Tendencias IA 2025-2035: Arquitecturas Neurosimbólicas, Modelos del Mundo y Fundamentación o “Grounding”

Introducción – Un repaso sobre ULOGIC/UMIND -

En los artículos anteriores de esta serie sobre ULOGIC/UMIND, hemos hecho un recorrido por los fundamentos de la lógica y filosofía de las matemáticas, el problema de las paradojas, y la búsqueda de soluciones.

Las conclusiones principales han sido tres:

(1) La lógica actual ha terminado siendo “una teoría matemática sobre polinomios dentro de estructuras matemáticas”, abandonando la misión original de ser una “ingeniería inversa del lenguaje”, ingenieria inversa para entender las reglas del razonamiento, la demostración, y la computación. La lógica debe cambiar de rumbo, y volver a entenderse como ingenieria inversa del lenguaje: explicitar las reglas (lo cual a su vez transforma el lenguaje en otro nuevo).

(2) Es esencial entender que existe un ÚNICO lenguaje, y es el que utilizamos para hablar, hacer ciencia, computación y metalingüística (no podemos salirnos de él). Es un lenguaje que hemos inventado en los últimos 3000 años, un lenguaje único que tiene capacidades MÚLTIPLES:

- Capacidad **discursiva** (hablar sobre el mundo)
- Capacidad **argumentativa-débil** (razonamientos imitativos),
- Capacidad **argumentativa-fuerte** (lógica y matemáticas),
- Capacidad **computacional** (definir procedimientos y ejecutarlos)
- Capacidad **metalingüística** (hablar sobre sí mismo, autodescribir las reglas).

(3) La aportación de ULOGIC es un primer paso para demostrar que ES POSIBLE definir un lenguaje exacto, regido por reglas que tiene las capacidades “completas” de nuestro lenguaje “borroso e impreciso”. En ULOGIC razonamiento y computación son lo mismo: procedimientos de manipulación de expresiones. Todo son expresiones (la ejecución de un algoritmo también es una expresión). Todas los resultados se estarían contenidos de forma exacta en una red mundial de TekDocs (un tipo de documento que a su vez también es una expresión). Además es posible dotar a ULOGIC de capacidad auto-meta-lingüística. Y adicionalmente permite una nueva fundamentación de las matemáticas (en particular una solución suficiente a los problemas de la teoría de conjuntos, inspirada en una revisión radical sobre qué són y cómo son las definiciones)

ULOGIC está inspirado y pensado por y para resolver problemas de la lógica, las matemáticas y la filosofía de la ciencia. Es una pieza de “ciencia base” en principio inspirada en cuestiones tan alejadas de la ingeniería como el problema de las paradojas conjuntistas clásicas.

ULOGIC, comparado con un lenguaje formal FOL, CIC, Coq, Lean, etc equivale a comparar una bicicleta con un avión: los principios de construcción son totalmente diferentes (de hecho, radicalmente opuestos) y el resultado es una potencia expresiva infinitamente superior.

La cuestión crucial (e inesperada) es que todo esto tendría una aplicación práctica: empieza a haber un consenso en el campo de la IA de que los enfoques de "escalar-datos-y-potencia" (el camino de los LLMs) es totalmente insuficiente por diseño. Se necesitan arquitecturas modulares con más ingredientes como capacidad de razonamiento exacto, representaciones del mundo y causalidad, e interconexión perceptual para "grounding" (anclar los símbolos con la experiencia).

En este nuevo panorama las arquitecturas **NeuroSimbólicas** (conectar sistemas-intuitivos como los LLMS con sistemas razonadores exactos) se plantean como imprescindibles en el camino (un camino que probablemente requiera más cosas).

Nuestra propuesta (en artículos anteriores) es que utilizar ULOGIC, un lenguaje con capacidad completa de razonamiento, computación y metalingüística, no es meramente una mejora incremental respecto al uso de FOL, Coq, Lean u otros lenguajes usados en arquitecturas NeuroSimbólicas, sino que es el germen para que los sistemas intuitivos (LLMs) puedan **desarrollar la intuición del razonamiento correcto**, además de obtener productos verificables y fiables.

Esto resolvería (por diseño) el problema de la **fiabilidad-verificabilidad-explicabilidad** inalcanzable para un LLMS: Fiable porque el Kernel-Ulogic ha verificado la corrección de los resultados basados en reglas (el sistema intuitivo propone, pero el Kernel dictamina). La explicabilidad es inmediata: los propios productos son su justificación (una demostración, paso a paso de algo, es la explicación de un resultado, análogo para algoritmos y sus resultados).

La confrontación de un LLMs con un Kernel-Ulogic basado en un lenguaje con suficiente potencia expresiva, permitiría que el LLMs desarrolle **la intuición del razonamiento**, computación y metalingüística, y que sus representaciones vectoriales abstractas internas sean modeladas y reflejen (de una forma abstracta) las reglas externas explícitas del lenguaje.

Esta dualidad de capacidades encaja perfectamente con los que los filosófos y los matemáticos han descrito desde siempre: Las matemáticas podrán ser un lenguaje con reglas de verificación, pero el "pensamiento" de un humano-matemático es algo caótico, inspirado en símiles, visiones, sueños intuitivos. Un matemático no consigue resultados porque siga las reglas, sino porque no las sigue. Un humano-matemático "ve los resultados" primero sin necesidad de demostrarlos (aunque normalmente esta intuición está equivocada, y por eso queda luego la dura tarea de probar que la intución es correcta, así es el juego). Es imposible, incluso para una máquina, explorar el infinito combinatorio de reglas aplicables en el juego de las matemáticas: Una máquina sólo podrá "descubrir matemáticas" si desarrolla un motor intuitivo para proponer resultados, propuestas que luego verificará de forma exacta el Kernel (otro componente del sistema).

Hay otro problema insoluble para los LLMS, el de la "**componibilidad**", el que los resultados de conocimiento sean reutilizables para genera más conocimiento, pero también usados para acelerar la adquisición y la comprensión de nuevos problemas. En ULOGIC/UMIND los

productos de la actividad quedan recogidos en **TekDocs**, que constituirían una red mundial de conocimiento exacto verificado. Esto resuelve por diseño el problema de la "componibilidad".

Esto nos lleva a una suposición plausible: ¿Podríamos usar los TekDocs como soporte para una **representación del conocimiento de sentido común**? Esto requeriría una extensión de ULOGIC en una versión v2, para añadir conceptos borrosos de cantidad, probabilidad, aproximación, plausibilidad, temporalidad, modalidad… Pero eso es un proyecto realizable, porque en el fondo es una enorme estructura de conceptos interrelacionados entre sí (eso son las estructuras matemáticas). Esto es expresable en ULOGIC. No difiere mucho de expresar "una teoría de grupos" sólo que con decenas de miles de conceptos interrelacionados. Todavía estaría por resolver el "cuello de botella de la generación" de esa red de conceptos y sus interrelaciones, pero es concebible un "traductor" que puede leer libros "de lenguaje natural" y expresarlos en "lenguaje-ULOGIC-extendido" incorporando esas interrelaciones entre conceptos.

Los problemas del "grounding" se solucionarían como habitualmente se propone con "sistemas perceptuales", pero con un enfoque particular: que esos sistemas realicen representaciones del mundo en espacios abstractos "con topología" y se **use el lenguaje unificado ULOGIC** para etiquetar dinámicamente regiones de ese espacio. Eso sería hablar sobre el mundo.

El problema de la planificación y fijación de objetivos es todavía una capa superior muy especultativa en este momento, pero si ULOGIC es capaz de expresar no sólo algoritmos sino también "procedimientos generales" **entonces es capaz de expresar "planes"** (porque un plan es un procedimiento general). ¿Necesitaríamos ir más allá y que fije objetivos? Quizá solo objetivos intermedios para consecución de objetivos generales que nosotros le planteemos.

Pero la pregunta principal es ¿qué queremos conseguir con ULOGIC/UMIND? ¿Queremos una "mente humana"? El problema de crear una mente con capacidades humanas es muy sencillo, lo resolvió la biología, cualquiera puede hacerlo, y se tardan 9 meses (se necesitan dos personas normalmente para conseguirlo). En definitiva: mentes humanas ya tenemos.

Lo que entendemos debería ser nuestro objetivo es **construir "mentes-artificiales-científicas" que nos ayuden en el razonamiento, la resolución de problemas y la investigación científica**, como asistentes pero también como "**co-descubridores fiables y explicables**".

Ese sería el objetivo de ULOGIC/UMIND.

Siguientes pasos y preguntas:

Inevitablemente surge la cuestión de si este camino podría ser una "solución posible hacia la AGI" (artificial general intelligence). Parece plausible que un componente neurolingüístico potente, cimentado sobre un lenguaje de expresividad superior como ULOGIC, sea el paso inicial para desde ahí construir todo lo demás. Por lo menos hay un indicio fuerte: imita milimétricamente la forma en que los científicos y los matemáticos "piensan", por un lado un componente intuitivo como motor creativo que alimenta al Kernel-exacto árbitro del juego.

¿Pero cuál es el panorama actual reciente sobre la posibilidad y caminos a explorar para construir una AGI según autores reconocidos en el campo? Tener una visión general al respecto es necesaria para intentar al menos comprender "el tamaño del elefante".

Lo que sigue es un informe recopilado desde diversas fuentes, suficientemente compacto y breve, pero al mismo tiempo con cierta extensión, que permite visualizar las tendencias en inteligencia artificial para los años 2025-2035 según las figuras y actores punteros actualmente en este campo.

