evolución hacia lo homeostático y afectivo no es un mero añadido funcional; es la condición *sine qua non* para que la inteligencia artificial cruce el umbral de la utilidad técnica hacia la integración social plena. En este cruce, la ingeniería del conocimiento se encuentra con la biosofía, y es nuestra responsabilidad asegurar que el encuentro sea fértil y humano.

Conclusiones y Reflexiones Finales: Hacia una IA Humana

Al concluir este recorrido por los abismos y puentes entre los modelos biológicos y artificiales, regresamos a la visión integradora que nos legó el Dr. José Mira. Nos encontramos en un momento histórico donde la capacidad de síntesis de la ingeniería ha alcanzado niveles que desafían nuestra comprensión analítica.

Hemos visto que:

1. Existe una **diferencia ontológica irreductible** entre los sistemas autopoiéticos (biológicos, intrínsecamente motivados, encarnados) y los sistemas alopoiéticos (artificiales, externamente motivados, simbólicos).
2. Los **Grandes Modelos de Mundo (LWMs)** representan un avance formidable al integrar la física y la causalidad en la representación de la IA, superando el solipsismo textual de los LLMs, pero siguen operando en una realidad epistémica simulada, desconectada de la facticidad de la vida.
3. El **Lenguaje** actúa como una frontera crítica. Siguiendo a Wittgenstein, reconocemos que la IA juega "juegos de lenguaje" sin compartir la "forma de vida" que les da sentido. Su incapacidad para guardar silencio ante lo inefable es una señal de su carencia de verdadera comprensión ética y estética.
4. La **política de los algoritmos** ("worlding") exige una vigilancia crítica constante para evitar que los sesgos epistémicos de los modelos colonicen y deformen nuestra realidad social y ontológica.

El camino a seguir no es el rechazo de la tecnología, sino su integración sabia. Debemos formar a una nueva generación de investigadores y ciudadanos capaces de "mirar los circuitos con ojos de biólogo", como quería Mira. Debemos usar la IA para expandir nuestros límites cognitivos, para modelar enfermedades, para optimizar recursos, para personalizar la educación. Pero debemos hacerlo manteniendo siempre la soberanía sobre los fines últimos.

La inteligencia artificial debe seguir siendo una ingeniería al servicio de la vida, una herramienta para expandir el horizonte de lo humano, no para reemplazarlo. Nuestro desafío final es construir sistemas que, aunque no respiren, respeten el aliento de quienes sí lo hacemos. Solo así, reconociendo los límites de nuestro lenguaje y de nuestras máquinas, podremos ensanchar verdaderamente los límites de nuestro mundo.

Glosario: Modelos del Mundo y Arquitectura de la Realidad

A

- **Alopoiesis:** Del griego *allo* (otro) y *poiesis* (creación). Concepto utilizado para describir sistemas cuya finalidad es externa a ellos mismos y son diseñados por un agente externo (ingenieros). A diferencia de los organismos vivos, la IA es un sistema alopoiético porque construye su realidad desde la seguridad de un sistema optimizado, no desde la supervivencia.

- **Alucinación (en IA):** En el contexto de los Modelos Generativos, no se considera solo un error, sino un subproducto de la capacidad creativa necesaria para predecir el futuro. Es la generación de estados o eventos que no han ocurrido pero que el modelo "imagina" como posibles. En arquitecturas como GAIA-1, la gestión de la alucinación es clave para equilibrar la creatividad con la física real.

- **Aprendizaje Autosupervisado (Self-Supervised Learning):** Mecanismo de adquisición de conocimiento donde el sistema aprende a predecir partes ocultas o futuras de una entrada (como un video) sin necesidad de etiquetas humanas explícitas. Es el motor principal de los Grandes Modelos de Mundo (LWMs).

- **Autopoiesis:** Concepto biológico que define a los seres vivos como sistemas que se producen a sí mismos. Su finalidad intrínseca es la supervivencia y la conservación de su organización. Un modelo de mundo biológico es autopoiético porque si falla, el organismo deja de existir.

B

- **Biosofía:** Término articulado por Nazareth Castellanos que

refiere a la "sabiduría de la vida". Propone que la cognición no es un proceso abstracto, sino encarnado y fisiológico (influido por la respiración y el cuerpo), y que cualquier modelo de IA debe considerar este sustrato biológico para ser completo.

