CAPÍTULO IV

Agentic AI y Multiagentic: ¿Estamos reinventando la rueda? Una reinterpretación epistemológica

Vicent Botti

Los términos "Agentic AI (IA agentiva)" y "Multiagentic AI (IA multiagentiva)" han ganado popularidad recientemente en los debates sobre inteligencia artificial generativa, a menudo utilizados para describir agentes de *software* autónomos y sistemas compuestos por dichos agentes. Sin embargo, estos conceptos no representan una nueva ontología de agentes, sino una reconfiguración tecnológica de ideas ya consolidadas en la teoría de agentes autónomos y sistemas multiagente (MAS) desde hace muchos años. El uso de estos términos confunde estos conceptos de moda con conceptos bien establecidos en la literatura de la IA: agentes autónomos y sistemas multiagente. El capítulo aboga por el rigor científico y tecnológico y por el uso de la terminología establecida del estado del arte en IA, incorporando la riqueza del conocimiento existente —incluidos los estándares para plataformas de sistemas multiagente, lenguajes de comunicación y algoritmos de coordinación/cooperación, tecnologías de acuerdo (negociación automática, argumentación, organizaciones virtuales, confianza, reputación, etc.)— en la nueva y prometedora ola de agentes de IA basados en LLM, para no terminar reinventando la rueda.

Palabras clave: inteligencia artificial, agentes autónomos, sistemas multiagente, IA Generativa, LLMs.

1. INTRODUCCIÓN

El campo de los sistemas multiagente comenzó a tomar forma a principios de la década de 1990, inspirado por la idea de que el futuro de la inteligencia artificial implicaría redes de entidades de IA autónomas —conocidas como agentes— que actuarían de forma independiente para cumplir las tareas asignadas por los usuarios. Un elemento central de esta visión era la expectativa de que estos agentes necesitarían interactuar entre sí para alcanzar sus objetivos, lo que impulsó una investigación significativa para dotarlos de capacidades sociales como la cooperación, la coordinación, la negociación, la argumentación, la confianza y la reputación. Este capítulo ofrece un análisis crítico de este mal uso conceptual del término Agentic (agentiva) y defiende la invariancia conceptual del paradigma de agentes autónomos frente a la evolución tecnológica, examina la continuidad conceptual entre los agentes autónomos clásicos y las nuevas arquitecturas basadas en los LLMs, integrando las aportaciones epistemológicas y neurosimbólicas. Revisamos los orígenes teóricos de lo "agentic (agentiva)" en las ciencias sociales (Bandura, 1986) y las nociones filosóficas de intencionalidad (Dennett, 1971), y luego resumimos los trabajos fundamentales sobre agentes inteligentes y sistemas multiagente de Wooldridge, Jennings y otros. Examinamos las arquitecturas clásicas de agentes —desde simples agentes reactivos hasta modelos de Creencia-Deseo-Intención (BDI)— y destacamos las propiedades clave (autonomía, reactividad, proactividad, capacidad social) que definen la agencia en la IA. A continuación, discutimos los desarrollos recientes en los grandes modelos de lenguaje (LLM) y las plataformas de agentes basadas en LLM, incluida la aparición de agentes de IA basados en LLM y los marcos de orquestación multiagente de código abierto. Argumentamos que el término "Agentic AI (IA agentiva)" se utiliza a menudo como un término de moda para lo que son esencialmente agentes de IA, y "Multiagentic AI (IA multiagentiva)" para lo que son sistemas multiagente. Esta confusión pasa por alto décadas de investigación en el campo de los agentes autónomos/sistemas multiagente. Se sostiene que la IA agentiva no redefine la agencia, sino que la reinterpreta en clave cognitivo-lingüística, manteniendo la estructura funcional de percepción, decisión y acción de la abstracción de agente definida en el área de los agentes autónomos/sistemas multiagente. Finalmente, se discuten los riesgos de la inflación terminológica y las oportunidades de una síntesis neurosimbólica para el futuro de la inteligencia artificial. El capítulo aboga por el rigor científico y tecnológico y por el uso de la terminología establecida del estado del arte en IA, incorporando la riqueza del conocimiento existente —incluidos los estándares para plataformas de sistemas multiagente, lenguajes de comunicación y algoritmos de coordinación/cooperación, tecnologías de acuerdo (negociación automatizada, argumentación, organizaciones virtuales, confianza, reputación, etc.)— en la nueva y prometedora ola de agentes de IA basados en LLM, para no terminar reinventando la rueda.

Los recientes avances en la IA generativa, en particular los grandes modelos de lenguaje (LLM), han llevado a un resurgimiento del interés en los agentes de *software* autóno-

mos: sistemas de IA que pueden percibir de forma autónoma su entorno, razonar y actuar para cumplir sus objetivos de diseño y realizar tareas en nombre de los usuarios. Existe un creciente entusiasmo en torno al desarrollo de agentes basados en LLM que aprovechan su inteligencia general para operar de forma autónoma. Visionarios como Bill Gates predijeron que, en un futuro cercano, todos tendremos un asistente personal de IA "muy superior a la tecnología actual", capaz de responder a solicitudes en lenguaje natural y realizar diversas tareas comprendiendo los objetivos del usuario (Gates, 2023). Gates señala que este *software*, que él y otros han imaginado durante décadas, por fin se está convirtiendo en algo práctico gracias a los avances en IA, y que tales "agentes" podrían revolucionar nuestra interacción con los ordenadores. Además, hay una creciente evidencia de que organizar sistemas como conjuntos de agentes especializados e interactivos puede ser un enfoque eficaz para abordar tareas complejas de resolución de problemas. Paralelamente, cierto discurso industrial, no siempre (hay muchos casos en los que la terminología se respeta), ha introducido nuevos términos como "Agentic AI (IA agentiva)" para describir sistemas de IA dotados de autonomía y toma de decisiones proactiva. Blogs y artículos técnicos diferencian entre "agentes de IA" y "Agentic AI (IA agentiva)", presentando esta última como un marco general donde operan múltiples agentes. Los sistemas basados en los LLM, capaces de generar lenguaje y planificar acciones, han sido denominados "agentes agentivos" (una tautología que debilita la precisión científica), sugiriendo un salto cualitativo en autonomía y razonamiento. De manera similar, el término "Multiagentic (multiagentiva)" se utiliza a veces para sistemas con múltiples agentes que interactúan, de forma análoga a los sistemas multiagente clásicos.

Esta aparición de la terminología "agentic" en la IA plantea preguntas críticas. ¿Son estos conceptos genuinamente nuevos o simplemente nuevas etiquetas para ideas ya consolidadas en la investigación de agentes autónomos? El campo de los agentes inteligentes y los sistemas multiagente (MAS) tiene una rica historia que se remonta a décadas, con sus propios fundamentos teóricos, arquitecturas e incluso conferencias, revistas y asociaciones científicas dedicadas a investigadores en el área. Antes de abrazar el hype de la "Agentic AI", es importante examinar si el término se está aplicando erróneamente como una palabra de moda para capacidades comprendidas desde hace mucho tiempo en la IA. En otras palabras, ¿estamos reinventando la rueda al equiparar la Agentic IA con los agentes inteligentes y los sistemas multiagentic con los sistemas multiagente? La llamada IA agentiva constituye una continuidad conceptual del paradigma de agentes autónomos, aunque potenciada por nuevas tecnologías, la incorporación de la tecnología de los LLM como medio para que un agente determine, a partir de sus percepciones, qué acciones desempeñar para alcanzar los objetivos que le hayan sido delegados. Por tanto, la discusión actual no gira en torno a la definición del concepto de agente, sino a la reinterpretación de su implementación en contextos neurosimbólicos y lingüísticos.

2. LOS ORÍGENES DE AGENTIC: AGENCIA EN PSICOLOGÍA Y FILOSOFÍA

El adjetivo "agentic" ganó prominencia formal fuera de la informática. La primera referencia al término "agentic" aparece en el contexto de la psicología social, en el trabajo de psicólogos como Richard H. Goffman y Susan Fiske en la década de 1970, aunque sus raíces provienen de conceptos anteriores relacionados con la agencia y la autonomía. El término proviene de "agencia" y se refiere a la capacidad de los individuos para actuar de forma autónoma, tomar decisiones y ejercer control sobre sus vidas. El concepto de agencia ha existido durante mucho más tiempo, pero el término "agentic" como adjetivo se popularizó en la literatura académica especialmente a través del trabajo de Albert Bandura. En su libro *Social Foundations of Thought and Action: A Social Cognitive Theory* (1986), Albert Bandura introdujo el término agentic en el contexto de la agencia humana (Bandura, 1986). En la teoría social cognitiva de Bandura, agentic se refiere a la capacidad de los individuos para actuar intencionadamente, tomar decisiones y ejercer control sobre su entorno. Esto enfatiza un aspecto proactivo y deliberativo del comportamiento humano: en resumen, la cualidad de tener agencia. Una persona con capacidad agentic (agencia) no solo reacciona a fuerzas externas; son originadores de acciones, persiguiendo objetivos por su propia voluntad. La introducción del término por parte de Bandura destacó la naturaleza proactiva y autorreguladora de la acción humana, subrayando que las personas son agentes de su propio cambio. Este concepto de agencia ha sido influyente en la psicología, sustentando teorías de autoeficacia y motivación, pero originalmente se aplicó a humanos, no a máquinas.