Parte I: El Espacio del Problema de la AGI y los Límites de la Escala

Sección 1: Formalizando la Inteligencia Artificial General

La búsqueda de la Inteligencia Artificial General (AGI, por sus siglas en inglés) representa el objetivo original y más ambicioso del campo de la IA: la creación de sistemas con una inteligencia de nivel humano, capaces de comprender, aprender y adaptarse a una amplia gama de tareas intelectuales.

1.1. Definiendo la Inteligencia: De la Optimización de Tareas a la Competencia General

Existen dos perspectiva, una eminentemente matemática y formal, define la inteligencia en términos de rendimiento y optimización. La otra, de inspiración más cognitiva, la define en términos de arquitectura y capacidades subyacentes.

La **perspectiva de la inteligencia universal**, propuesta por Shane Legg y Marcus Hutter, define formalmente la inteligencia como la capacidad de un agente para alcanzar objetivos en una amplia gama de entornos. Esta definición se basa en el marco teórico de AIXI, un agente de aprendizaje por refuerzo bayesiano que, en teoría, es óptimamente inteligente.
Este enfoque es valioso por su formalismo y generalidad, ya que no es antropocéntrico y abarca desde agentes simples hasta superinteligencias. Sin embargo, su naturaleza es puramente conductual y externa; mide el rendimiento sin especificar la arquitectura interna. La búsqueda de aproximaciones a esta inteligencia universal conduce naturalmente a la hipótesis de la escala: si la inteligencia es rendimiento óptimo, entonces escalar los datos, los parámetros y la computación debería aproximar mejor dicho óptimo.

En contraste, la **perspectiva de los sistemas cognitivos**, defendida por investigadores como Ben Goertzel, define la AGI en función de sus capacidades cognitivas de tipo humano. Según Goertzel, una AGI debe poseer "un grado razonable de autocomprensión y autocontrol autónomo, y tener la capacidad de resolver una variedad de problemas complejos en una variedad de contextos, y de aprender a resolver nuevos problemas que no conocía en el momento de su creación". Esta definición no se centra en el rendimiento externo, sino en la arquitectura interna y la flexibilidad cognitiva. No se trata solo de *qué* hace el sistema, sino de *cómo* lo hace. Esta visión implica que la inteligencia requiere módulos específicos para la autorrepresentación, la planificación y el aprendizaje, lo que conduce directamente a la hipótesis de la arquitectura cognitiva.

1.2. Una Enumeración Pragmática de las Capacidades Fundamentales de la AGI

Sintetizando la investigación en el campo, podemos establecer un conjunto de capacidades necesarias, aunque quizás no suficientes, que cualquier sistema que aspire a ser una AGI debe demostrar. Estas capacidades servirán como criterios de evaluación para las arquitecturas discutidas en este informe.

- **Razonamiento y Estrategia:** La habilidad de emplear la lógica, la estrategia, resolver acertijos y tomar decisiones bajo incertidumbre. Esto implica ir más allá de la simple coincidencia de patrones para realizar inferencias deductivas, inductivas y abductivas.
- **Representación del Conocimiento:** La capacidad de construir, mantener y utilizar un modelo interno rico y coherente del mundo, que incluya conocimiento de sentido común. Este modelo debe representar entidades, sus propiedades, sus relaciones y las reglas que gobiernan sus interacciones.
- **Planificación y Abstracción:** La capacidad de establecer y perseguir objetivos, descomponer problemas complejos en sub-problemas manejables y operar en múltiples niveles de abstracción. Una AGI debe ser capaz de formular planes a largo plazo y ajustarlos dinámicamente.
- **Aprendizaje y Adaptación:** La capacidad de aprender de manera eficiente y continua a partir de la experiencia, con una mínima intervención humana. Esto incluye el aprendizaje por transferencia (*transfer learning*), el aprendizaje de pocos ejemplos (*few-shot learning*) y la auto-enseñanza, permitiendo al sistema adaptarse a situaciones y dominios novedosos para los que no fue explícitamente programado.
- **Comunicación Fundamentada (*Grounded*):** La habilidad de utilizar el lenguaje natural de una manera que esté significativamente conectada con el mundo real y las representaciones internas del sistema. El lenguaje no debe ser un mero juego de símbolos, sino una herramienta para describir y razonar sobre el mundo.
- **Interacción Corpórea (*Embodied Interaction*):** Aunque no es un requisito universalmente aceptado, la capacidad de percibir el mundo a través de múltiples modalidades sensoriales (vista, oído, tacto) y de actuar sobre el entorno físico se considera cada vez más un catalizador crucial para el desarrollo de una inteligencia verdaderamente general. La cognición corpórea postula que la inteligencia se desarrolla a través de la interacción con el mundo físico.

El debate entre la definición de inteligencia universal y la de sistemas cognitivos no es meramente académico; define las estrategias de investigación:

- La primera sugiere que la AGI podría "emerger" de la escala,
- mientras que la segunda sostiene que debe ser "diseñada" con una arquitectura específica.

La evidencia acumulada sobre las limitaciones de los modelos a gran escala apoya firmemente la segunda visión, haciendo de la arquitectura cognitiva el camino más viable hacia la AGI

Sección 2: La Hipótesis de la Escala y sus Deficiencias Fundamentales

La era actual de la IA (2016-2025) está dominada por la hipótesis de la escala, la creencia de que la inteligencia general puede surgir simplemente aumentando el tamaño de los modelos, la cantidad de datos de entrenamiento y el poder computacional.

Los Modelos de Lenguaje Grandes (LLMs) son la máxima expresión de esta hipótesis. A pesar de sus impresionantes capacidades en la generación de lenguaje y el reconocimiento de patrones, que los sitúan como sofisticados sistemas de "Pensamiento de Sistema 1" —rápidos, intuitivos y basados en hábitos—, sus limitaciones no son meros problemas de ingeniería que se resolverán con más escala, sino fallos arquitectónicos fundamentales que impiden su progresión hacia la AGI.

El análisis de estas limitaciones revela que no son fallos aislados, sino una cascada de síntomas que emanan de una causa raíz: la ausencia de un modelo del mundo estructurado, causal y fundamentado en la realidad (sumado a la incapacidad de razonamiento riguroso real)

2.1. Limitación Central 1: El Problema del Anclaje de Símbolos (*Symbol Grounding*)

El problema del anclaje de símbolos, formulado por Stevan Harnad, pregunta cómo los símbolos dentro de un sistema formal pueden adquirir un significado intrínseco, en lugar de ser meramente "parásitos" de los significados que existen en nuestras mentes.

Un LLM opera en un sistema de símbolos no anclados; las palabras y los tokens que manipula no tienen una conexión inherente con los objetos, propiedades o relaciones del mundo real. Son parte de un "carrusel de símbolo a símbolo", donde el significado de un símbolo se define únicamente en términos de otros símbolos.

Esta falta de anclaje es la deficiencia más profunda de los LLMs. Significa que, a un nivel fundamental, no "comprenden" los conceptos que procesan. Sus operaciones son puramente sintácticas, basadas en las relaciones estadísticas aprendidas de un corpus masivo de texto, sin acceso a la semántica del mundo real que ese texto describe.

NOTA: Los tres párrafos anteriores son una “descripción en lenguaje habitual” del problema. Ahora bien, en mi opinión, la comprensión del “significado es algo más complejo” y tiene como mínimo tres niveles (por lo menos):

1. **Significado intensional-débil**, que unas palabras tienen en su relación con otras, dentro de un sistema con capacidad “discursiva-narrativa”. Si sabes “unir palabras” es que has “entendido el significado intensional”.

2. **Significado intensional-fuerte** que unas palabras tienen en su relación con otras y con las reglas de producción exactas del sistema (que premite ver relaciones lógicas más allá de las meramente narrativas). Si además de unir palabras, tienes coherencia lógica,

has “entendido el significado intensional-fuerte. (Imposible si no sabes que existen las reglas lógicas)

3. **Significado denotacional**, que las palabras tienen porque el agente que las usa tiene sistemas perceptuales que conecta las percepciones (estados internos) con el lenguaje. Imposible si no hay sistemas perceptuales.

Es incorrecto decir que los LLMS no entienden el significado. Claro que lo entienden, pero el significado intensional-débil, careciendo de significado intensional fuerte y por supuesto de significado denotacional (inexistente si no hay sistemas perceptuales)

2.2. Limitación Central 2: Ausencia de Modelos del Mundo Coherentes

Como consecuencia directa de la falta de anclaje, los LLMs no construyen ni mantienen modelos del mundo explícitos, dinámicos e interpretables. Un modelo del mundo es una representación interna de las entidades, sus estados y las reglas que gobiernan sus interacciones. Los LLMs, en cambio, construyen un mapa estadístico masivo de co-ocurrencias de tokens.

El ajedrez sirve como un ejemplo paradigmático de esta falla. Un LLM puede recitar las reglas del ajedrez, describir aperturas famosas e incluso predecir el siguiente movimiento más probable en una partida conocida, porque toda esta información existe como texto en sus datos de entrenamiento.

Sin embargo, si se le presenta una configuración de tablero novedosa o se le pregunta por las implicaciones de una regla de ajedrez no estándar, su rendimiento se degrada catastróficamente. No puede "razonar" sobre el tablero porque no posee una representación interna, manipulable y coherente del mismo; solo puede recuperar patrones textuales asociados con el ajedrez.

Aunque algunas investigaciones sugieren que los LLMs pueden desarrollar "modelos del mundo internos" para tareas específicas, estos son implícitos, frágiles y no están disponibles para el razonamiento deliberado y de propósito general que requiere la AGI. No se puede señalar una estructura de datos dentro de un LLM y decir: "aquí es donde se almacena el estado del tablero de ajedrez".