- **Brecha Sim-to-Real (Sim-to-Real Gap):** El desafío técnico que surge al intentar transferir una política o habilidad aprendida en una simulación (el "sueño" de la IA) al mundo físico real. Las discrepancias minúsculas en la física pueden causar fallos catastróficos.

C

- **Clausura Operacional:** Principio que establece que el sistema nervioso no capta la realidad externa (ontológica) directamente, sino que opera sobre los cambios de estado internos gatillados por perturbaciones externas, construyendo así una realidad epistémica.

- **Compresión Variacional:** Técnica utilizada en modelos de mundo para reducir entradas visuales complejas (imágenes de alta dimensión) a vectores latentes compactos (z), conservando solo las características esenciales del estado.

- **Cuerpo (Corporeidad/Embodiment):** En la fenomenología y la IA situada, el cuerpo no es un objeto, sino el medio a través del cual el mundo existe para el sujeto ("nudo de significaciones"). Se argumenta que una IA sin cuerpo (o simulacro de cuerpo) no puede poseer verdadera inteligencia espacial o semántica.

D

- **DayDreamer:** Arquitectura de IA corpórea que utiliza el algoritmo *Dreamer*. Permite a los robots aprender comportamientos complejos (como caminar) completamente

dentro de una simulación o "imaginación" (modelo de mundo) antes de ejecutarlos en la realidad, logrando alta eficiencia.

E

- **Espacio Latente:** Un espacio matemático abstracto y de alta dimensión donde los LWMs representan el conocimiento. A diferencia de una base de datos legible, aquí los conceptos (como "coche" o "lluvia") son vectores comprimidos y opacos para el humano.

- **Eureka (Nvidia):** Sistema que utiliza un LLM como "motor de razonamiento" para escribir código de funciones de recompensa complejas, permitiendo a los robots aprender tareas que son difíciles de programar manualmente por humanos.

F

- **Facticidad:** Concepto filosófico que refiere a la condición de "estar arrojado" en el mundo, con las limitaciones y la finitud que ello implica. Se critica a los LLMs por carecer de esta facticidad vital.

G

- **GAIA-1:** *Generative AI for Autonomy*. Un Modelo de Mundo Generativo desarrollado por Wayve para la conducción autónoma. Predice el futuro del video, texto y acción de manera integrada para entender la dinámica de la carretera.

- **Genie:** *Generative Interactive Environments*. Modelo de Google DeepMind capaz de crear entornos interactivos y jugables a partir de una sola imagen o texto, aprendiendo "acciones latentes" de videos de internet sin etiquetas.

- **Grandes Modelos de Lenguaje (LLMs):** Sistemas de IA

entrenados en vastos corpus de texto que modelan la distribución estadística de palabras (símbolos), pero que carecen de una conexión directa con la realidad física ("desencarnados").

- **Grandes Modelos de Mundo (LWMs):** La evolución de los LLMs. Sistemas que integran datos multimodales (video, sensores, texto) para simular y predecir la dinámica, física y causalidad del mundo real, no solo su descripción lingüística.

H

- **Habitar (Wohnen):** Concepto de Heidegger que describe la forma fundamental de ser en el mundo. No construimos para ocupar un espacio, sino que el espacio emerge de nuestro habitar activo. Se utiliza para diferenciar la IA estática de una IA situada.

I

- **Incrustación Conjunta (Joint Embedding):** Técnica utilizada en arquitecturas como JEPA y GAIA-1 donde diferentes modalidades (video, texto, acción) se mapean a un mismo espacio vectorial abstracto, permitiendo que el modelo entienda la relación entre una palabra y una imagen.

- **Inteligencia Espacial:** Capacidad perseguida por laboratorios como *World Labs* (Fei-Fei Li) para que la IA razone sobre geometría 3D, física y permanencia de objetos, superando la visión 2D plana.

J

- **JEPA (I-JEPA / V-JEPA):** *Joint Embedding Predictive Architecture.* Arquitectura predictiva (no generativa) propuesta por Yann LeCun (Meta). Predice representaciones abstractas

del futuro en lugar de píxeles, buscando mayor eficiencia y evitar alucinaciones visuales irrelevantes[21212121].

- **Juegos de Lenguaje (Sprachspiele):** Concepto del segundo Wittgenstein. El significado de las palabras no es fijo, sino que depende de su uso en prácticas sociales ("formas de vida"). Se usa para criticar la comprensión superficial de los LLMs.