Si buscamos en el diccionario, encontramos el siguiente significado de agentic:

- "Agentic" es un adjetivo derivado del sustantivo *agency*, y se usa comúnmente en psicología, sociología y educación. Se refiere a tener la capacidad de actuar de forma independiente, tomar decisiones y ejercer control sobre la propia vida y circunstancias.

- Definición (en contexto):

  - Que exhibe agencia; capaz de acción intencional y toma de decisiones.

  - Que actúa como un agente (es decir, iniciando y dirigiendo el propio comportamiento).

En paralelo, filósofos como Daniel C. Dennett exploraban cómo las atribuciones de agencia e intencionalidad podían aplicarse a máquinas y otros sistemas no humanos. El artículo de Dennett de 1971, *Sistemas Intencionales*, articuló la idea de que a menudo interpretamos y predecimos el comportamiento de sistemas complejos tratándolos como si tuvieran creencias, deseos e intenciones (Dennett, 1971). Dennett llamó sis-

tema intencional a aquel cuyo comportamiento puede explicarse y predecirse con éxito atribuyéndole cualidades mentalistas como creencias y deseos, de forma similar a la psicología popular que usamos para los humanos (Dennett, 1971). Esto proporciona una base filosófica para hablar de agentes de IA: incluso si un termostato o un programa no tienen literalmente deseos o creencias, tratarlo como un agente intencional (con objetivos, conocimiento, etc.) puede ser una abstracción útil para comprender su comportamiento (Dennett, 1971). La teoría de Dennett tendió así un puente entre la psicología humana y las entidades artificiales, sugiriendo que cualquier sistema cuyo comportamiento sea suficientemente complejo y dirigido puede ser visto en términos intencionales (Dennett, 1971).

Las nociones de Bandura (Bandura, 1986) y Dennett (Dennett, 1971) proporcionan un telón de fondo para la agencia como concepto: la agencia implica una acción intencional y dirigida a un objetivo, ya sea en humanos o en sistemas artificiales. Cuando llamamos a un sistema de IA agentivo, en el sentido estricto, queremos decir que exhibe agencia, es decir, que puede actuar de forma autónoma con la intención de alcanzar objetivos. En la psicología humana, esto es incontrovertido, pero en la IA plantea la pregunta: ¿qué se necesita para que un sistema artificial sea considerado un agente? El campo de la IA ha respondido a esto a lo largo de los años definiendo la abstracción del agente inteligente, que examinaremos a continuación. Baste decir que, para cuando el término agentic comenzó a aplicarse a la IA en la década de 2020, la comunidad ya tenía un profundo conocimiento de lo que significa que un sistema tenga agencia.

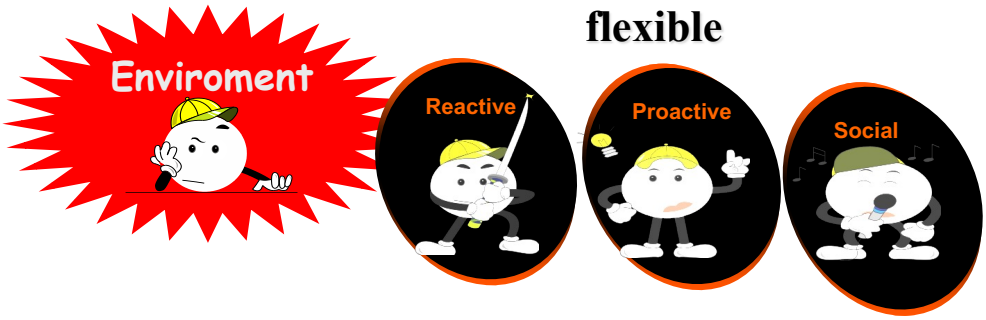
3. AGENTES INTELIGENTES: DEFINICIONES Y PROPIEDADES CLAVE

En 1991, Yoav Shoham propuso un nuevo paradigma de programación, la "Programación Orientada a Agentes" (Shoham, 1993). El paradigma promueve una visión social de la computación, en la que múltiples agentes interactúan entre sí.

En inteligencia artificial, un agente inteligente es fundamentalmente una entidad autónoma que percibe su entorno y actúa sobre él para alcanzar sus objetivos. Una definición clásica de Wooldridge y Jennings (1995) describe un agente inteligente como "un sistema informático situado en un entorno, capaz de actuar de forma flexible y autónoma para cumplir sus objetivos de diseño" (figura 1). Esto abarca varios aspectos importantes. El agente está situado en un mundo (físico o virtual) donde recibe entradas (percepciones) a través de sensores e influye en el mundo a través de actuadores. La autonomía implica que opera sin intervención humana directa, tomando sus propias decisiones sobre qué acciones realizar (Wooldridge, 2002). La acción flexible significa que el agente puede adaptarse a condiciones cambiantes y perseguir sus objetivos de manera robusta.

FIGURA 1

DEFINICIÓN DEL AGENTE DE WOOLDRIDGE Y JENNINGS



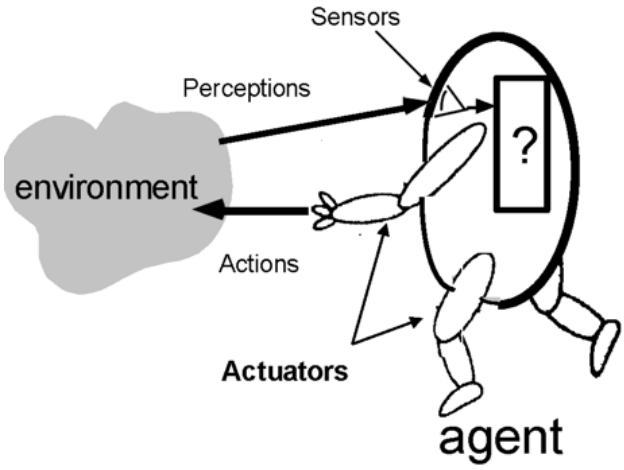
Fuente: Elaboración propia.

Wooldridge amplió que la acción flexible e inteligente implica propiedades como la reactividad, la proactividad y la capacidad social (Wooldridge & Jennings, 1995). Estas, junto con la autonomía, a menudo se citan como propiedades clave de los agentes inteligentes (Jennings *et al.,* 1998). La autonomía significa que los agentes funcionan de forma independiente, tomando decisiones y ejecutando acciones sin necesidad de supervisión humana directa; esta autonomía les permite realizar tareas y adaptarse en tiempo real a entornos cambiantes. La reactividad implica que el agente percibe su entorno y responde a los cambios de manera oportuna. La proactividad significa que puede tomar la iniciativa —un comportamiento orientado a objetivos— no solo reaccionando al entorno, sino también haciendo acciones dirigidas a objetivos para cumplir sus metas. La capacidad social se refiere a la habilidad de un agente para interactuar con otros agentes (o humanos), comunicándose, cooperando o compitiendo según sea necesario. Estos atributos distinguen a un agente inteligente de un mero programa que actúa de manera predefinida. Es importante destacar que estos conceptos se establecieron a mediados de la década de 1990, lo que subraya que el concepto de agentes de IA no es nuevo; los investigadores han estudiado durante mucho tiempo qué hace que un agente sea "inteligente".

Para ilustrar estas ideas, podemos visualizar un modelo simple de un agente inteligente interactuando con su entorno (figura 2). El agente recibe percepciones del entorno a través de sensores, procesa estas entradas (utilizando su razonamiento interno o mecanismo de toma de decisiones) y luego produce acciones a través de actuadores para cambiar el entorno. El ciclo percibir-pensar-actuar es el núcleo de cualquier agente. Por ejemplo, un agente termostato percibe la temperatura actual, "decide" si coincide con

FIGURA 2

MODELO SIMPLE DE UN AGENTE INTERACTUANDO CON SU ENTORNO



Fuente: Elaboración propia.

el punto de ajuste deseado y actúa encendiendo o apagando la calefacción. Un ejemplo más complejo es un agente robótico: podría percibir el mundo a través de cámaras y sensores de distancia, razonar sobre sus objetivos (por ejemplo, navegar a una ubicación) y actuar sobre su sistema de propulsión. En todos los casos, la noción de agencia implica que el sistema vincula continuamente la percepción con la acción en busca de objetivos.

El cuadro con el signo de interrogación en la figura 2 es donde incorporaremos la inteligencia del agente, es decir, aquellas técnicas de IA que nos permiten determinar, a partir de las percepciones del agente, qué acciones realizar en el entorno para alcanzar sus objetivos. Estas técnicas nos permitirán interpretar sus percepciones, deliberar para determinar qué objetivos cumplir, razonar para determinar qué acciones le permitirán alcanzar sus objetivos, etc. Esto implica el uso de técnicas de reconocimiento de formas, procesamiento del lenguaje natural, planificación, los LLM, etc.

El agente se concibe como una abstracción tecnológica independiente del mecanismo de razonamiento: puede basarse en lógica simbólica, planificación, sistemas basados en reglas, sistemas basados en casos, aprendizaje, redes neuronales, etc. Por tanto, los agentes actuales basados en LLM no rompen con este modelo, sino que lo amplían mediante nuevas formas de inferencia lingüística y contextual basadas en la tecnología de los LLM. La IA agentiva no introduce una nueva categoría ontológica, sino que reinterpreta la noción clásica de agente bajo un nuevo marco de implementación, centrado en la generación y comprensión del lenguaje natural. Los nuevos agentes basados en los LLM incorporan una técnica conocida como cadena

de pensamiento (CdP) razonamiento, para mejorar sus habilidades de resolución de problemas. Este método alienta a los modelos a dividir los problemas en pasos intermedios, imitando la forma en que los humanos piensan en un problema de manera lógica esto es algo que ya se propuso en la inteligencia artificial distribuida. Esto no supone un razonamiento genuino, como lo entendemos en IA; se asemeja más a un proceso de resolución de problemas de búsqueda en un espacio de soluciones, una técnica clásica de la IA.