2.3. Limitación Central 3: Fracaso del Razonamiento Causal y Lógico

La ausencia de un modelo del mundo que separe las entidades de sus mecanismos causales subyacentes hace que el razonamiento causal robusto sea imposible para los LLMs. Utilizando la "Escalera de la Causalidad" de Judea Pearl como marco, los LLMs operan casi exclusivamente en el primer peldaño: la Asociación. Son expertos en identificar correlaciones en el texto (p. ej., "el humo se asocia con el fuego"), pero fallan sistemáticamente en los peldaños superiores:

- **Peldaño 2 (Intervención):** No pueden predecir de forma fiable los resultados de una acción o intervención (p. ej., "¿qué pasaría si prevengo el fuego?").

- **Peldaño 3 (Contrafácticos):** No pueden razonar sobre escenarios hipotéticos que contradicen los hechos (p. ej., "si no hubiera habido fuego, ¿habría humo?").

Esta incapacidad se manifiesta de varias maneras. Los LLMs confunden correlación con causalidad, se ven fuertemente influenciados por el orden temporal de los eventos en el texto (asumiendo que lo que se menciona primero es la causa), y su rendimiento se degrada cuando una narrativa contradice el conocimiento paramétrico almacenado en sus pesos. Esto conduce a inconsistencias lógicas y a una incapacidad para llevar a cabo un razonamiento robusto de múltiples pasos, un requisito indispensable para la inteligencia general.

2.4. Limitación Central 4: Generalización Sistemática y Composicionalidad

La inteligencia humana se caracteriza por la **composicionalidad sistemática**: la capacidad de comprender y producir un número infinito de expresiones novedosas combinando un conjunto finito de elementos conocidos (palabras, conceptos) de acuerdo con reglas. Esta capacidad es la base de la generalización robusta.

Los LLMs carecen de esta habilidad. Su "conocimiento" no es una gramática composicional de conceptos, sino un mapa estadístico plano. Aprenden a reconocer combinaciones de alto nivel que son frecuentes en los datos de entrenamiento, pero luchan por generalizar a combinaciones novedosas que son semánticamente válidas pero estadísticamente raras. Por ejemplo, un modelo que ha visto "hombre persigue perro" y "mujer monta a caballo" puede no ser capaz de interpretar correctamente "mujer persigue a caballo". Su generalización es, por tanto, frágil y poco fiable fuera de la distribución de sus datos de entrenamiento.

2.5. Limitación Central 5: Falta de Fiabilidad y Barreras Prácticas

Estas profundas fallas arquitectónicas se manifiestan en una serie de problemas prácticos que hacen que los LLMs, en su forma actual, no sean aptos para ser el núcleo de una AGI.

- **Alucinaciones:** La generación de información plausible pero fácticamente incorrecta es una consecuencia directa de la falta de anclaje y de modelos del mundo. Sin un mecanismo para verificar los hechos contra un modelo de realidad coherente, el modelo simplemente genera la secuencia de tokens más probable, que puede o no corresponder a la verdad.

- **Olvido Catastrófico:** La arquitectura de los LLMs los hace inherentemente susceptibles a olvidar información previamente aprendida cuando se les ajusta (*fine-tuning*) en nuevas tareas. Un verdadero agente AGI debe ser capaz de aprender de forma continua e incremental sin degradar el conocimiento existente, una capacidad que los LLMs no poseen.

- **Ineficiencia de Datos y Cómputo:** Las propias "leyes de escala" que han impulsado el éxito de los LLMs también apuntan a un camino insostenible. La necesidad de cantidades astronómicas de datos y de una potencia computacional cada vez mayor para obtener mejoras marginales sugiere que este enfoque se enfrenta a rendimientos decrecientes y a barreras físicas y económicas.

En resumen, los fallos de los LLMs no son errores aislados que puedan ser parcheados individualmente con más datos o soluciones ad-hoc como la Generación Aumentada por Recuperación (RAG).

Son síntomas interconectados de una enfermedad arquitectónica central. La falta de anclaje impide la construcción de modelos del mundo; la ausencia de modelos del mundo impide el razonamiento causal; y la falta de un modelo causal y composicional impide la generalización sistemática.

Esta diagnosis unificada apunta a una conclusión ineludible: el camino hacia la AGI no pasa por escalar los LLMs, sino por un cambio de paradigma fundamental hacia arquitecturas que diseñen explícitamente sistemas para el **anclaje**, la modelización del **mundo** y el **razonamiento**.

Limitación Fundamental	Descripción	Causa Arquitectónica Raíz	Investigadores/Artículos Clave
Falta de Anclaje de Símbolos	Los símbolos (tokens) carecen de significado intrínseco; su "comprensión" se basa en relaciones estadísticas con otros símbolos, no con el mundo real.	El modelo se entrena exclusivamente con datos textuales, sin conexión perceptual o de acción con un entorno.	Harnad (1990) , Bender & Koller (2020)
Ausencia de Modelos del Mundo	Incapacidad para construir y mantener representaciones internas coherentes, dinámicas y manipulables de entidades, estados y sus reglas.	Consecuencia de la falta de anclaje. Sin símbolos con significado, no se puede construir un modelo semántico del mundo, solo un mapa estadístico del lenguaje.	Gary Marcus , Yann LeCun
Fallo en Razonamiento Causal	Confunde correlación con causalidad. No puede razonar sobre intervenciones o contrafácticos. Su rendimiento se degrada con narrativas no cronológicas.	Ausencia de un modelo del mundo que separe las entidades de los mecanismos causales. Opera en el nivel de asociación, no de intervención.	Judea Pearl , Kıcıman et al. (2023)
Generalización No	Dificultad para	El conocimiento no	Lake & Baroni (2023) ,

Limitación Fundamental	Descripción	Causa Arquitectónica Raíz	Investigadores/Artículos Clave
Sistemática	generalizar a combinaciones novedosas de conceptos conocidos (falta de composicionalidad). Generalización frágil fuera de la distribución de entrenamiento.	está estructurado de forma composicional, sino como un mapa estadístico plano de patrones.	Fodor & Pylyshyn (1988)
Alucinaciones	Generación de información plausible pero fácticamente incorrecta o inconsistente.	Consecuencia directa de la falta de anclaje y de un modelo del mundo contra el cual verificar la veracidad de las afirmaciones generadas.	Lin et al. (2022) , Weidinger et al. (2022)
Olvido Catastrófico	El ajuste fino en nuevas tareas degrada o destruye el conocimiento previamente aprendido.	La actualización de los pesos del modelo para una nueva tarea interfiere con las representaciones distribuidas del conocimiento antiguo.	Kirkpatrick et al. (2017), Li & Hoiem (2017)

Parte II: Paradigmas Fundamentales para las Arquitecturas Cognitivas

La crítica a la hipótesis de la escala nos obliga a buscar alternativas constructivas.

Varios de los investigadores más influyentes en IA han propuesto planos para arquitecturas cognitivas que abordan directamente las deficiencias de los modelos monolíticos.

Estas visiones, aunque diversas en sus detalles de implementación, convergen en un conjunto de tres principios fundamentales:

- **modularidad**,
- la centralidad de los **modelos del mundo**
- y la necesidad de integrar el aprendizaje con el **razonamiento**.

Sección 3: Planos para Arquitecturas Cognitivas

3.1. Yann LeCun: El Modelo del Mundo como Componente Central

Yann LeCun postula que el camino hacia la inteligencia autónoma no reside en el aprendizaje por refuerzo tradicional, que es ineficiente y requiere un número masivo de ensayos, sino en el **aprendizaje predictivo basado en modelos del mundo**. Un agente inteligente debe ser capaz de predecir las consecuencias de sus acciones y las de otros, lo que le permite razonar y planificar de manera eficiente. Su propuesta es una arquitectura cognitiva modular, donde cada componente es diferenciable y, por tanto, entrenable mediante métodos basados en gradientes. La arquitectura propuesta por LeCun consta de varios módulos interconectados :

1. **Módulo de Percepción:** Este módulo recibe entradas sensoriales brutas (imágenes, audio, texto) y las transforma en representaciones abstractas del estado actual del mundo.

2. **Modelo del Mundo:** Es el corazón del sistema. Recibe la representación del estado actual del módulo de percepción y una secuencia de acciones hipotéticas del módulo Actor. Su función es predecir los estados futuros del mundo. Este modelo debe ser capaz de manejar la incertidumbre inherente al mundo real, prediciendo múltiples futuros plausibles.

3. **Módulo de Coste:** Este es el motor de motivación intrínseca del agente. Evalúa el "malestar" o coste asociado a un estado del mundo. Se compone de dos sub-módulos:
    - **Coste Intrínseco:** Un coste inmutable, programado, que representa impulsos básicos (p. ej., evitar daños).
    - **Crítico (*Critic*):** Un módulo entrenable que aprende a predecir el coste futuro acumulado a partir de un estado dado, permitiendo una planificación a más largo plazo.

4. **Módulo Actor:** Su función es generar secuencias de acciones que minimicen el coste predicho por el Modelo del Mundo y el Módulo de Coste. Realiza una búsqueda u optimización en el espacio de secuencias de acciones para encontrar la más adecuada.

5. **Memoria a Corto Plazo:** Mantiene y actualiza el estado del mundo, tanto el percibido como el predicho, sirviendo como un espacio de trabajo para el razonamiento.

La tecnología clave que LeCun propone para implementar el Modelo del Mundo es la **Arquitectura Predictiva de Incrustación Conjunta (JEPA, *Joint Embedding Predictive Architecture*)**. A diferencia de los modelos generativos que intentan predecir cada detalle de la entrada (p. ej., cada píxel de una imagen futura), lo que es computacionalmente costoso e innecesario, JEPA opera en un espacio de representación abstracto. Codifica una entrada (p. ej., un fotograma de vídeo) en una representación abstracta y luego predice la representación abstracta de una entrada futura (p. ej., el siguiente fotograma). Al predecir en este espacio latente, el modelo puede ignorar detalles irrelevantes y difíciles de predecir, centrándose en la semántica esencial del mundo, lo que lo hace mucho más eficiente y robusto para aprender

modelos del mundo.