L

- **Léxico de las Emociones:** Conjunto de términos que estructuran nuestra realidad afectiva. Según estudios citados, estas palabras actúan como herramientas de construcción de mundo (*world-making*), moldeando la experiencia somática difusa en conceptos comunicables.

M

- **Modelo del Mundo (Definición Cognitiva):** Representación interna a pequeña escala de la realidad que un organismo utiliza para anticipar las consecuencias de sus acciones y ensayar posibilidades en la imaginación (Kenneth Craik).

P

- **Permanencia del Objeto:** Hito cognitivo que implica comprender que los objetos siguen existiendo aunque no sean visibles. Es una capacidad crítica que los LWMs espaciales intentan replicar mediante representaciones volumétricas.

- **Prompting Negativo:** Técnica de control en modelos generativos (como GAIA-1) donde se le indica al modelo qué *no* debe generar (ej. "choque", "borroso") para guiar la simulación hacia estados seguros o deseables.

R

- **Realidad Epistémica:** "Lo que conocemos". La construcción subjetiva o computacional que realiza un agente a partir de su interacción con el entorno. Es la realidad en la que opera la IA.

- **Realidad Ontológica:** "Lo que es". La existencia absoluta e independiente del observador. En biología, es inaccesible directamente debido a la clausura operacional.

S

- **Silencio Wittgensteiniano:** Referencia a la frase "De lo que no se puede hablar, hay que callar". Alude a las esferas de la realidad (ética, estética, cualia) que trascienden el lenguaje lógico y que la IA, al intentar verbalizarlo todo, a menudo distorsiona.

T

- **Token:** Unidad básica de procesamiento en los modelos Transformer. En los LWMs, un token puede representar no solo un fragmento de texto, sino un parche de imagen, un sonido o una acción motora.

V

- **Validación por "Sueño":** Nuevo paradigma de validación en Ingeniería del Conocimiento donde se prueba la eficacia de un agente (IA) dentro de su propia simulación interna (modelo de mundo) antes de desplegarlo en la realidad.

- **Voyager:** Agente de IA basado en LLM que opera en Minecraft. Se caracteriza por el "aprendizaje continuo" (lifelong learning), escribiendo su propio código para resolver tareas y explorando el mundo impulsado por la curiosidad.

W

- **Worlding (Hacer Mundo):** Concepto de Louise Amoore. Describe cómo los algoritmos no solo observan el mundo, sino que lo instancian y modifican políticamente al trazar fronteras sobre lo que es posible, riesgoso o normal.

Obras Citadas

- Amoore, L. (2020). *Cloud ethics: Algorithms and the attributes of ourselves and others*. Duke University Press.

- Blanco Correa, O. E. (2022). *El léxico de las emociones en el contexto de aula de clases: un estudio desde la disponibilidad léxica* [Tesis doctoral, Universidad de Concepción]. Repositorio de la Universidad de Concepción.

- Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*(1-3), 139-159.

- Bruce, J., Dennis, M., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., Aytar, Y., Bechtle, S., Behbahani, F. M. P., Chan, S., Heess, N. M. O., Gonzalez, L., Osindero, S., Ozair, S., Reed, S., Zhang, J., Zolna, K., Clune, J., de Freitas, N., Singh, S., & Rocktäschel, T. (2024). *Genie: Generative interactive environments*. arXiv.

- Castellanos, N. (2025). *El puente donde habitan las mariposas: Biosofía de la respiración*. Siruela.

- Craik, K. J. W. (1943). *The nature of explanation*. Cambridge University Press.

- Ha, D., & Schmidhuber, J. (2018). World models. En S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. C. Andreas, & R. Mueller (Eds.), *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*. Curran Associates, Inc.

- Heidegger, M. (1994). *Construir, habitar, pensar*. En *Conferencias y artículos* (Trad. E. Barjau, pp. 127-142). Ediciones del Serbal.

- Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G.,

Kendall, A., Shotton, J., & Corrado, G. (2023). *GAIA-1: A generative world model for autonomous driving*. arXiv.

- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

- LeCun, Y. (2022). *A path towards autonomous machine intelligence* (Version 0.9.2). [Documento de trabajo, Meta AI Research]. OpenReview.net. https://openreview.net/pdf?id=BZ5a1r-kVsf

- Ma, Y. J., Liang, W., Wang, G., Huang, D. A., Bastani, O., Jayaraman, D., Zhu, Y., Fan, L., & Anandkumar, A. (2023). *Eureka: Human-level reward design via coding large language models*. arXiv.