Vale la pena señalar que el libro de texto de IA de Russell y Norvig (2020) identifica esta interacción agente-entorno como el concepto fundamental de la IA, incluso definiendo la IA misma como el estudio de los agentes inteligentes. A finales de la década de 1990, el paradigma de agente se había vuelto central en la investigación de la IA, dando lugar a subcampos como los agentes autónomos y los sistemas multiagente. Esto llevó a la creación de espacios dedicados como el International Workshop on Agent Theories, Architectures, and Languages (ATAL), la International Conference on Autonomous Agents (AGENTS), la International Conference on Multi-Agent Systems (ICMAS) y su fusión en la Autonomous Agents and Multiagent Systems Conference (AAMAS) a partir de 2002. Sociedades profesionales como The International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS) y The European Association for Multi-Agent Systems (EURAMAS), The Foundation for Intelligent Physical Agents (FIPA) y revistas como Autonomous Agents and Multi-Agent Systems (Springer) se establecieron para avanzar en la teoría y práctica de los agentes. El área de agentes y sistemas multiagente ha estado presente durante muchos años en conferencias de inteligencia artificial, en particular en las más prestigiosas como IJCAI (International Joint Conference on Artificial Intelligence), AAAI (Association for the Advancement of Artificial Intelligence) y ECAI (European Conference on AI). En resumen, para cuando los blogs tecnológicos actuales comenzaron a hablar de "Agentic AI", la comunidad de investigación de IA ya había construido un campo vibrante y en evolución en torno a estos conceptos.

La literatura distingue entre tres niveles analíticos (Ferber, 1999): conceptual, arquitectónico y tecnológico. El nivel conceptual define las propiedades esenciales del agente; el arquitectónico describe su estructura interna (reactiva, deliberativa o híbrida); y el tecnológico implementa dichas funciones mediante diversas técnicas. La IA agentiva opera principalmente en este último nivel, reemplazando/complementando el razonamiento simbólico o el razonamiento práctico (Bratman, 1999 [1987]) por modelos de lenguaje que permiten deliberación en lenguaje natural y adaptabilidad contextual. El nivel conceptual sigue siendo el mismo, aunque el soporte computacional evolucione.

4. SISTEMAS MULTIAGENTE

Aunque un solo agente inteligente puede ser poderoso, muchos problemas del mundo real requieren múltiples agentes que cooperen (o compitan) para ser resueltos. Un sis-

tema multiagente consta de múltiples agentes autónomos que interactúan —ya sea de forma colaborativa o competitiva— para alcanzar objetivos individuales o colectivos (Wooldridge, 2002). En tales sistemas, ningún agente individual tiene capacidades suficientes para lograr los objetivos generales por sí solo; esta interacción puede implicar comunicación, negociación, cooperación en subtareas o competencia por recursos compartidos, dependiendo del escenario.

A principios de la década de 2000, los SMA habían madurado como un área de investigación distinta, abordando los desafíos de coordinación, comunicación y organización de múltiples agentes. Los investigadores desarrollaron estándares y lenguajes para permitir la interoperabilidad entre agentes de diferentes desarrolladores. Por ejemplo, la Fundación para Agentes Físicos Inteligentes (FIPA) comenzó a emitir estándares para protocolos de comunicación de agentes en 1996 (FIPA, 1996).

Esta área de investigación se basa en un nuevo paradigma de computación, la computación como interacción (Shoham, 1993). En este paradigma, la computación ocurre a través de la comunicación entre entidades computacionales; la computación es una actividad inherentemente social, no solitaria, lo que da lugar a nuevas formas de concebir, diseñar, desarrollar y gestionar sistemas computacionales.

Cuando hablamos de comunicación, no podemos olvidar iniciativas como el Esfuerzo de Compartición de Conocimiento (*Knowledge Sharing Effort*). Esta iniciativa fue lanzada alrededor de 1990 por ARPA (la Asociación de Proyectos de Investigación Avanzada del Departamento de Defensa de EE. UU.), con el apoyo de organizaciones de investigación como ASOFR, NSF y NRI. Su objetivo principal era promover la interoperabilidad en la comunicación de conocimiento entre sistemas inteligentes.

Como resultado de este proyecto, se concluyó que la comunicación efectiva entre agentes distribuidos requiere tres niveles de lenguaje:

1. *Sintaxis*: Se requiere un lenguaje para representar el conocimiento de manera formal y estructurada.

2. *Semántica*: Se requiere un lenguaje para definir ontologías, lo que permite describir el significado del conocimiento representado.

3. *Pragmática*: Se requiere un lenguaje para la comunicación entre agentes, centrándose en cómo se intercambian y manipulan los mensajes de conocimiento.

El diseño de estos lenguajes de comunicación de agentes se inspiró en teorías lingüísticas sobre los actos de habla (Austin, 1962; Searle, 1969):

- J. L. Austin (1962) – Cómo hacer cosas con palabras
- J. R. Searle (1969) – Actos de habla

Estos trabajos sentaron las bases para comprender la pragmática de la comunicación, es decir, cómo se utilizan las palabras para realizar acciones.

Los primeros lenguajes de comunicación de agentes se basaron en la teoría de los actos de habla de la filosofía del lenguaje. Concretamente, se propusieron lenguajes como KQML (*Knowledge Query and Manipulation Language*) (Finin *et al.,* 1994) y FIPA-ACL (FIPA, 1996) para estandarizar cómo los agentes hacen preguntas, se informan mutuamente o negocian. También se desarrollaron formatos comunes de representación del conocimiento, como KIF (*Knowledge Interchange Format*) (Genesereth & Fikes, 1992) para la sintaxis del contenido y lenguajes como ONTOLINGUA (Ontolingua, s.f.) para definir ontologías compartidas para la semántica. Estos esfuerzos fueron el intento de la comunidad de SMA de garantizar que agentes heterogéneos pudieran trabajar juntos, un reconocimiento de que la capacidad social (una de las propiedades de los agentes) requería métodos de comunicación estandarizados.

En los sistemas multiagente, los agentes que los constituyen deben interactuar para resolver el problema para el que fueron diseñados; estas interacciones pueden ser colaborativas, cuando los agentes son benevolentes, o competitivas, cuando los agentes son egoístas. Esto se traduce en la necesidad de métodos que permitan a los agentes llegar a acuerdos o alinear sus intenciones, lo que lleva a conceptos como consenso distribuido, trabajo en equipo o subastas para la asignación de tareas (*Agreement Computing)* (Sierra *et al.*, 2011).

En respuesta a esta necesidad de alcanzar acuerdos, surgieron las tecnologías de acuerdo (*Agreement Technologies*, AT) (Ossowski *et al.*, 2012). Estas se refieren a sistemas informáticos en los que agentes de *software* autónomos negocian entre sí, generalmente en nombre de personas, para llegar a acuerdos mutuamente aceptables. La autonomía, la interacción, la movilidad y la apertura son características relevantes desde una perspectiva teórica y práctica.

En Europa, por ejemplo, a finales de la década de 2000 se lanzó un gran programa de investigación sobre tecnologías de acuerdo (*Agrement Techologies-Cost Action*) para investigar estos temas. Todo esto subraya que, para 2010, los problemas de coordinación de múltiples actores de IA —desde la comunicación y la cooperación hasta la gobernanza— ya estaban siendo estudiados activay parcialmente resueltos.

En resumen, la comunidad de IA ha tenido durante mucho tiempo definiciones claras de lo que es un agente de IA y lo que es un sistema multiagente. Estas definiciones

giran en torno a la autonomía, la acción intencional y, para los SMA, la interacción entre agentes. A continuación, resumimos las características clave de cada concepto:

- Agente inteligente: una entidad autónoma y flexible situada en un entorno, con propiedades de autonomía, reactividad, proactividad y capacidad social.

- Sistema multiagente: una colección de agentes autónomos que interactúan en un entorno común, posiblemente cooperando o compitiendo, lo que requiere mecanismos de comunicación y coordinación.

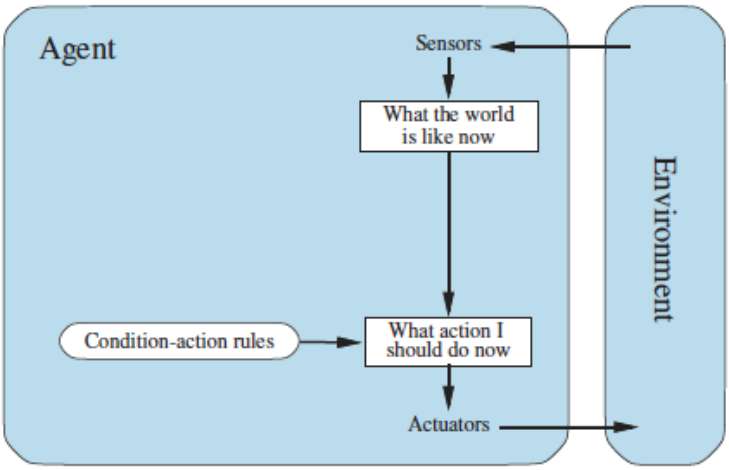
A partir de estos conceptos, se puede deducir que esta es un área bien establecida en el campo de la IA. A continuación, nos centraremos en cómo se construyen los agentes de IA, es decir, las arquitecturas que permiten a los agentes exhibir un comportamiento inteligente.