3.2. Yoshua Bengio: Cognición de Sistema 2 y el Prior de Conciencia

Yoshua Bengio aborda el problema de la AGI a través del prisma de la psicología cognitiva, utilizando la distinción de Daniel Kahneman entre el **Sistema 1** y el **Sistema 2** de pensamiento.

- **Sistema 1:** Es rápido, intuitivo, inconsciente y paralelo. Corresponde a las capacidades actuales del *deep learning*, como el reconocimiento de patrones.
- **Sistema 2:** Es lento, lógico, secuencial y consciente. Implica razonamiento, planificación y manipulación de conceptos abstractos. Bengio argumenta que alcanzar la AGI requiere desarrollar un robusto Sistema 2.

La pieza central de la propuesta de Bengio es el **Prior de Conciencia (*Consciousness Prior*)**. Este no es un componente arquitectónico, sino un poderoso sesgo inductivo para el aprendizaje de representaciones. La hipótesis se inspira en la Teoría del Espacio de Trabajo Global de la conciencia, que postula que la conciencia actúa como un cuello de botella: de un vasto conjunto de procesamientos inconscientes, unos pocos elementos son seleccionados por la atención y se "transmiten" globalmente para condicionar el procesamiento posterior.

Formalmente, el Prior de Conciencia postula que los "pensamientos conscientes" son combinaciones de baja dimensión de unos pocos conceptos de alto nivel. Una frase como "el gato está sobre la alfombra" involucra pocos conceptos (gato, estar sobre, alfombra) pero hace una afirmación fuerte y probablemente verdadera sobre el mundo. Esto implica que la distribución de probabilidad conjunta sobre los conceptos de alto nivel tiene la estructura de un **gráfico de factores disperso (*sparse factor graph*)**. En este grafo, cada factor (que representa una dependencia) conecta solo un pequeño número de variables (conceptos), pero la dependencia puede ser muy fuerte. Esta estructura es inherentemente composicional y permite una generalización combinatoria, abordando una de las principales debilidades de los LLMs.

Para implementar los mecanismos de muestreo y búsqueda necesarios para el razonamiento de Sistema 2, Bengio y su equipo han desarrollado las **Redes de Flujo Generativo (GFlowNets)**. Las GFlowNets son una clase de modelos generativos diseñados para muestrear objetos composicionales (como gráficos, ecuaciones o planes) con una probabilidad proporcional a una función de recompensa o energía. A diferencia de los métodos MCMC que pueden tardar en mezclar, las GFlowNets aprenden una política para construir estos objetos de forma secuencial, lo que las hace muy adecuadas para tareas de búsqueda estructurada en los vastos espacios combinatorios que caracterizan el razonamiento de Sistema 2.

3.3. Geoffrey Hinton: Paradigmas Alternativos de Aprendizaje y Representación

Aunque Geoffrey Hinton no ha propuesto una arquitectura cognitiva completa y unificada como LeCun o Bengio, su trabajo reciente ofrece componentes alternativos cruciales que podrían ser fundamentales para cualquier futura AGI.

1. **El Algoritmo Forward-Forward (FF):** Hinton ha criticado la plausibilidad biológica de la retropropagación (*backpropagation*) y ha propuesto el algoritmo FF como alternativa. FF reemplaza las pasadas hacia adelante y hacia atrás de la retropropagación con dos pasadas hacia adelante: una con datos "positivos" (reales) y otra con datos "negativos" (generados o incorrectos). Cada capa de la red tiene su propio objetivo local: maximizar una "métrica de bondad" (p. ej., la suma de las actividades neuronales al cuadrado) para los datos positivos y minimizarla para los negativos. Al normalizar la actividad entre capas, se obliga a cada capa a aprender nuevas características. Este aprendizaje local y sin retropropagación podría ser una forma mucho más eficiente y biológicamente plausible de entrenar las redes neuronales profundas que componen los módulos de percepción y actuación en una arquitectura AGI.

2. **Redes de Cápsulas (*Capsule Networks*):** Las Redes Neuronales Convolucionales (CNNs) logran invarianza a la traslación a través del *pooling*, que descarta información de posición precisa. Las Redes de Cápsulas son una alternativa diseñada para preservar la información espacial y manejar las relaciones jerárquicas parte-todo de manera más robusta. Una "cápsula" es un grupo de neuronas cuyo vector de actividad representa los parámetros de instanciación de una entidad (como su posición, orientación, tamaño, etc.). Las cápsulas de nivel superior hacen predicciones para las poses de las cápsulas de nivel inferior, y se activan si múltiples predicciones coinciden, un proceso llamado "enrutamiento por acuerdo" (*routing by agreement*). Esta arquitectura tiene una composicionalidad y una equivariancia (la capacidad de entender cómo las transformaciones en el objeto afectan a su representación) incorporadas, lo que podría dar lugar a módulos de percepción mucho más robustos y generalizables que los basados en CNNs estándar.

3.4. Ben Goertzel: Un Marco Integrador y de Código Abierto para la AGI

Ben Goertzel y el proyecto OpenCog ofrecen una visión pragmática y constructiva de la AGI, materializada en el marco de código abierto **OpenCog Hyperon**. La filosofía central es la "sinergia cognitiva": la idea de que la inteligencia general no surgirá de un único algoritmo maestro, sino de la interacción cooperativa y dinámica de múltiples algoritmos y paradigmas de IA.
La arquitectura de Hyperon se basa en dos componentes clave:

1. **Atomspace:** Es el almacén de conocimiento universal del sistema. Técnicamente, es un metagrafo ponderado y etiquetado, una estructura de datos extremadamente flexible capaz de representar conocimiento de diversas formas: declarativo (hechos), procedimental (programas), lingüístico, de atención, de objetivos, etc.. A diferencia de una base de datos tradicional, el Atomspace está diseñado para la manipulación y transformación dinámica por parte de los algoritmos de IA.

2. **MeTTa (Meta-Type Talk):** Es el lenguaje de programación del sistema. MeTTa es un lenguaje funcional, probabilístico y fuertemente tipado, diseñado específicamente para que diversos algoritmos de IA (redes neuronales, razonadores lógicos, algoritmos evolutivos) puedan operar sobre el Atomspace e interactuar entre sí. Permite que un algoritmo manipule o incluso reescriba a otro, facilitando la emergencia de procesos

cognitivos complejos y la auto-organización del sistema.

La visión de Goertzel es menos una arquitectura fija y más un ecosistema para que la inteligencia emerja. Proporciona un lenguaje y un espacio de trabajo comunes donde diferentes especialistas (un reconocedor de imágenes neuronal, un probador de teoremas lógicos) pueden colaborar para resolver problemas que ninguno podría resolver por sí solo.

RESUMEN:

Al analizar estas cuatro visiones, emerge un consenso implícito. A pesar de las diferencias en la terminología y los mecanismos de implementación propuestos (JEPA, Prior de Conciencia, Atomspace), todos estos pioneros convergen en la necesidad de una arquitectura modular y de inspiración cognitiva. Rechazan la idea de que la AGI surgirá de una caja negra monolítica entrenada de extremo a extremo.

En su lugar, proponen sistemas compuestos por módulos especializados que interactúan: un módulo para la percepción, otro para el razonamiento y la planificación, y un sistema de motivación. La pregunta central de la investigación en AGI ya no es "escala vs. arquitectura", sino **"¿qué arquitectura?".**

El siguiente paso es detallar cómo se puede construir el módulo de razonamiento, el componente que dota a estos sistemas de "pensamiento".

Sección 4: Integración Neurosimbólica: La Síntesis del Aprendizaje y el Razonamiento

Si las arquitecturas cognitivas modulares son el "qué" del camino hacia la AGI, la integración neurosimbólica (NeSy) es el "cómo".

NeSy es el paradigma tecnológico que proporciona el puente crucial entre el mundo continuo, sub-simbólico y basado en datos de las redes neuronales y el mundo discreto, estructurado y basado en la lógica del razonamiento y la planificación.

Es la tecnología que permite construir los módulos de "Sistema 2" o los "motores de razonamiento" que las arquitecturas de LeCun, Bengio y Goertzel exigen.

4.1. Principios Fundamentales de la IA Neurosimbólica (NeSy)

La IA neurosimbólica surge del reconocimiento de que los dos paradigmas históricos de la IA, el conexionismo (redes neuronales) y el simbolismo (lógica, reglas), son complementarios en sus fortalezas y debilidades.

- **Las redes neuronales** destacan en el aprendizaje de patrones a partir de datos brutos, no estructurados y ruidosos, pero son opacas ("cajas negras"), requieren grandes cantidades de datos y carecen de capacidades de razonamiento explícito y verificable.
- **La IA simbólica** sobresale en el razonamiento explícito, la planificación y la explicabilidad,

ya que opera sobre reglas y conocimiento formal. Sin embargo, es frágil ante datos ruidosos o incompletos y sufre del "cuello de botella de la adquisición de conocimiento", ya que las reglas a menudo deben ser codificadas manualmente por expertos.

NeSy busca combinar lo mejor de ambos mundos para crear sistemas que sean simultáneamente robustos en el aprendizaje y rigurosos en el razonamiento, más transparentes, eficientes en el uso de datos y capaces de realizar inferencias complejas de múltiples pasos.

4.2. Una Taxonomía de Arquitecturas NeSy

Para comprender las diversas formas en que se pueden integrar los componentes neuronales y simbólicos, la taxonomía propuesta por Henry Kautz es un marco de referencia influyente y ampliamente adoptado. Esta taxonomía clasifica las arquitecturas NeSy según el flujo de control y la naturaleza de su integración.