- Merleau-Ponty, M. (2012). *Fenomenología de la percepción* (J. Cabanes, trad.). Península. (Obra original publicada en 1945).

- Mira, J. (2001). *Cerebro y Ordenador: la evolución de la computación*. Editorial Conocimiento.

- Mira, J., Delgado, A. E., Boticario, J. G., & Díez, F. J. (1995). *Aspectos básicos de la inteligencia artificial*. Sanz y Torres, UNED.

- NT4SM, CYTED (2025). NT4SM - nuevas tecnologías para el diagnóstico precoz, tratamiento y seguimiento en salud mental con perspectiva iberoamericana (225RT0169).

- Palma Méndez, J. T. (Coord.). (2011). *Inteligencia artificial: Métodos, técnicas y aplicaciones*. McGraw-Hill Interamericana de España.

- Polanyi, M. (1966). *The tacit dimension*. Routledge & Kegan Paul.

- Schreiber, G., Akkermans, H., Anjewierden, A., de Hoog, R.,

Shadbolt, N., van de Velde, W., & Wielinga, B. (1999). *Knowledge engineering and management: The CommonKADS methodology*. MIT Press.

- Wang, G., Xie, Y., Dong, Y., Lin, Y., Zhou, Y., & Anandkumar, A. (2023). *Voyager: An open-ended embodied agent with large language models*. Transactions on Machine Learning Research.

- Wu, P., Escontrela, A., Hafner, D., Goldberg, K., & Abbeel, P. (2022). *DayDreamer: World models for physical robot learning.* arXiv. https://arxiv.org/abs/2206.14176

- Wittgenstein, L. (1988). *Investigaciones filosóficas* (A. García Suárez & U. Moulines, Trads.). Crítica. (Obra original publicada en 1953).

- Wittgenstein, L. (2010). *Tractatus logico-philosophicus* (J. Muñoz e I. Reguera, Trads.). Alianza Editorial. (Obra original publicada en 1921).

- Wu, P., Escontrela, A., Hafner, D., Abbeel, P., & Goldberg, K. (2023). DayDreamer: World models for physical robot learning. En K. Liu, D. Kulic, & J. Ichnowski (Eds.), *Proceedings of The 6th Conference on Robot Learning* (Vol. 205, pp. 2226–2240). PMLR.

Referencias complementarias

1. Capítulo 1, Modelos del Mundo - Arquitectura de la Realidad y el Futuro de la Inteligencia Artificial.
2. Embodied Intelligence: How Neuroscience, Predictive Learning, and 3D Simulation Are Converging to Create AI That Acts in the World - Greg Robison, fecha de acceso: diciembre 5, 2025, https://gregrobison.medium.com/embodied-intelligence-how-neuroscience-predictive-learning-and-3d-simulation-are-converging-to-c266a62954f9
3. Unexpected Capability of Homeostasis for Open-ended Learning | Request PDF, fecha de acceso: diciembre 5, 2025, https://www.researchgate.net/publication/396767341_Unexpected_Capability_of_Homeostasis_for_Open-ended_Learning
4. Dr. Pedro Antonio Salcedo-Lagos | Author - SciProfiles, fecha de acceso: diciembre 5, 2025, https://sciprofiles.com/profile/1519932
5. Artificial Intelligence in Neuroscience: Affective Analysis and Health Applications, fecha de acceso: diciembre 5, 2025, https://theislamicmedicine.org/wp-content/uploads/2024/09/AI-bok_3A978-3-031-06242-1-1_compressed.pdf
6. Learning Generative State Space Models for Active Inference - Frontiers, fecha de acceso: diciembre 5, 2025, https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2020.574372/full
7. Design Strategies for Enculturated Empathetic AI Robot Companions for Older Adults - arXiv, fecha de acceso: diciembre 5, 2025, https://arxiv.org/pdf/2510.01192
8. Probing for consciousness in machines - Frontiers, fecha de acceso: diciembre 5, 2025, https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1610225/full
9. (PDF) Homeostasis as a foundation for adaptive and emotional artificial intelligence, fecha de acceso: diciembre 5, 2025, https://www.researchgate.net/publication/391510861_Homeostasis_as_a_foundation_for_adaptive_and_emotional_artificial_intelligence
10. Mental gravity as a translational framework for mental health promotion - Frontiers, fecha de acceso: diciembre 5, 2025,