5. ARQUITECTURAS DE AGENTES INTELIGENTES

Stuart Russell y Peter Norvig, en su libro *Inteligencia Artificial: Un Enfoque Moderno* (Russell & Norvig, 2020) clasifican los agentes inteligentes según su capacidad para percibir, modelar, planificar y aprender. La evolución va desde simples agentes reactivos hasta agentes con capacidades de aprendizaje autónomo, y definen cinco arquitecturas de agentes abstractas que describimos brevemente a continuación.

FIGURA 3

ARTIFICIAL INTELLIGENCE: A MODERN APPROACH



Fuente: Russell & Norvig (2020).

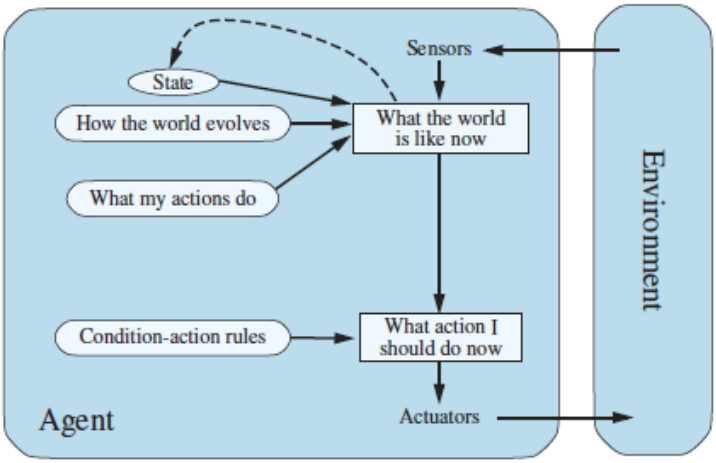
Agentes reactivos simples: Actúan directamente basándose en las percepciones actuales, utilizando reglas de condición → acción. No tienen memoria ni un modelo interno del entorno.

Agentes basados en modelos (memoria): Mantienen un modelo interno para representar el estado del entorno. Actualizan este modelo basándose en las percepciones recibidas.

Agentes basados en objetivos: Seleccionan acciones basadas en objetivos específicos. Usan técnicas de búsqueda y planificación para alcanzar estos objetivos.

FIGURA 4

ARTIFICIAL INTELLIGENCE: A MODERN APPROACH



Fuente: Russell & Norvig (2020).

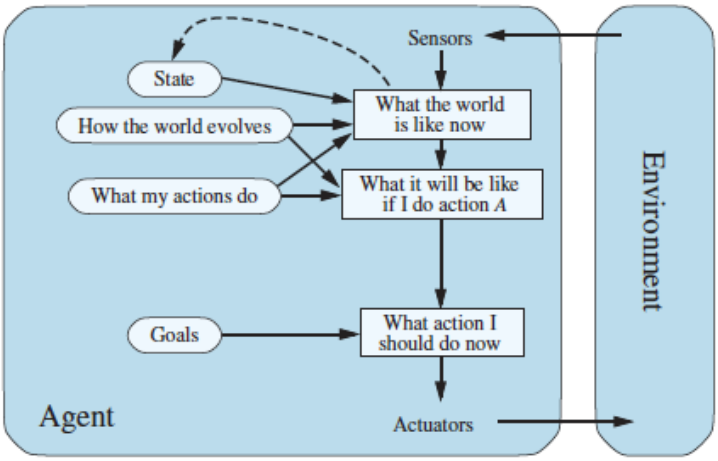
Agentes basados en utilidad: Generalizan los agentes basados en objetivos utilizando una función de utilidad para evaluar diferentes resultados y seleccionar el más preferido.

Agentes de aprendizaje: Capaces de mejorar su rendimiento a lo largo del tiempo a través de la experiencia. Incluyen módulos de acción, crítica y aprendizaje.

Diseñar un agente inteligente implica especificar cómo procesa la información desde la percepción hasta la acción. A lo largo de los años, los investigadores de IA han propuesto diversas arquitecturas de agentes, es decir, modelos subyacentes que definen la estructura interna y los procesos de toma de decisiones de un agente. Van desde los sistemas reactivos simples a los deliberativos complejos. A continuación, repasamos

FIGURA 5

ARTIFICIAL INTELLIGENCE: A MODERN APPROACH



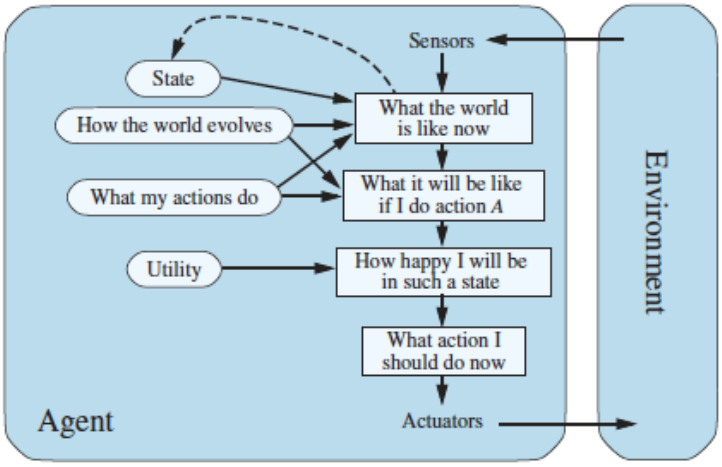
Fuente: Russell & Norvig (2020).

varios de los principales paradigmas arquitectónicos, ya que su comprensión aclarará cómo encajan (o no) los llamados sistemas "agentivos" con los enfoques tradicionales.

Agentes reactivos (reflejos): El tipo más simple de arquitectura de agente es un agente reactivo, también llamado agente reflejo simple. Este agente selecciona acciones basa-

FIGURA 6

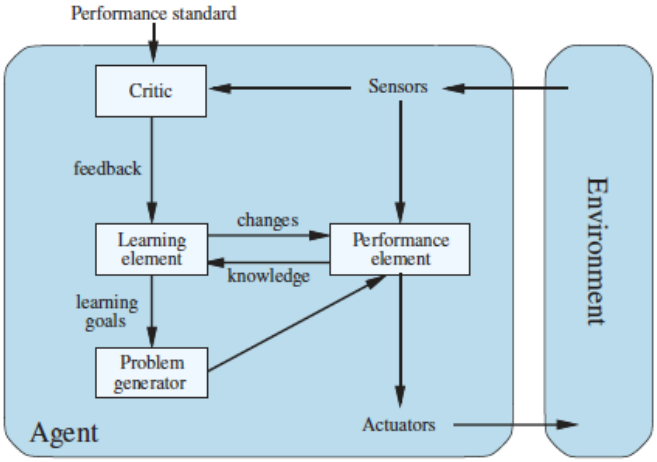
ARTIFICIAL INTELLIGENCE: A MODERN APPROACH



Fuente: Russell & Norvig (2020).

FIGURA 7

ARTIFICIAL INTELLIGENCE: A MODERN APPROACH



Fuente: Russell & Norvig (2020).

das en la percepción actual, ignorando la historia perceptiva. Implementa un mapeo directo de situaciones a acciones, básicamente un conjunto de reglas de condición-acción. Si la condición coincide con el estado actual del entorno, el agente ejecuta la acción asociada; si ninguna regla coincide, no hace nada. Los agentes reactivos no planifican explícitamente ni modelan el mundo, sino que actúan en tiempo real siguiendo una lógica "si-entonces". Este paradigma cobró importancia a mediados de los 80 como reacción a las limitaciones de la planificación simbólica pura en IA. La arquitectura de subsunción de Rodney Brooks (Brooks, 1986) es un ejemplo. En este modelo, el sistema de control del agente se construye en capas de comportamiento (por ejemplo, evitación de obstáculos, deambulación, búsqueda de objetivos), donde los comportamientos reactivos de bajo nivel pueden anular a los de alto nivel cuando sea necesario. Esto tuvo una gran influencia en la robótica basada en el comportamiento, ya que demostró que pueden surgir comportamientos eficaces a partir de redes de módulos simples y reflejos, sin un razonamiento simbólico centralizado. La ventaja de los agentes reactivos es su velocidad y robustez en entornos dinámicos (sin deliberaciones complejas que los ralenticen); la desventaja es que los agentes puramente reactivos tienen dificultades con los comportamientos estratégicos a largo plazo, ya que carecen de un modelo interno del mundo o de objetivos explícitos.

Agentes deliberativos (simbólicos): En el otro extremo del espectro están los agentes deliberativos, a veces llamados agentes de razonamiento deductivo (especialmente en la literatura temprana). Estos agentes mantienen un modelo simbólico del mundo y utilizan el razonamiento simbólico (lógica) para decidir sus acciones. En una arquitectura deliberativa, el agente tiene representaciones explícitas de creencias sobre el mundo y

utiliza la deducción lógica o la búsqueda para idear un plan de acción que alcance sus objetivos. Este enfoque se remonta a los inicios de la IA (años 50) y continuó en los años 70 y 80 en sistemas que intentaban demostrar teoremas o deducir planes a partir de principios formales. Un agente puramente deductivo podría, por ejemplo, utilizar la demostración automática de teoremas para deducir la mejor acción dado un objetivo formal y una base de conocimiento formal, un método que es correcto por diseño pero a menudo inviable desde el punto de vista computacional en entornos complejos y dinámicos. A finales de la década de 1980, se reconoció que los agentes puramente simbólicos eran frágiles y lentos en la práctica; los entornos de la vida real cambian más rápido de lo que pueden replanificarse, y el problema del encuadre (determinar qué cambia y qué permanece igual después de una acción) hace que los enfoques lógicos simples sean poco prácticos. Esta constatación dio lugar al movimiento de agentes reactivos antes mencionado, así como a arquitecturas que hibridan ambos extremos. Ejemplos de este tipo de arquitectura son AGENT0 y PLACA.