1. **Symbolic[Neuro]:** En esta arquitectura, el control principal es simbólico. Un sistema de razonamiento simbólico (como un motor de búsqueda o un planificador) invoca a una red neuronal como una subrutina para realizar una tarea específica que es difícil de formalizar con reglas, como la evaluación de patrones.
    - *Mecanismo de Interacción:* El componente simbólico pasa una consulta al componente neuronal y recibe una salida (p. ej., una puntuación, una clasificación) que luego utiliza en su proceso de razonamiento.
    - *Ejemplo Canónico:* **AlphaGo** y **AlphaGeometry**. En AlphaGo, un algoritmo de búsqueda de árbol Monte Carlo (MCTS, el componente simbólico) explora el árbol de juego. Para evaluar la calidad de una posición del tablero, llama a una red neuronal (el componente neuronal) que devuelve una estimación de la probabilidad de ganar desde esa posición. En AlphaGeometry, un motor de deducción simbólico, cuando se atasca, solicita a un modelo de lenguaje (el componente neuronal) que sugiera construcciones geométricas auxiliares prometedoras para avanzar en la prueba.

2. **Neuro|Symbolic:** Esta es una arquitectura de tipo pipeline o secuencial. Un componente neuronal actúa como un frontend de percepción, procesando datos brutos (p. ej., una imagen) y extrayendo una representación simbólica estructurada. Esta representación se pasa luego a un componente simbólico backend que realiza el razonamiento.
    - *Mecanismo de Interacción:* La red neuronal traduce la entrada no estructurada a un formato simbólico (p. ej., un grafo de escena, una fórmula lógica). El razonador simbólico opera exclusivamente sobre esta salida simbólica.
    - *Ejemplo Canónico:* **Respuesta a Preguntas Visuales (VQA)**. Un sistema VQA puede usar una CNN para detectar objetos en una imagen ("gato", "sofá") y sus relaciones espaciales ("está sobre"). Esta información se codifica en un grafo de escena. Luego, un razonador lógico puede responder a una pregunta como "¿De qué color es el objeto sobre el que está el gato?" consultando este grafo.

3. **Neuro:Symbolic→Neuro:** En este enfoque, el conocimiento simbólico se utiliza para guiar o estructurar el proceso de entrenamiento de una red neuronal. El componente simbólico no forma parte del sistema en tiempo de inferencia.
    - *Mecanismo de Interacción:* El conocimiento simbólico se "compila" en la arquitectura de la red o en los datos de entrenamiento. Por ejemplo, se puede usar un sistema de álgebra simbólica para generar millones de pares de problemas y soluciones para entrenar una red neuronal para que resuelva ecuaciones. O las reglas lógicas pueden usarse para generar etiquetas de datos adicionales o para imponer restricciones en la arquitectura de la red.

4. **NeuroSymbolic (Integración Profunda):** Esta es una de las formas más estrechas de integración. El conocimiento simbólico, típicamente en forma de lógica de primer orden, se integra directamente en el proceso de aprendizaje de la red neuronal, a menudo como un término de regularización en la función de pérdida.
    - *Mecanismo de Interacción:* Las reglas lógicas se traducen a un formato diferenciable. Durante el entrenamiento, la función de pérdida no solo mide el error en los datos etiquetados, sino también el grado en que las predicciones de la red violan las reglas lógicas. Esto obliga a la red a aprender representaciones (incrustaciones) que son consistentes con el conocimiento simbólico.
    - *Ejemplo Canónico:* **Logic Tensor Networks (LTN)**. En LTN, un axioma como \forall x (\text{gato}(x) \rightarrow \text{mamifero}(x)) se convierte en una pérdida que penaliza al modelo si la "veracidad" de que un objeto es un mamífero es menor que la "veracidad" de que es un gato.

5. **Neuro:** Esta arquitectura es la inversa de Symbolic[Neuro]. El control principal es neuronal. Un modelo neuronal, típicamente un agente de RL o un LLM, aprende a invocar a un motor de razonamiento simbólico externo como si fuera una herramienta.
    - *Mecanismo de Interacción:* El modelo neuronal genera una consulta a un sistema externo (una calculadora, un motor de búsqueda, una base de conocimiento) y luego incorpora la respuesta en su propio proceso de generación o toma de decisiones.
    - *Ejemplo Canónico:* Los **LLMs agenticos** que aprenden a usar herramientas. Un LLM al que se le pregunta "¿Cuál es la raíz cuadrada de la población de París?" puede aprender a generar una llamada a una API de búsqueda para encontrar la población y luego una llamada a una calculadora para realizar la operación matemática, en lugar de intentar adivinar la respuesta.

4.3. Marcos de Referencia e Implementaciones Clave de NeSy

Más allá de la taxonomía, varios marcos de software concretos han demostrado el poder de los enfoques NeSy en la práctica.

- **DeepProbLog:** Este marco integra la programación lógica probabilística (ProbLog) con el aprendizaje profundo a través del concepto de **predicado neural**. Un predicado neural

vincula un predicado lógico a una red neuronal. Por ejemplo, en la tarea de sumar dígitos de MNIST, se puede definir un predicado digit(Image, Number). La red neuronal toma una imagen Image y produce una distribución de probabilidad sobre los posibles números Number. Este hecho probabilístico puede ser utilizado dentro de un programa lógico más amplio. El ejemplo clásico es addition(I1, I2, Sum) :- digit(I1, N1), digit(I2, N2), Sum is N1 + N2. Aquí, dos redes neuronales (una para cada dígito) realizan la percepción, y el programa lógico realiza el razonamiento simbólico (la suma). Todo el sistema puede entrenarse de extremo a extremo, donde el error en la suma final se retropropaga para mejorar tanto las redes de clasificación de dígitos como los parámetros probabilísticos del programa lógico.

- **Logic Tensor Networks (LTN):** LTN proporciona un lenguaje, llamado **Real Logic**, para fundamentar la lógica de primer orden en espacios vectoriales continuos, haciéndola compatible con el aprendizaje profundo. En LTN, los símbolos lógicos se asignan a elementos del framework de deep learning:
    - Las **constantes** se convierten en tensores (vectores de características).
    - Los **predicados** se convierten en modelos neuronales (clasificadores) que toman tensores de entrada y producen un valor de verdad en el intervalo $$, utilizando semántica de lógica difusa.
    - Los **axiomas lógicos** se traducen en un término de pérdida diferenciable. El objetivo del entrenamiento es minimizar esta pérdida, lo que obliga al modelo a aprender incrustaciones para los conceptos (p. ej., "cisne", "blanco") que sean consistentes con las reglas lógicas.

- **Neural Logic Machines (NLM):** Mientras que DeepProbLog y LTN a menudo utilizan reglas lógicas predefinidas, las NLM son una arquitectura neuronal diseñada para el **aprendizaje inductivo de reglas "lifted"** (reglas que contienen variables) a partir de ejemplos. Las NLM operan sobre representaciones tensoriales de predicados y utilizan operadores neuronales que emulan las operaciones lógicas (AND, OR, NOT) y los cuantificadores (EXISTE, PARA TODO). Al entrenarse en tareas como ordenar una lista pequeña, una NLM puede aprender la regla simbólica subyacente del ordenamiento y luego generalizarla a listas mucho más largas, demostrando una capacidad de generalización sistemática que los modelos puramente neuronales no logran.

Estos marcos demuestran que la integración neurosimbólica no es solo una idea teórica.

Proporciona las herramientas de ingeniería de software y los fundamentos algorítmicos para construir los módulos de razonamiento de las arquitecturas cognitivas. Un módulo de percepción (como una JEPA) puede producir representaciones abstractas, que luego se convierten en la entrada para un motor de razonamiento NeSy.

Por ejemplo, el predicado digit(X, N1) en DeepProbLog es una interfaz Neuro|Symbolic perfecta: la red neuronal realiza la percepción (de la imagen a la distribución de probabilidad sobre los dígitos), y esto se convierte en un hecho simbólico para el programa lógico. NeSy es, por tanto, el detalle de implementación necesario para cualquier arquitectura AGI modular viable.

Marco NeSy	Principio Fundamental	Categoría de Kautz	Fortalezas	Limitaciones
DeepProbLog	**Predicados Neuronales:** Las redes neuronales definen las probabilidades de los hechos en un programa lógico probabilístico.	Neuro	Symbolic	Maneja de forma nativa la incertidumbre; integra el razonamiento lógico y probabilístico; entrenable de extremo a extremo.
Logic Tensor Networks (LTN)	**Lógica Diferenciable:** Los axiomas de la lógica de primer orden se convierten en una función de pérdida para regularizar el aprendizaje de incrustaciones.	NeuroSymbolic	Impone consistencia lógica en los espacios de incrustaciones; mejora la eficiencia de los datos al usar conocimiento previo; flexible.	La interpretación de los valores de verdad difusos puede ser menos intuitiva; la elección de los operadores difusos afecta el rendimiento.
Neural Logic Machines (NLM)	**Inducción de Reglas Lifted:** Arquitectura neuronal diseñada para aprender reglas lógicas abstractas (con variables) a partir de ejemplos de entrada/salida.	Neuro:Symbolic→Neuro (aprende reglas que luego podrían compilarse)	Demuestra una fuerte generalización sistemática a problemas de mayor escala; aprende la estructura lógica en lugar de tenerla predefinida.	La escalabilidad a reglas de alta aridad y lógicas complejas es un desafío; la extracción de reglas legibles por humanos a partir de los pesos es un problema abierto.
AlphaGeometry	**Búsqueda Simbólica Guiada por Red Neuronal:** Un motor de deducción simbólico utiliza un LLM para proponer construcciones auxiliares prometedoras.	Symbolic[Neuro]	Resuelve problemas de razonamiento extremadamente complejos que están fuera del alcance de los métodos puramente neuronales o simbólicos.	Arquitectura altamente especializada; requiere un motor simbólico específico del dominio.
Agentes LLM con Herramientas	**Invocación de Herramientas por LLM:** Un LLM aprende a generar llamadas a herramientas externas (calculadoras, APIs, bases de datos) para obtener información precisa.	Neuro	Reduce las alucinaciones al anclar las respuestas en fuentes de datos verificables; combina la flexibilidad del lenguaje con la precisión de los sistemas simbólicos.	Depende de la capacidad del LLM para aprender a usar las herramientas correctamente; la sobrecarga de la llamada a la API puede ser un problema.