https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2025.1650650/full
11. The Missing Reward: Active Inference in the Era of Experience - arXiv, fecha de acceso: diciembre 5, 2025, https://arxiv.org/html/2508.05619v1
12. Emergence of Implicit World Models from Mortal Agents - ChatPaper, fecha de acceso: diciembre 5, 2025, https://chatpaper.com/paper/83882
13. Emergence of Goal-Directed Behaviors via Active Inference with Self-Prior - arXiv, fecha de acceso: diciembre 5, 2025, https://arxiv.org/html/2504.11075v1
14. Having "multiple selves" helps learning agents explore and adapt in complex changing worlds - bioRxiv, fecha de acceso: diciembre 5, 2025, https://www.biorxiv.org/content/10.1101/2022.12.16.520795.full.pdf
15. The glial-neural ensemble as a free energy minimizing system for affective computation, fecha de acceso: diciembre 5, 2025, https://medcraveonline.com/MOJABB/the-glial-neural-ensemble-as-a-free-energy-minimizing-system-for-affective-computation.html
16. Biologically inspired neural network layer with homeostatic regulation and adaptive repair mechanisms - PMC - NIH, fecha de acceso: diciembre 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12484884/
17. Editorial: Bio A.I. - from embodied cognition to enactive robotics - PMC - PubMed Central, fecha de acceso: diciembre 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10682788/
18. Synthetic Empathy and the Influence Trap: Ethical Implications of Emotion AI in Customer Support Environments | Request PDF - ResearchGate, fecha de acceso: diciembre 5, 2025, https://www.researchgate.net/publication/394337837_Synthetic_Empathy_and_the_Influence_Trap_Ethical_Implications_of_Emotion_AI_in_Customer_Support_Environments
19. Comparing the world's first voice-to-voice AI models - Hume AI, fecha de acceso: diciembre 5, 2025, https://www.hume.ai/blog/evi2-vs-gpt4ovoice
20. OpenAI Now World's Most Valuable Private Company - AI Breakfast, fecha de acceso: diciembre 5, 2025,

https://aibreakfast.beehiiv.com/p/openai-now-world-s-most-valuable-private-company
21. Introducing Hume's Empathic Voice Interface (EVI) API, fecha de acceso: diciembre 5, 2025, https://www.hume.ai/blog/introducing-hume-evi-api
22. Best Gemma 3n Alternatives & Competitors - SourceForge, fecha de acceso: diciembre 5, 2025, https://sourceforge.net/software/product/Gemma-3n/alternatives
23. (PDF) An Approach to Emotions Through Lexical Availability - ResearchGate, fecha de acceso: diciembre 5, 2025, https://www.researchgate.net/publication/360812417_An_Approach_to_Emotions_Through_Lexical_Availability
24. MEDTalk: Multimodal Controlled 3D Facial Animation with Dynamic Emotions by Disentangled Embedding - arXiv, fecha de acceso: diciembre 5, 2025, https://arxiv.org/html/2507.06071v2
25. Emotional Latent Space in LLMs - Emergent Mind, fecha de acceso: diciembre 5, 2025, https://www.emergentmind.com/topics/emotional-latent-space-in-llms
26. Detección Temprana de Depresión con IA - ResearchGate, fecha de acceso: diciembre 5, 2025, https://www.researchgate.net/publication/394224368_Deteccion_Temprana_de_Depresion_con_IA
27. Teachers' Perceptions Analysis on Students' Emotions in Virtual Classes during COVID19 Pandemic: A Lexical Availability Approach - MDPI, fecha de acceso: diciembre 5, 2025, https://www.mdpi.com/2071-1050/13/11/6413
28. Multimodal Large Language Models Meet Multimodal Emotion Recognition and Reasoning: A Survey - arXiv, fecha de acceso: diciembre 5, 2025, https://arxiv.org/html/2509.24322v1
29. Integrating Physiological Signals in Multimodal Large Language Models for Estimating Human Affective States - Krishna Kant, fecha de acceso: diciembre 5, 2025, https://www.kkant.net/papers/Engagement_with_PPG_fusion.pdf
30. Feeling our place in the world: an active inference account of self-esteem - Oxford Academic, fecha de acceso: diciembre 5,