Agentes híbridos: Las arquitecturas híbridas intentan combinar los puntos fuertes de los enfoques reactivo y deliberativo. Un patrón común es un diseño de dos (o tres) capas: una capa reactiva se encarga de las respuestas de bajo nivel y en tiempo crítico al entorno, mientras que una capa deliberativa se ocupa de la planificación de alto nivel y la gestión de objetivos. Las dos capas deben coordinarse; por ejemplo, la capa reactiva puede anular los planes para hacer frente a amenazas inmediatas o, a la inversa, la capa deliberativa puede suprimir algunos comportamientos reactivos cuando se persigue un plan a largo plazo. Una de las primeras arquitecturas híbridas famosas fue la arquitectura TouringMachines (Ferguson, 1992), que tenía capas para las habilidades reactivas, la planificación y una capa de supervisión para gestionar los conflictos. Otros ejemplos son InteRRaP y las arquitecturas 3T (de tres niveles). En la década de 1990, el diseño híbrido de agentes era la norma para los robots complejos y los agentes de *software*: reconocía que ningún método es suficiente para todos los aspectos de la inteligencia. El enfoque híbrido reconoce la necesidad de reactividad para manejar los cambios en tiempo real y de deliberación para manejar el razonamiento estratégico. Muchos agentes de IA modernos siguen implícitamente este enfoque por capas, aunque no estén explícitamente etiquetados como tales.

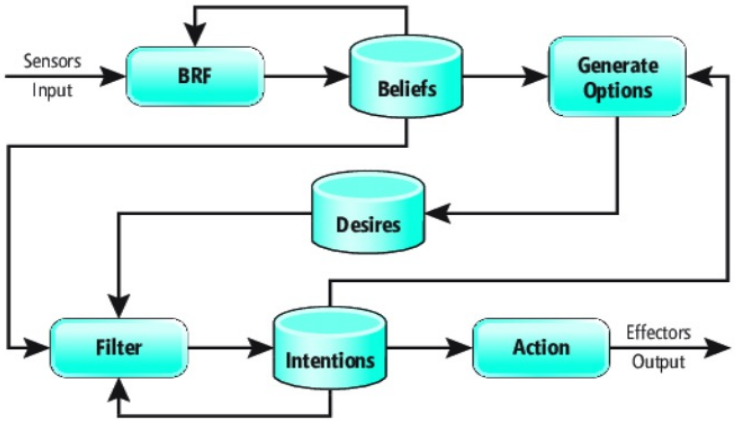
Agentes de Creencia-Deseo-Intención (Belief-Desire- Intention BDI): Una arquitectura particularmente importante, que se origina en el trabajo de Michael Bratman (Bratman, 1999 [1987]) sobre el razonamiento práctico humano y fue adaptada para la IA en la década de 1990. La arquitectura BDI proporciona un marco teórico estructurado para la deliberación del agente, está definida por los siguientes conceptos:

- Creencias (B): el estado de información del agente, lo que cree sobre el mundo.
- Deseos (D): los objetivos o situaciones que al agente le gustaría lograr.

- Intenciones (I): el subconjunto de deseos con los que el agente se ha comprometido.

FIGURA 8

ESQUEMA DE LA ARQUITECTURA DE UN AGENTE BDI



Nota: Esquema de la arquitectura de un agente BDI, que muestra cómo las percepciones del entorno son procesadas por la función de revisión de creencias (BRF) para actualizar las creencias del agente, que junto con las intenciones existentes informan la generación de nuevas opciones (posibles deseos). A continuación, un filtro selecciona qué deseos adoptar como Intenciones actuales, y el agente actúa para lograr estas intenciones. Este ciclo de observación, actualización de creencias, deliberación (opciones → intenciones) y acción se repite continuamente.

Fuente: https://learn.microsoft.com/en-us/archive/msdn-magazine/2019/january/machine-learning-leveraging-the-beliefs-desires-intentions-agent-architecture

El modelo BDI define un bucle de control en el que el agente actualiza continuamente sus creencias, genera opciones (deseos) basadas en esas creencias y en las intenciones actuales, filtra (delibera) esos deseos para elegir con cuáles comprometerse (intenciones) y, a continuación, planifica para determinar qué acciones desempeñar para satisfacer las intenciones y, finalmente, actúa para cumplir las intenciones. Esto se asemeja mucho a la forma en que los humanos razonamos sobre las acciones: tenemos muchos deseos potenciales en cualquier momento, pero los filtramos en función de las prioridades y la viabilidad para convertirlos en intenciones, que luego intentamos ejecutar. Un agente BDI no vuelve a planificar a ciegas desde cero en cada momento, sino que mantiene su compromiso con las intenciones hasta que se cumplen determinadas condiciones (objetivo alcanzado, objetivo imposible o sustituido por un objetivo de mayor prioridad), lo que proporciona un equilibrio entre la persistencia proactiva y la adaptabilidad reactiva.

La arquitectura BDI formaliza el razonamiento práctico en los agentes, permitiéndoles equilibrar la reactividad (mediante actualizaciones perceptivas) con el comportamiento orientado a objetivos (mediante intenciones). Se ha utilizado como base para muchos lenguajes y plataformas de programación de agentes (por ejemplo, PRS, JACK, JAM). El énfasis del modelo BDI en el estado mental del agente (actitudes informativas y motivacionales) conecta con la teoría de sistemas intencionales de Dennett, proporcionando una forma rigurosa de diseñar agentes que parecen tener creencias, deseos e intenciones; de hecho, en BDI, estas son estructuras de datos explícitas.

Además de las arquitecturas presentadas, existen otras muchas (agentes impulsados por eventos, agentes de redes neuronales, etc.), pero las comentadas ilustran los principales paradigmas. Está claro que, a principios de la década de 2000, los investigadores de IA habían desarrollado un sofisticado conjunto de herramientas para construir agentes. Las arquitecturas reactivas ofrecían velocidad, las deliberativas previsión, las híbridas intentaban combinar lo mejor de ambas y las BDI añadían claridad filosófica y estructura práctica a las estrategias de compromiso.

Es en este contexto en el que consideramos la reciente oleada de "Agentic AI". Cuando los comentarios actuales hablan de sistemas de IA agentiva, debemos preguntarnos: ¿cuál de estas arquitecturas establecidas (o combinaciones) se está utilizando realmente? ¿Los actuales "agentes autónomos basados en los LLM" representan algo fundamentalmente novedoso, o podrían ser una implementación de, por ejemplo, un ciclo BDI con un LLM que se encarga de parte de la toma de decisiones? Para explorar esta cuestión, examinamos ahora qué significa la Agentic AI en el contexto de las tendencias actuales de la IA generativa.

6. DE LOS LLM A LA "AGENTIC AI": LA NUEVA OLA DE AGENTES AUTÓNOMOS

En el año 2023 se produjo una explosión de interés por encadenar y orquestar grandes modelos lingüísticos para realizar tareas autónomas orientadas a objetivos. El lanzamiento de potentes LLM (GPT-3, GPT-4, GPTo, Gemini, Deepsek, Llama, etc.) capaces de razonar y generar código llevó a los desarrolladores a crear sistemas como AutoGPT, BabyAGI, LangChain, GraphChain, AutoGen, CAMEL, ADK y muchos otros, que han permitido la creación de agentes y SMA en los que los agentes son LLM capaces de razonar, planificar y comunicarse entre sí utilizando lenguaje natural. Estos SMA basados en LLM aprovechan la inteligencia colectiva de múltiples agentes especializados para abordar tareas complejas de forma más eficiente que un único agente. Un conjunto de LLM puede asumir funciones especializadas definidas por instrucciones, colaborando mediante el intercambio de información y la coordinación de acciones en lenguaje natural, de forma similar a la interacción humana. Sin embargo, el éxito de los LLM en tareas complejas requiere una intervención humana considerable para guiar la genera-

ción de textos y la toma de decisiones. Por tanto, la coordinación autónoma entre LLM plantea varios retos, como garantizar que los agentes cooperen de forma que se alineen con los objetivos humanos. Otro hito importante es que se han desarrollado también propuestas de estandarización como MCP (*Model Context Protocol*) desarrollado por Anthropic (estándar abierto que permite a los desarrolladores crear conexiones seguras y bidireccionales entre sus fuentes de datos y herramientas basadas en inteligencia artificial) o A2A (Agent2Agent Protocol, un estándar abierto que permite a los agentes de IA comunicarse y colaborar entre diferentes plataformas y marcos, independientemente de las tecnologías subyacentes). Aunque en esta nueva ola de los agentes basados en LLM aún queda mucho trabajo de estandarización, que debería basarse en los trabajos ya realizados en el área de los sistemas multiagente, son iniciativas importantes que van en el camino adecuado.

Esta tendencia ha ido acompañada de la popularización del término "Agentic AI (IA agentiva)". Pero, ¿qué significa exactamente este término?