Parte III: Un Plano para un Sistema de AGI Híbrido

El análisis de las limitaciones de los modelos actuales y la exploración de paradigmas alternativos convergen en una conclusión clara: el camino hacia la AGI requiere un diseño deliberado, basado en principios cognitivos.

No será un sistema monolítico, sino una arquitectura híbrida y modular. Esta parte final sintetiza los hallazgos anteriores para proponer un plano concreto para un sistema de este tipo, destacando sus pilares fundamentales y los desafíos que aún quedan por resolver.

Sección 5: Los Pilares Esenciales de una Inteligencia Generalizable

Cualquier arquitectura que aspire a la AGI debe construirse sobre tres pilares conceptuales que abordan directamente las deficiencias fundamentales de los modelos actuales. Estos pilares no son componentes opcionales, sino prerrequisitos para una inteligencia robusta y adaptable.

5.1. El Imperativo del Anclaje (*Grounding*)

El problema del anclaje de símbolos es quizás el obstáculo más fundamental para la IA actual. Para que los símbolos que un sistema manipula (ya sean palabras, tokens o nodos en un grafo) tengan un significado genuino, deben estar conectados causalmente con el mundo real. Un sistema entrenado únicamente con texto, por muy vasto que sea el corpus, permanece atrapado en el "carrusel de símbolo a símbolo", incapaz de comprender verdaderamente los referentes de su lenguaje.

La solución a este problema es el **anclaje multimodal**. Una AGI debe aprender de y sobre el mundo a través de múltiples modalidades sensoriales —visión, audición, tacto, etc.— además del lenguaje. La visión proporciona información sobre objetos, sus propiedades espaciales y sus interacciones; el sonido informa sobre eventos y sus dinámicas; el lenguaje proporciona una estructura simbólica para describir y razonar sobre estas percepciones. Es la correlación y la coherencia entre estas modalidades lo que permite que los símbolos se anclen en la experiencia perceptual. Un sistema que aprende que la palabra "gato", la imagen de un gato y el sonido "miau" co-ocurren consistentemente comienza a construir una representación multimodal y anclada del concepto *gato*, superando la superficialidad de los LLMs.

5.2. La Centralidad de los Modelos del Mundo

Una vez que un sistema tiene símbolos anclados, puede comenzar a construir un **modelo del mundo**. Este no es simplemente un almacén estático de hechos, sino una representación interna, dinámica y simulable del entorno y de sí mismo. Un modelo del mundo robusto es esencial para las capacidades cognitivas de alto nivel:

- **Predicción:** Permite al agente anticipar los estados futuros del mundo como consecuencia de eventos o de sus propias acciones. Esto es fundamental para cualquier forma de planificación.
- **Planificación:** Al utilizar el modelo del mundo como un simulador interno, el agente puede "imaginar" los resultados de diferentes secuencias de acciones sin tener que ejecutarlas

en el mundo real, lo que permite una toma de decisiones eficiente y segura.

- **Relleno de Información:** El modelo puede inferir información faltante o no observable. Si el agente ve humo, su modelo del mundo puede inferir la alta probabilidad de un fuego oculto.
- **Creatividad y Contrafácticos:** Un modelo del mundo permite al agente razonar sobre escenarios hipotéticos, una piedra angular de la creatividad y la resolución de problemas avanzada.

La arquitectura JEPA de LeCun está explícitamente diseñada para aprender estos modelos del mundo de manera eficiente, al predecir en un espacio de representación abstracto en lugar de en el espacio sensorial de alta dimensión.

5.3. La Primacía del Razonamiento Causal

Para que un modelo del mundo sea verdaderamente útil para una inteligencia general, debe ser **causal**, no meramente correlacional. Como argumenta Judea Pearl, la verdadera inteligencia reside en la capacidad de subir por la "Escalera de la Causalidad". Un agente debe ser capaz no solo de observar asociaciones (Peldaño 1), sino también de razonar sobre los efectos de las intervenciones (Peldaño 2) y de imaginar resultados contrafácticos (Peldaño 3).
Un modelo del mundo causal representa explícitamente las relaciones de causa y efecto, permitiendo al agente responder preguntas como "¿*Por qué* ocurrió esto?" y "¿Qué debo hacer para que ocurra aquello?". Esta capacidad es un prerrequisito para la planificación robusta, el diagnóstico de fallos y la toma de decisiones estratégicas.

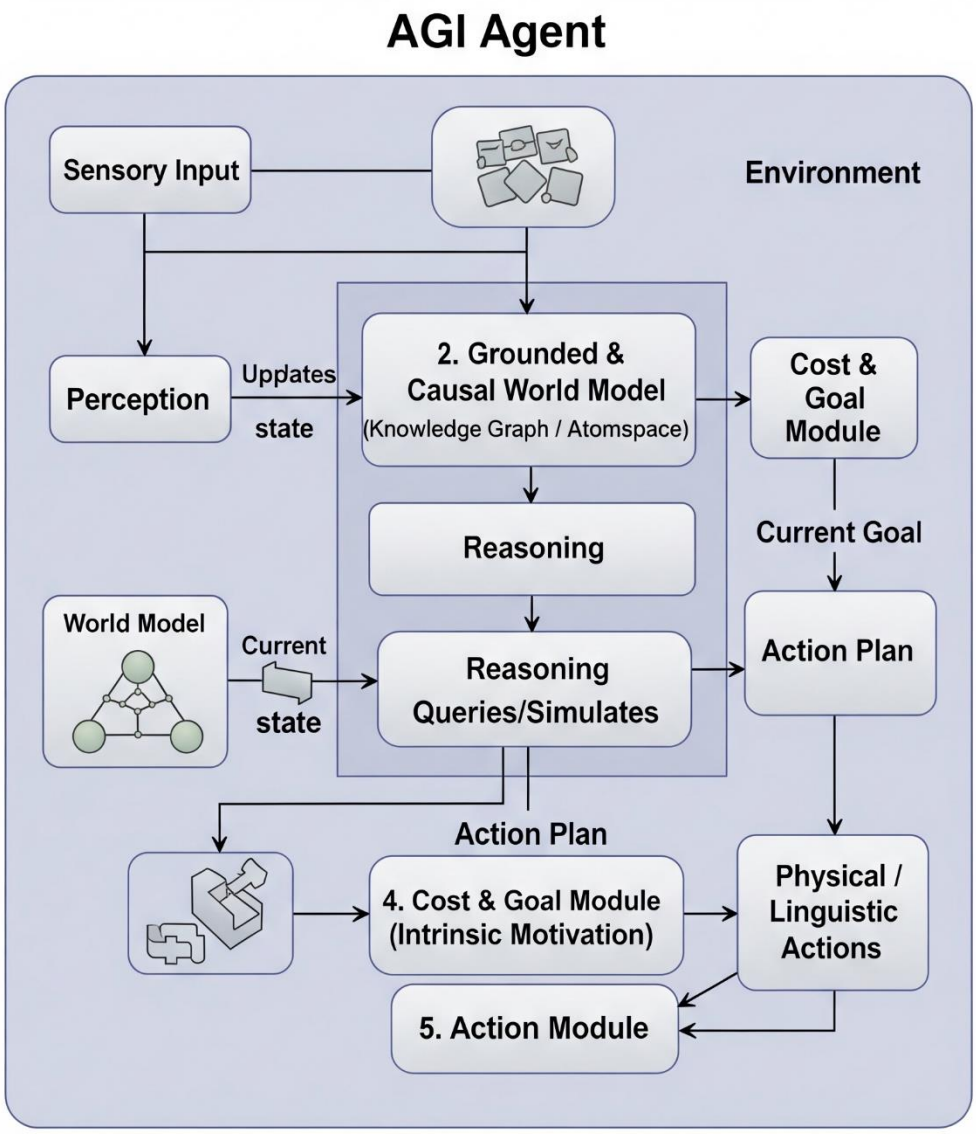
La **IA Causal Neurosimbólica** es el campo emergente dedicado a construir estos modelos. En estas arquitecturas, las relaciones causales se representan a menudo en una estructura simbólica (como un grafo de conocimiento causal o un programa lógico), mientras que los componentes neuronales se utilizan para aprender los parámetros de estos modelos causales a partir de datos de observación e intervención, o para realizar inferencias sobre ellos.
Estos tres pilares —anclaje, modelos del mundo y causalidad— no son independientes. El anclaje multimodal proporciona los datos brutos para aprender un modelo del mundo, y este modelo del mundo debe tener una estructura causal para permitir un razonamiento y una planificación verdaderamente inteligentes. Juntos, forman la base sobre la que se debe construir una arquitectura AGI modular.

Sección 6: Una Arquitectura Modular Propuesta para la AGI

Sintetizando los principios de los pioneros de la IA y las tecnologías habilitadoras de NeSy, podría proponerse un ESQUEMA tentativo para una arquitectura modular para la AGI. Esta arquitectura no es un modelo monolítico, sino un sistema de componentes especializados que interactúan en un ciclo cognitivo continuo.