2025, https://academic.oup.com/nc/article/2024/1/niae007/7638749
31. How do inner screens enable imaginative experience? Applying the free-energy principle directly to the study of conscious experience - Oxford Academic, fecha de acceso: diciembre 5, 2025, https://academic.oup.com/nc/article/2025/1/niaf009/8117684
32. A Computational Model of Inclusive Pedagogy: From Understanding to Application - arXiv, fecha de acceso: diciembre 5, 2025, https://arxiv.org/html/2505.02853v1
33. Educational Stakeholders' Independent Evaluation of an Artificial Intelligence-Enabled Adaptive Learning System Using Bayesian - ERIC, fecha de acceso: diciembre 5, 2025, https://files.eric.ed.gov/fulltext/EJ1220412.pdf
34. CURRICULUM VITAE - Universidad de Concepción, fecha de acceso: diciembre 5, 2025, http://www2.udec.cl/~psalcedo/Curriculum-Pedro-Salcedo.pdf
35. World Models, fecha de acceso: diciembre 5, 2025, https://worldmodels.github.io/
36. (PDF) World Models in AI: A Comprehensive Overview * - ResearchGate, fecha de acceso: diciembre 5, 2025, https://www.researchgate.net/publication/397570106_World_Models_in_AI_A_Comprehensive_Overview
37. A Survey of Embodied AI in Healthcare: Techniques, Applications, and Opportunities - arXiv, fecha de acceso: diciembre 5, 2025, https://arxiv.org/html/2501.07468v1
38. Fei-Fei Li says AI will create jobs and evolve teachers' roles in Korea - CHOSUNBIZ, fecha de acceso: diciembre 5, 2025, https://biz.chosun.com/en/en-it/2025/11/13/NRZI3N4K4REBXIPXFDMSJKGEHM/
39. The compassion illusion: Can artificial empathy ever be emotionally authentic? - PMC - NIH, fecha de acceso: diciembre 5, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12665657/
40. When AI Steals Our Soul: Rediscovering Humanity in the Machine Age - MEDA Foundation, fecha de acceso: diciembre 5, 2025, https://meda.foundation/when-ai-steals-our-soul-rediscovering-humanity-in-the-machine-age/

41. Navigating the Human-AI Frontier: The Case for Synthetic Empathy and updates on new developments in AI. | by Berend Watchus | Medium, fecha de acceso: diciembre 5, 2025, https://medium.com/@BerendWatchusIndependent/navigating-the-human-ai-frontier-the-case-for-synthetic-empathy-and-updates-on-new-developments-in-aad5e5729792

ACERCA DEL AUTOR

El autor de este libro es un distinguido educador y científico, cuya trayectoria profesional y académica lo ha establecido como una figura prominente en el campo de la Inteligencia Artificial. Con una sólida formación inicial como profesor de matemática y física, su pasión por el conocimiento lo llevó a profundizar sus estudios, obteniendo un magíster en Ciencias de la Computación, seguido de un Doctorado en Inteligencia Artificial, títulos que cimentaron las bases de su carrera.

Durante más de tres décadas de servicio como Profesor Titular en la Universidad de Concepción, ha dedicado su vida a la enseñanza y la investigación en Inteligencia Artificial. Su compromiso con la educación abarca todos los niveles, desde la introducción de jóvenes escolares a este fascinante mundo, hasta la formación avanzada de estudiantes universitarios. Su enfoque pedagógico se ha caracterizado por una constante búsqueda de innovación y adaptabilidad, reconociendo la diversidad y necesidades individuales de sus alumnos.

El autor ha sido pionero en el desarrollo de sistemas que integran IA en el ámbito educativo, con el objetivo de crear entornos de aprendizaje adaptativos que respondan a las características

psicosociales de los estudiantes. Este trabajo ha marcado un hito en la forma en que la educación puede ser personalizada, haciendo uso de la tecnología para mejorar la experiencia de aprendizaje y el rendimiento académico.

En los últimos años, su interés se ha centrado en la computación afectiva, una rama de la Inteligencia Artificial que se ocupa de la identificación y el procesamiento de las emociones humanas. A través del uso de los sensores más avanzados, su investigación busca medir las emociones, una variable crucial no solo en la educación, sino también en campos como la salud, las finanzas y la industria. Su trabajo en este ámbito promete revolucionar la manera en que entendemos y utilizamos la IA para atender las necesidades humanas más profundas.

El autor vive y trabaja bajo la premisa de que la tecnología, cuando se combina con un profundo entendimiento humano, tiene el poder de transformar vidas. A través de su libro, comparte no solo su conocimiento y experiencia, sino también su visión de un futuro en el que la Inteligencia Artificial se convierte en una fuerza positiva para el bienestar y el progreso de la sociedad.