En el ámbito industrial, el término "Agentic AI" (Softude, 2025; Silfverskiöld, 2025) se refiere generalmente a los sistemas de IA que pueden decidir y ejecutar acciones de forma autónoma para alcanzar un objetivo de alto nivel, a menudo dividiendo el objetivo en subtareas y posiblemente coordinando múltiples componentes o subagentes. Implica un grado de iniciativa (proactividad) y planificación a largo plazo que va más allá de las respuestas puntuales o la generación de texto. Por ejemplo, un asistente de Agentic AI no solo podría responder a una consulta del usuario, sino también tomar medidas proactivas (por ejemplo, programar una reunión, enviar correos electrónicos, ejecutar transacciones) sin indicaciones explícitas para cada paso. En esencia, esto describe un agente inteligente en el sentido clásico. El agente percibe un objetivo, decide de forma iterativa "¿Qué subobjetivo o acción debe realizar a continuación para alcanzarlo?", posiblemente invoca herramientas externas o API, interactúa con otros agentes, supervisa el progreso, etc. El entusiasmo en torno a estas capacidades en 2023-2024 es real, impulsado por impresionantes demostraciones de agentes basados en LLM, pero conceptualmente, esto puede identificarse claramente con arquitecturas y conceptos de agentes autónomos que existen desde hace décadas.

Es revelador ver cómo las fuentes modernas distinguen entre "Agentic AI" y "agentes de IA". Según un artículo de IBM (IBM, 2023), "la Agentic AI es el framework; los agentes de IA son los componentes básicos dentro de ese framework". En otras palabras, la Agentic AI se refiere a un sistema global (quizás un ecosistema de módulos que incluye uno o más agentes) diseñado para resolver problemas con una supervisión mínima, mientras que un agente de IA es un componente autónomo específico que realiza tareas individuales dentro de dicho sistema. Softude (2025) describe de manera similar a un agente de IA como "un bloque de construcción de Agentic AI... programado para funcionar de forma autónoma", con bucles de percepción, decisión y acción, mientras que la Agentic AI se refiere a sistemas con un alto grado de agencia que "establecen

objetivos de forma proactiva, toman decisiones y actúan con una intervención humana mínima". Los sistemas denominados de IA agentiva implementan los principios clásicos de agencia usando modelos fundacionales. Los agentes basados en los LLM ejecutan tareas complejas mediante razonamiento lingüístico, planificación y acceso a herramientas externas. Esta tendencia no constituye un nuevo paradigma, sino una ampliación tecnológica del marco deliberativo tradicional. En este sentido, la IA agentiva se podría entender como una instancia moderna del modelo BDI (*Belief–Desire–Intention*) (Jao & Georgeff, 1991 y 1995), donde las creencias son *embeddings* contextuales, los deseos son prompts y las intenciones, secuencias de acciones generadas.

La clave es que la Agentic AI a menudo implica no un solo agente, sino un conjunto orquestado de agentes y herramientas, junto con una arquitectura que les permite coordinarse dinámicamente. En términos de computación en la nube, un sistema de Agentic AI podría considerarse como un conjunto o flujo de trabajo (a veces llamado un "sistema de orquestación de agentes"), mientras que un agente inteligente es un actor autónomo individual dentro de él.

En IBM (2023), el ejemplo de IBM para la IA agencial es un sistema inteligente de energía doméstico que gestiona múltiples agentes dispositivos (IoT). La Agentic AI (el sistema global) comprende el objetivo general del propietario (eficiencia energética) y coordina los agentes de IA individuales —el agente termostato, el agente iluminación, los agentes electrodomésticos—, cada uno con su propia tarea y objetivo específicos. Este ejemplo se corresponde con el concepto de sistema multiagente: el hogar inteligente tiene múltiples agentes (termostato, luces, etc.) que, en este caso, cooperan entre sí. De hecho, no se introduce nada conceptualmente exótico al llamarlo "agente": se trata esencialmente de un sistema de control multiagente inteligente, un tema ampliamente tratado en la investigación sobre IA distribuida.

Entonces, ¿por qué esta nueva terminología? Probablemente se deba en parte al *marketing* o al ciclo natural de las ideas que son redescubiertas por nuevas comunidades. El resurgimiento del interés por los "agentes inteligentes" entre los desarrolladores se produjo tras la llegada de los LLM avanzados, que proporcionaron un nuevo conjunto de herramientas para implementar comportamientos de agentes (por ejemplo, utilizando LLM para interpretar objetivos y generar planes en lenguaje natural). Para comercializar esto, términos como "Agentic AI" ganaron popularidad, tal vez para señalar un cambio de los modelos de IA estáticos (que generan resultados de forma pasiva) a los sistemas de IA que son participantes activos y orientados a objetivos (agentes). Sin embargo, lo irónico es que las nociones subyacentes de autonomía e interacción multiagente no son nada nuevas.

Como se ha comentado anteriormente, también hemos asistido a una proliferación de plataformas y bibliotecas de código abierto para facilitar la construcción de estos agentes basados en LLM. De hecho, en 2024-2025, han surgido docenas de nuevas

plataformas basadas en LLM, muchos de ellas lanzadas en el plazo de un año. Algunos ejemplos son AutoGen, AgentScope, CrewAI, CamelAI, ADK, OpenAI, 8n8, LangChain, AutoGen, Hugging Face Transformers Agents, LangGraph, Agenta, Minichain, Auto-GPT, BabyAGI y muchos más. Estas plataformas integran modelos LLM que permiten el desarrollo de agentes basados en LLM para razonar, planificar y actuar en entornos abiertos. Su funcionamiento se basa en tres capacidades clave:

- Razonamiento: Los LLM generan cadenas de pensamiento para razonar antes de actuar.
- Memoria: Integran recuerdos episódicos y a largo plazo (almacenes vectoriales, registros) para contextualizar las decisiones.
- Uso de herramientas: Utilizan herramientas externas (calculadoras, navegadores, API) a través de indicaciones estructuradas.

Proporcionan abstracciones para definir herramientas, gestionar instrucciones y conectar múltiples agentes basados en LLM.

Plataformas como LangChain, AutoGen y CrewAI obtuvieron rápidamente decenas de miles de estrellas en GitHub (barras doradas), lo que indica un gran interés por parte de los desarrolladores, mientras que muchas plataformas más recientes entraron en escena a finales de 2024 con una creciente aceptación. Esto refleja una rápida expansión de las herramientas y bibliotecas para crear agentes inteligentes autónomos, a menudo centrados en la orquestación de LLM con herramientas externas. La tendencia subraya el entusiasmo de la industria por las soluciones de IA basadas en agentes, pero también conlleva el riesgo de fragmentación y repetición de ideas conocidas en la investigación académica sobre agentes.

La proliferación de marcos sugiere que muchos se enfrentan a retos similares: cómo permitir que un agente de IA utilice herramientas (API, bases de datos), cómo mantener la memoria de interacciones pasadas, cómo coordinar múltiples agentes o encadenar subtareas, etc. Se trata, en esencia, de implementaciones de lo que la comunidad de investigación sobre agentes reconocería como planificación, memoria/base de conocimientos y coordinación multiagente. Por ejemplo, un patrón de diseño común es un agente "planificador" basado en LLM que puede generar agentes "ejecutores" para subtareas, que luego informan de sus resultados, lo que recuerda a la planificación jerárquica de tareas o a las organizaciones de agentes maestro-esclavo de la literatura sobre MAS.

Otro ejemplo: el concepto de dar acceso a un agente LLM a herramientas (como búsquedas web o calculadoras) y esperar que decida cuándo usar cada una de ellas, es análogo a un agente que razona sobre sus acciones y elige un efector o subrutina apro-

piados. Clásicamente, esto podría asignarse al repertorio de acciones del agente y al acceso al entorno. Hoy en día, los investigadores están inventando técnicas de instrucción (por ejemplo, "ReAct", que entrelaza el razonamiento con el uso de herramientas) para lograr estos comportamientos con los LLM, de forma similar a trabajos anteriores sobre sistemas de planificación de agentes que incluían acciones tanto internas (razonamiento) como externas (ejecución de una llamada API).

El verdadero valor añadido de los nuevos sistemas basados en LLM reside en la flexibilidad y los conocimientos que encapsulan los grandes modelos preentrenados. El hecho de que una instancia de GPT-4, Gemini, Llama, Deepsek o cualquier otro modelo pueda analizar una instrucción arbitraria, descomponerla en pasos y generar código o texto para llevarlos a cabo, supone un gran avance en cuanto a capacidad en comparación con, por ejemplo, un agente BDI codificado de forma rígida que solo conoce una biblioteca fija de planes. Esto ha permitido crear agentes mucho más polivalentes. Por ejemplo, un LLM bien diseñado puede servir como interfaz universal entre los objetivos del lenguaje natural y las acciones formales, funcionando como el "cerebro" del agente que razona a un alto nivel de forma abierta. Esto no era posible con una buena fiabilidad antes del aprendizaje profundo, por lo que el entusiasmo en torno a la Agentic AI está justificado: estos sistemas pueden hacer cosas con las que los sistemas de agentes anteriores tenían dificultades, gracias a la potencia de los LLM.

La velocidad con la que se están creando herramientas de desarrollo de sistemas de agentes y multiagente basados en LLM es también un factor tremendamente positivo. Durante la evolución del campo de los agentes y los sistemas multiagentes, la disponibilidad de plataformas de desarrollo (JADE, JACK, MADKit, ZEUS, Magentix, etc.) evolucionó muy lentamente, mientras que hoy en día tenemos una amplia gama entre la que elegir, y siguen aumentando.