Diagrama 1: El Ciclo Cognitivo de la AGI

El siguiente diagrama ilustra la arquitectura de alto nivel, mostrando los módulos principales y sus interacciones. Este diseño sería una síntesis (posible aproximada) de las propuestas de LeCun, Bengio y Goertzel:



Descripción de los Módulos:

1. **Percepción Multimodal (Sistema 1):** Este es el portal del agente hacia el mundo. Ingiere flujos de datos sensoriales brutos y no estructurados. Su función es transformar estos datos en representaciones abstractas, estructuradas y disentangled. Para ello, utilizaría arquitecturas de aprendizaje auto-supervisado como **JEPA** de LeCun para aprender representaciones predictivas , o **Redes de Cápsulas** de Hinton para capturar jerarquías parte-todo de forma robusta. La salida de este módulo no son etiquetas, sino representaciones vectoriales de alto nivel de los objetos, eventos y sus propiedades detectados.

2. **Modelo del Mundo Causal y Anclado:** Este es el núcleo cognitivo del agente, su base de "comprensión". Es una base de conocimiento dinámica y estructurada que almacena el estado del mundo. La implementación más probable sería un **Grafo de Conocimiento Causal** o un metagrafo como el **Atomspace** de OpenCog. Las representaciones abstractas del módulo de Percepción se utilizan para actualizar continuamente este modelo. Almacena no solo qué entidades existen y cuáles son sus propiedades (anclaje), sino también las reglas causales que gobiernan sus interacciones (causalidad).

3. **Motor de Razonamiento y Planificación NeSy (Sistema 2):** Este es el módulo de deliberación. Recibe un objetivo del Módulo de Coste y el estado actual del Modelo del Mundo. Su tarea es formular un plan para alcanzar el objetivo. Es intrínsecamente un sistema híbrido neurosimbólico que combina:
    - ***Razonamiento Simbólico****:* Utiliza un motor de inferencia lógico (como un probador de teoremas o un planificador simbólico) para generar planes de alto nivel y garantizar la corrección lógica.
    - ***Políticas Neuronales****:* Una red neuronal, entrenada mediante aprendizaje por refuerzo, guía la búsqueda a través del vasto espacio de posibles planes, podando ramas poco prometedoras, de forma análoga a la red de políticas de AlphaGo.
    - ***Simulación en el Modelo del Mundo****:* Antes de ejecutar un plan, este motor lo "simula" en el Modelo del Mundo para predecir sus consecuencias, permitiendo la corrección de errores y la optimización antes de actuar. Este módulo implementaría múltiples estrategias NeSy, como **Neuro** para consultar herramientas especializadas (p. ej., un solucionador matemático) o **Neuro|Symbolic** para razonar sobre los estados percibidos.

4. **Módulo de Coste y Objetivos:** Este es el sistema de motivación intrínseca del agente, como lo describe LeCun. No se basa en recompensas externas, sino en objetivos internos para guiar el comportamiento. Estos objetivos pueden ser de bajo nivel (p. ej., "evitar daño") o de alto nivel (p. ej., "minimizar la incertidumbre sobre el mundo", lo que crea un impulso de curiosidad, o "responder a la pregunta del usuario"). Es este módulo el que inicia el proceso de razonamiento al proporcionar un objetivo al motor de planificación.

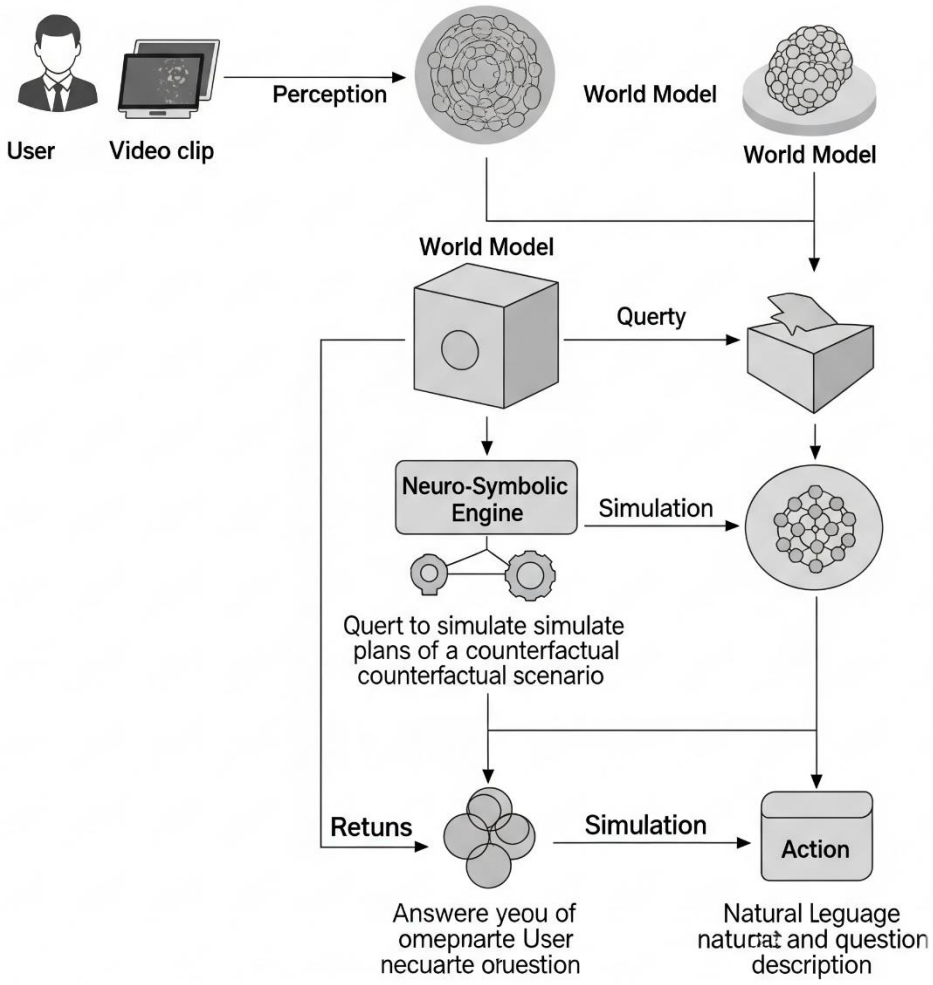
5. **Módulo de Acción:** Es el efector del agente. Traduce los planes abstractos generados por el Motor de Razonamiento en acciones concretas en el entorno. Estas acciones pueden ser de control motor para un robot o de generación de lenguaje para un agente conversacional. La ejecución de estas acciones modifica el entorno, lo que a su vez

genera nuevos datos sensoriales, cerrando así el ciclo percepción-acción.

Diagrama 2: Flujo de Información para el Razonamiento Causal Anclado

Para concretar cómo operaría esta arquitectura, consideremos el flujo de información para responder a una pregunta contrafáctica, una tarea que está fuera del alcance de los LLMs actuales.

Consulta: Un usuario muestra al agente un vídeo de una bola de billar roja golpeando una bola azul, y pregunta: *"¿Qué habría pasado si la bola roja hubiera sido más pesada?"*



Pasos del Flujo:

1. **Entrada y Percepción:** El Módulo de Percepción procesa tanto el vídeo como el texto. Identifica las entidades ("bola roja", "bola azul"), sus propiedades (color, movimiento) y los eventos (colisión).
2. **Anclaje y Actualización del Modelo del Mundo:** La información percibida se utiliza para instanciar y actualizar el Modelo del Mundo Causal. El modelo ahora contiene representaciones ancladas de las bolas y las reglas físicas (causales) de las colisiones.
3. **Formulación del Objetivo:** El Módulo de Coste/Objetivos interpreta la pregunta como una solicitud de inferencia contrafáctica, estableciendo el objetivo de "responder a la pregunta del usuario".
4. **Razonamiento y Simulación:** El Motor de Razonamiento NeSy recibe este objetivo. Crea una copia temporal del estado del mundo. En esta copia, realiza una **intervención**, modificando la propiedad masa de la entidad bolaRoja. Luego, utiliza las reglas causales del Modelo del Mundo para simular hacia adelante y predecir el nuevo resultado de la colisión.
5. **Generación de la Respuesta:** El resultado de la simulación (p. ej., un nuevo vector de velocidad para la bola azul) se pasa al Módulo de Acción, que lo traduce de nuevo a lenguaje natural para responder al usuario.

Una capacidad crucial que distingue a una verdadera AGI de un sistema estático es el **meta-aprendizaje** o la auto-organización.

La arquitectura descrita no debe ser fija. Un agente verdaderamente general debe aprender a orquestar sus propios recursos cognitivos.

Diferentes problemas requieren diferentes estrategias de razonamiento. Una pregunta simple podría resolverse directamente por el módulo de percepción (Sistema 1), mientras que un problema complejo de planificación requeriría una deliberación extensa por parte del motor de razonamiento (Sistema 2).

La AGI debería aprender a asignar dinámicamente sus recursos computacionales y a seleccionar la estrategia de razonamiento más adecuada para la tarea en cuestión. Esto implica un nivel superior de control, quizás un "Configurador" como el propuesto por LeCun , que optimiza no solo el conocimiento del agente sobre el mundo, sino también sus propios procesos de resolución de problemas. Este es un paso hacia una inteligencia que no solo aprende, sino que aprende a aprender.

Sección 7: Conclusión: Problemas Abiertos y Trayectorias Futuras

7.1. Recapitulación del Argumento

Este informe ha argumentado que la Inteligencia Artificial General no es un destino probable en el camino del escalado de modelos monolíticos como los LLMs. Las limitaciones de estos sistemas —su falta de anclaje, la ausencia de modelos del mundo coherentes, sus fallos en el razonamiento causal y su incapacidad para la generalización sistemática— no son defectos superficiales, sino fallas arquitectónicas fundamentales.