Sin embargo, a nivel conceptual, los sistemas de Agentic AI siguen siendo agentes inteligentes. Del mismo modo, un sistema multiagentic (término que se utiliza ocasionalmente para referirse a escenarios multiagente LLM) es, por definición, un sistema multiagente.

7. MAL USO DE LA TERMINOLOGÍA Y CONFUSIÓN CONCEPTUAL

De lo que hemos discutido hasta ahora, podemos abordar el meollo del asunto: el mal uso de los términos "agentic (agentiva)" y "multiagentic (multiagentiva)" en el discurso actual sobre la IA. Muchos artículos y comunicados de prensa recientes utilizan estas palabras como si denotaran una nueva tecnología, cuando en realidad están describiendo capacidades o diseños de sistemas examinados durante mucho tiempo bajo las banderas de los agentes inteligentes y los sistemas multiagente. Esta confusión conceptual puede ser problemática por varias razones:

- Oculta la riqueza de la investigación existente. Al no reconocer que la Agentic AI se refiere esencialmente a agentes autónomos, los nuevos profesionales pueden pasar por alto décadas de literatura sobre cómo diseñar y evaluar estos sistemas. Como resultado, pueden repetir errores del pasado o volver a trabajar en problemas ya resueltos. De hecho, una de las principales motivaciones de este artículo es la preocupación de que ignorar el trabajo anterior nos lleve a "reinventar la rueda" al aplicar la Agentic AI a nuevos ámbitos.

- Crea confusión terminológica. Tradicionalmente, en IA, el adjetivo "agentic" rara vez se utilizaba; el término "agencia" ha sido más frecuente; los investigadores simplemente hablaban de sistemas basados en agentes, agentes autónomos, etc. Ahora vemos frases como "agentic AI agents", que son redundantes (literalmente, "agentes de IA similares a agentes"). El término "multiagentic" es aún más problemático, ya que intenta adjetivar "multiagente". Decir "sistema multiagentic" es un neologismo innecesario cuando el término "sistema multiagente" es claro y está bien definido. Esta jerga puede confundir a los recién llegados y enturbiar las comunicaciones, especialmente entre la industria y el mundo académico. Por ejemplo, un ingeniero de *software* podría decir: "Hemos creado una IA multiagente para hacer X", mientras que un investigador de IA lo interpretaría como "un sistema multiagente para X". Es de esperar que se refieran a lo mismo, pero la terminología divergente puede impedir la transferencia de conocimientos.

- Puede exagerar la novedad. Cuando se acuña un nuevo término, a veces se da a entender que el concepto en sí es nuevo. Esto puede dar lugar a expectativas exageradas o a un reconocimiento injustificado. Consideremos que afirmaciones como "La Agentic AI va un paso más allá al permitir la toma de decisiones y la ejecución de tareas de forma autónoma", es una descripción acertada, pero que se aplica igualmente a los agentes inteligentes desde la década de 1990. O afirmaciones como "La Agentic AI se refiere a sistemas en los que colaboran múltiples agentes de IA... A diferencia de un único agente inteligente, que opera de forma aislada", básicamente describe un sistema multiagente. Sin contexto histórico, sin rigor científico, los lectores pueden pensar que estas capacidades fueron inventadas por los laboratorios de IA actuales.

Seamos claros: el término "agentic" no es un concepto nuevo en la IA, tiene raíces en las ciencias sociales y cognitivas y fue adoptado posteriormente (moderadamente) en la IA, pero las ideas subyacentes han sido parte del desfile de agentes de la IA durante mucho tiempo. Cuando los documentos técnicos de 2023 hablan de "Agentic AI", generalmente se refieren a lo que los investigadores de IA llamarían simplemente un agente inteligente/agente autónomo o un sistema basado en agentes. En otras palabras, deberíamos llamar a las cosas por su nombre: si es un agente de IA, llámalo

agente; si es un sistema multiagente, llámalo así. La IA agentiva no sustituye a los agentes autónomos/ sistemas multiagente, sino que los reinterpreta como ecosistemas de agentes lingüísticos, donde la interacción se da en lenguaje natural y la coordinación se logra mediante mecanismos de argumentación y diálogo.

Para ser claros, no hay que menospreciar o restar importancia a los recientes logros de la ingeniería en la implementación de agentes basados en LLM con nuevas herramientas de IA. Los avances son reales, pero encajan dentro de un marco conceptual ya existente. Al reconocer esto, los desarrolladores e investigadores pueden beneficiarse enormemente de las lecciones ya aprendidas. Por ejemplo:

La comunidad MAS ha estudiado los comportamientos emergentes en los sistemas multiagente. Algunos de los primeros experimentos con sistemas multiagente basados en LLM han observado comportamientos sorprendentes (positivos o negativos) cuando varios agentes LLM interactúan (por ejemplo, quedarse atascados en bucles de preguntas entre ellos). Muchos de ellos podrían entenderse utilizando enfoques de teoría de juegos o de coordinación desarrollados para MAS.

Las cuestiones relacionadas con la confianza y la seguridad en los agentes autónomos (como que un agente de IA se vuelva incontrolable o cause daños no deseados) se han debatido durante mucho tiempo en el contexto de la autonomía limitada y la autonomía ajustable. Existen marcos para incorporar restricciones o normas éticas en la toma de decisiones de los agentes. En lugar de empezar desde cero, las nuevas iniciativas de Agentic AI pueden adaptar estos marcos a los agentes basados en LLM.

La importancia de los estándares no puede subestimarse. A finales de la década de 1990, los investigadores en agentes se dieron cuenta de que, sin interfaces estándar, era difícil conseguir que agentes de diferentes fuentes trabajaran juntos. Esto cobra especial importancia cuando hablamos de sistemas multiagente abiertos en los que interactúan agentes de diferentes desarrolladores que pueden entrar y salir del sistema multiagente. En la actual proliferación de plataformas de agentes, también observamos fragmentación: una plataforma puede no comunicarse fácilmente con otra. La industria de la IA haría bien en recordar el valor de los estándares. Es alentador que estén surgiendo algunas iniciativas, como el Protocolo de Contexto de Modelos (MCP) de Anthropic en 2023 (Anthropic, 2023), cuyo objetivo es estandarizar la forma en que los asistentes de IA (agentes) interactúan con los sistemas externos. O A2A (Agent2Agent Protocol) (A2A Protocol, 2024), un estándar abierto que permite a los agentes de IA comunicarse y colaborar entre diferentes plataformas y marcos, independientemente de las tecnologías subyacentes. Se trata de un paso en la dirección correcta, pero se necesita más colaboración para evitar una situación en la que cada plataforma de agentes esté aislada de las demás.

Otro aspecto que se confunde en los medios es la equiparación excesiva del concepto de agente con los LLM. Si bien el entusiasmo actual se debe en gran medida a las capacidades de los LLM, el concepto de agente no exige el uso de un LLM. Un agente podría utilizar la lógica tradicional basada en reglas, una política de aprendizaje por refuerzo o cualquier combinación de técnicas de IA. De hecho, es probable que las mejores soluciones combinen varias técnicas de IA; por ejemplo, un enfoque híbrido de IA en el que el razonamiento simbólico (para la planificación a largo plazo o para garantizar las restricciones) complemente métodos subsimbólicos como las redes neuronales para la percepción o el procesamiento del lenguaje natural. Algunos han denominado a esta convergencia IA neurosimbólica. Las arquitecturas de agentes son un lugar natural para dicha convergencia: un agente podría utilizar un modelo neuronal para interpretar una situación y un planificador simbólico para decidir una secuencia de acciones, o un LLM para proporcionar capacidad de conversación al agente. Al reintegrar la investigación sobre agentes con la IA moderna, podemos lograr sistemas que realmente exhiban una inteligencia robusta.

8. HACIA LA INTEGRACIÓN: ABRAZANDO EL LEGADO DE LA INVESTIGACIÓN EN AGENTES EN LA IA MODERNA

¿Cómo podemos avanzar de manera que nos beneficiemos tanto de los nuevos avances como de los fundamentos establecidos? En primer lugar, la comunidad de IA (tanto investigadores como profesionales) debe esforzarse por utilizar una terminología clara y mantener la coherencia conceptual, siempre basada en el rigor científico. Si está creando un sistema en el que interactúan múltiples componentes basados en LLM, tenga en cuenta que ha creado un sistema multiagente y consulte la bibliografía sobre MAS para obtener orientación sobre el diseño y los posibles escollos. Si su aplicación implica un ciclo de toma de decisiones autónomo, considérela un agente inteligente y aproveche las arquitecturas de agentes conocidas (reactivas, BDI, etc.) para estructurarla. En publicaciones y debates, el uso de términos estándar (quizás con una nota que indique que se ajustan al término popular "Agentic AI") facilitará la interacción entre los investigadores experimentados en agentes y la nueva generación de desarrolladores entusiasmados con los agentes basados en LLM.

Utilizar el término IA agentiva (Agentic AI) no es incorrecto, si entendemos la IA agentiva como un estadio de integración neurosimbólica que articula el conocimiento estructurado y el razonamiento generativo, donde se integran los avances de la IA generativa en las teorías de agentes y sistemas multiagente preexistentes. El término sistemas multiagentivos (multiagentic systems) es, en mi opinión, más cuestionable y no tiene rigor científico; debería ser utilizado el término sistemas multiagente.