La alternativa más prometedora, respaldada por un consenso emergente entre los principales investigadores del campo, es el diseño deliberado de **arquitecturas cognitivas modulares**. Estas arquitecturas se inspiran en los principios de la cognición humana y animal, separando la percepción del razonamiento y centrando la inteligencia en un modelo del mundo interno, dinámico y causal. La tecnología clave para realizar los componentes de razonamiento de estas arquitecturas es la **IA Neurosimbólica**, que fusiona el aprendizaje robusto de las redes neuronales con el rigor y la explicabilidad del razonamiento simbólico.

El plano propuesto en este informe —un ciclo cognitivo que integra la percepción multimodal, un modelo del mundo causal, un motor de razonamiento neurosimbólico, un sistema de motivación intrínseca y un módulo de acción— representa una síntesis de estas ideas. Es un camino hacia una IA que no solo reconoce patrones, sino que comprende, razona y planifica.

7.2. Desafíos Clave sin Resolver

A pesar de la claridad de esta trayectoria, el camino hacia la AGI está plagado de desafíos de investigación fundamentales que deben ser abordados.

- **Aprendizaje Causal Escalable:** ¿Cómo puede un agente aprender de manera eficiente y robusta modelos causales del mundo a partir de la observación y la interacción, especialmente en entornos complejos y de alta dimensión? Los métodos actuales de descubrimiento causal a menudo no escalan bien o requieren suposiciones fuertes.
- **El Problema de la Representación:** ¿Cuál es el formato óptimo para el modelo del mundo? ¿Cómo podemos salvar de la manera más efectiva la brecha entre las representaciones continuas y distribuidas de las redes neuronales y las estructuras discretas y composicionales de la lógica simbólica? Aunque marcos como LTN y DeepProbLog ofrecen soluciones, la integración sigue siendo un área de investigación activa.
- **El Problema de la Función Objetivo:** ¿Cómo definimos las funciones de coste o los objetivos intrínsecos que conduzcan a un comportamiento seguro, alineado y beneficioso sin ser excesivamente restrictivos? Un agente verdaderamente autónomo debe ser capaz de desarrollar sus propios sub-objetivos, pero esto plantea profundas cuestiones de seguridad y alineación.
- **Integración NeSy Eficiente y Escalable:** Aunque existen muchos marcos NeSy, su integración puede ser computacionalmente costosa y compleja. El desarrollo de

sistemas híbridos más fluidos y eficientes, que no introduzcan cuellos de botella computacionales, es crítico para la viabilidad práctica de estas arquitecturas.

7.3. Observaciones Finales

El viaje hacia la Inteligencia Artificial General no es una carrera de velocidad basada en el escalado de un único paradigma, sino un programa de investigación complejo e interdisciplinario.

Requiere la síntesis de ideas del aprendizaje profundo, la IA simbólica, la ciencia cognitiva, la causalidad y la filosofía. Las arquitecturas híbridas, modulares y de inspiración cognitiva, impulsadas por la integración neurosimbólica, representan la frontera más prometedora de este esfuerzo.

Construir máquinas que no solo calculen, sino que piensen, sigue siendo el mayor desafío de nuestro tiempo, y es en esta síntesis de aprendizaje y razonamiento donde reside la esperanza más fundada de lograrlo.

7.3. Conclusión: ¿La Piedra Rosetta de la IA?

De todo lo anterior se desprende que las arquitecturas NeuroSimbólicas van a estar en el centro de todos los intentos de crear una AGI, o por lo menos sistemas de AI mejores que los actuales, en el futuro previsible 2025-2035.

Parece que el primer paso, de una compleja y larga secuencia, va a ser integrar la intuición de los LLMs con los lenguajes formales y la lógica, para poder conseguir capacidad de razonamiento potente y suficiente.

Si esto es así, entonces la “lógica” reinventada y redefinida como “ingenieria inversa del lenguaje universal que usamos para hacer discursos, razonamiento y computación” será un campo fundamental y central de investigación para encontrar la solución.

Y en tal caso, un lenguaje como ULOGIC (y sus mejoras futuras) podría ser la Piedra Rosetta que permita sentar los cimientos de todo el edificio.

En palabras de Arquímedes de Siracusa: **δῶς μοι πᾶ στῶ καὶ τὰν γᾶν κινάσω** (“dadme un punto de apoyo y moveré el mundo”)

REFERENCIAS

1. Arquitecturas Cognitivas y Modelos del Mundo (Visión de LeCun)

- **Fuente:** LeCun, Y. (2022). *A Path Towards Autonomous Machine Intelligence*. OpenReview.
    - **URL:** https://openreview.net/pdf?id=BZ5a1r-kVsf
    - **Comentario:** This is the foundational paper by Yann LeCun that outlines his vision for a modular, cognitive architecture for AGI, centered on a predictive world model. It details the roles of the Perception, World Model, Cost, and Actor modules, arguing that this is the path to move beyond the limitations of current models. It's the primary source for the architecture proposed in Section 3.1.
- **Fuente:** Assran, M., et al. (2023). *Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture*. arXiv.
    - **URL:** https://arxiv.org/pdf/2301.08243
    - **Comentario:** This paper provides the technical details for the Joint-Embedding Predictive Architecture (JEPA), which LeCun proposes as the key technology for building the World Model component. It explains how JEPA works in an abstract representation space to make predictions, which is more efficient than predicting every pixel.

2. Cognición de Sistema 2 y Razonamiento (Visión de Bengio)

- **Fuente:** Bengio, Y., et al. (2019). *The Consciousness Prior*. arXiv.
    - **URL:** https://arxiv.org/pdf/1709.08568
    - **Comentario:** This is a key paper by Yoshua Bengio introducing the "Consciousness Prior". It formalizes the idea that high-level, conscious thoughts involve a sparse combination of concepts, which maps to a sparse factor graph.
- **Fuente:** Bengio, Y., et al. (2021). *GFlowNets for AI-Driven Scientific Discovery*. arXiv.
    - **URL:** https://arxiv.org/pdf/2307.13524
    - **Comentario:** This paper details Generative Flow Networks (GFlowNets), the mechanism proposed by Bengio's lab to perform structured search and sampling. It's the technical answer to How a System 2 module might explore the vast combinatorial space of reasoning. It's designed to generate compositional objects like graphs or explanations.

3. Limitaciones de los LLMs: Causalidad y Grounding

- **Fuente:** Kıcıman, E., et al. (2023). *Causal Reasoning and Large Language Models: A Survey*. arXiv.
    - **URL:** https://arxiv.org/pdf/2305.00050
    - **Comentario:** This is a comprehensive survey about the causal reasoning failures of LLMs. It uses Judea Pearl's "Ladder of Causation" as a framework to show that LLMs are stuck at the level of association and fail at intervention and counterfactuals

- **Fuente:** Bender, E. M., & Koller, A. (2020). *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
    - **URL:** https://aclanthology.org/2020.acl-main.463.pdf
    - **Comentario:** This is a highly influential paper that provides a rigorous linguistic and philosophical argument for the symbol grounding problem in LLMs. It argues that models trained only on text (form) cannot, by definition, learn meaning in the way humans do. This paper is a cornerstone for the arguments you make about the "symbol-to-symbol carousel" and the different levels of meaning.
- **Fuente:** Lake, B. M., & Baroni, M. (2023). *Human-like systematic generalization through a meta-learning neural network*. Nature.
    - **URL:** https://www.nature.com/articles/s41586-023-06668-3
    - **Comentario:** This paper directly addresses the lack of systematic generalization and compositionality in standard neural networks. It presents an architecture that learns to generalize in a more human-like way, confirming that this is a recognized and critical limitation of current approaches and a major area of active research.

4. Arquitecturas Neurosimbólicas (NeSy) y Frameworks

- **Fuente:** d'Avila Garcez, A. S., et al. (2022). *Neurosymbolic AI: The 3rd Wave*.
    - **URL:** https://www.scitepress.org/Papers/2022/107775/107775.pdf
    - **Comentario:** A great high-level overview of the field of Neurosymbolic AI. It positions NeSy as the "Third Wave" of AI, following handcrafted rules and statistical learning. It validates the central thesis that combining neural and symbolic approaches is the most promising path forward.
- **Fuente:** Kautz, H. (2022). *The Third AI Summer, and its Contradictions*. AI Magazine.
    - **URL:** https://ojs.aaai.org/index.php/aimagazine/article/view/19485/19253
    - **Comentario:** This is the source for the influential taxonomy of Neuro-Symbolic architectures. It provides the definitions for Symbolic[Neuro], Neuro|Symbolic, etc., and serves as a formal framework for classifying the different integration strategies discussed.
- **Fuente:** De Raedt, L., et al. (2020). *From Statistical Relational AI to Neuro-Symbolic AI*.
    - **URL:** https://arxiv.org/pdf/2003.08316
    - **Comentario:** This paper provides a detailed look into the principles behind systems like DeepProbLog. It explains the concept of the "neural predicate" and how probabilistic logic programming can be integrated with deep learning.

5. Marcos Integradores de AGI (Visión de Goertzel)

- **Fuente:** Goertzel, B., et al. (2023). *The OpenCog Hyperon AGI Architecture*. arXiv.
    - **URL:** https://arxiv.org/pdf/2310.18318
    - **Comentario:** This is the primary technical paper describing the OpenCog Hyperon framework. It details the two core components: the Atomspace as a flexible knowledge metagraph and the MeTTa programming language designed for cognitive synergy. It is the definitive source for the ideas presented in Section 3.4, advocating for an integrative, multi-algorithm approach to AGI.