En segundo lugar, debemos incorporar activamente los resultados probados del área de los agentes en los nuevos marcos de agentes basados en LLM. Por ejemplo, el

concepto de un lenguaje de comunicación entre agentes podría ser muy relevante si empezamos a tener múltiples agentes basados en LLM que se comunican entre sí. Actualmente, la mayor parte de la comunicación entre agentes en los sistemas LLM se realiza en lenguaje natural ad hoc (literalmente, un agente da instrucciones a otro en lenguaje natural). Esto está bien para la creación de prototipos, pero a medida que aumenta la complejidad, podría redescubrirse la necesidad de un intercambio más estructurado (para evitar malentendidos y reducir la verbosidad). Los investigadores de la década de 1990 ya desarrollaron protocolos de comunicación basados en actos de habla; estos podrían inspirar esquemas de solicitud más eficientes o el paso de mensajes basados en API entre agentes. Trabajos recientes como el protocolo de comunicación autorreferencial (*Self-Referential Communication,* SRC) para agentes LLM (Barak, 2024) o Agora (*A Scalable Communication Protocol for Networks of Large Language Models*) (Marro *et al.,* 2024) se hacen eco de estas ideas, aunque expresadas en términos modernos.

Del mismo modo, las estrategias de coordinación y competencia de los MAS (como los algoritmos de asignación de tareas, los protocolos de consenso, los mecanismos de votación, la negociación automática, la argumentación automática, etc.) pueden transferirse a los sistemas de IA basados en agentes LLM para gestionar múltiples agentes que trabajan en subproblemas. Si, por ejemplo, se crea un equipo de agentes de IA basados en LLM para actuar como fuerza de ventas virtual (uno busca clientes potenciales, otro escribe correos electrónicos, otro programa llamadas), surge un problema de asignación y programación de recursos multiagente. En lugar de un enfoque ingenuo de prueba y error, se podrían aplicar los algoritmos de coordinación MAS existentes.

Pero no olvidemos que en muchos otros ámbitos problemáticos tendremos que desarrollar sistemas multiagente abiertos y, para ello, tendremos que tener en cuenta los resultados y conceptos ya desarrollados en el campo de los sistemas multiagente, principalmente las normas, en general las tecnologías de acuerdo, para desarrollar coreografías de agentes.

Por otro lado, la comunidad de investigación sobre agentes también debe actualizar sus herramientas aprovechando las ventajas de los LLM y el aprendizaje automático moderno. Las arquitecturas clásicas de agentes a menudo tenían dificultades con la adquisición de conocimientos y la flexibilidad, áreas en las que los LLM destacan. Un agente BDI con un planificador basado en LLM o un módulo de percepción basado en LLM podría ser mucho más potente que un agente BDI puramente simbólico. La integración del aprendizaje en las arquitecturas de los agentes (para que estos mejoren con la experiencia) ha sido un reto desde hace mucho tiempo. Las técnicas actuales de aprendizaje por refuerzo y aprendizaje con pocos ejemplos podrían finalmente hacerlo posible en la práctica. En resumen, la sinergia entre la sabiduría de los agentes tradicionales y las nuevas capacidades de la IA puede ampliar las fronteras de los sistemas autónomos.

9. CONCLUSIÓN

La "Agentic AI" se ha convertido en una palabra de moda que ha capturado la imaginación del mundo tecnológico, pero, como hemos argumentado, se trata esencialmente de un cambio de nombre de conceptos ya establecidos sobre agentes inteligentes autónomos, contextualizados por los recientes avances en IA generativa. El entusiasmo que suscitan los agentes autónomos basados en LLM está justificado por sus impresionantes nuevas capacidades, pero no debe llevarnos a ignorar el vasto conjunto de conocimientos sobre agentes y sistemas multiagente desarrollados a lo largo de décadas. El término "agentic", tomado de las ciencias sociales, destaca el comportamiento intencional y orientado a objetivos, precisamente el foco de la investigación sobre agentes inteligentes desde la década de 1990. Del mismo modo, los escenarios "multiagentic" son simplemente sistemas multiagente con otro nombre. Utilizar estos términos como si denotaran ideas novedosas corre el riesgo de generar confusión y repetir trabajos y errores del pasado.

Al revisar los fundamentos teóricos (la agencia de Bandura, los sistemas intencionales de Dennett) y los logros previos de la comunidad de IA (definiciones, propiedades, arquitecturas y plataformas para agentes autónomos), hacemos hincapié en que la rueda ya se ha inventado en lo que respecta a los agentes autónomos. Afortunadamente, es una rueda bastante buena, capaz de avanzar con el impulso de la IA generativa. Animamos a los profesionales actuales a basarse en investigaciones previas. En términos prácticos, esto implica adoptar la terminología correcta, consultar la bibliografía sobre sistemas de agentes/multiagentes para obtener orientación e integrar principios de diseño probados (como estándares de comunicación, arquitecturas de agentes, tecnologías de acuerdo, etc.) en los nuevos sistemas de IA.

El término "agentiva" proviene de la psicología social (Bandura, 1986) y denota la capacidad de actuar intencionalmente (Dennett, 1971). Su adopción en la IA moderna busca resaltar la supuesta intencionalidad de los agentes lingüísticos. Sin embargo, esta resemantización introduce redundancia conceptual, al atribuir "agencialidad" a entidades (agentes) que ya la poseen por definición. La llamada IA agentiva no introduce una nueva forma de agencia, sino un nuevo medio para expresarla lingüísticamente. El riesgo radica en la inflación terminológica: "agentes agentivos" o "IA multiagentiva" son tautologías que debilitan la precisión científica.

Por el contrario, también reconocemos que las nuevas tecnologías a menudo requieren adaptaciones de ideas antiguas. La comunidad de agentes debe aprovechar las oportunidades que ofrecen los LLM y el aprendizaje profundo para mejorar las capacidades de razonamiento y aprendizaje de los agentes. El resultado de combinar estos enfoques serán agentes autónomos verdaderamente avanzados, robustos, adaptables y capaces de abordar tareas complejas del mundo real. De hecho, como sugiere una visión, el futuro podría estar en los agentes neurosimbólicos, sistemas que combinan la percep-

ción y la comprensión del lenguaje impulsadas por la IA generativa con el razonamiento simbólico y la planificación. En mi opinión, esta integración de agentes, IA generativa, IA débil e IA responsable (RAI) es el camino hacia la IA neurosimbólica o IA híbrida, que es el siguiente paso en la evolución de la IA y también probablemente el futuro cercano de la IA "con cuerpo".

En conclusión, el campo de la IA debe tener cuidado de no perder de vista su propia historia en medio del entusiasmo actual. Los conceptos de agentes y sistemas multiagente siguen siendo tan relevantes en 2025 como lo fueron en 1995, incluso si los implementamos hoy con herramientas mucho más potentes. En lugar de acuñar términos ambiguos como "Agentic AI", deberíamos llamar a estos sistemas por su nombre y construir sobre los cimientos ya establecidos. Al hacerlo, no solo reconocemos el mérito donde corresponde, sino que también aceleramos el progreso, evitando volver sobre lo ya conocido y centrándonos en innovaciones genuinas. La rueda de los agentes inteligentes no necesita reinventarse; necesita perfeccionarse y potenciarse con los motores de la IA generativa. El camino que tienen por delante los agentes autónomos y sistemas multiagente es apasionante y, al unir los conocimientos del pasado con la tecnología actual, podemos garantizar que se construya sobre una ruta ya trazada, que nos lleve a destinos más allá de lo que las generaciones anteriores de IA fueron capaces de alcanzar.

La IA agentiva no representa una ruptura o innovación conceptual, sino una evolución tecnológica del paradigma de agentes autónomos y sistemas multiagente. Su relevancia reside en la capacidad de integrar los avances de la IA generativa dentro de marcos de acción autónoma preexistentes. La continuidad teórica entre los agentes clásicos y los sistemas basados en LLM es evidencia de la madurez del campo. El desafío actual no es definir nuevos tipos de agentes, sino preservar la coherencia teórica y evitar la fragmentación conceptual. Podemos ver la IA agentiva como un estadio de integración neurosimbólica que articula el conocimiento estructurado y el razonamiento generativo. Esta visión coincide con la necesidad, ya señalada por Wooldridge (2002) y Ferber (1999), de dotar a los sistemas multiagente de capacidades cognitivas más ricas sin perder su base formal. En consecuencia, la IA agentiva no sustituye a los agentes autónomos/SMA, sino que los reinterpreta como ecosistemas de agentes lingüísticos, donde la interacción se da en lenguaje natural y la coordinación se logra mediante mecanismos de argumentación y diálogo. El futuro de la disciplina dependerá de la capacidad para combinar la tradición formal de la teoría de agentes con las nuevas herramientas generativas y multimodales.

Referencias

Anthropic. (2023). Model Context Protocol (MCP) – Standard proposal by Anthropic for connecting AI assistants with external systems.

Austin, J. L., and Urmson, J. O. (1975). *How to do things with words* (2nd ed.). Cambridge, Mass.: Harvard University Press. ISBN 978-0674411524.

Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice- Hall. (Introduced the term "agentic" in context of human agency).

Bratman, M. E. (1999) [1987]. *Intention, Plans, and Practical Reason*. CSLI Publications. ISBN 1-57586-192-5.

Brooks, R. (1986). A Robust Layered Control System for a Mobile Robot. *IEEE Journal on Robotics and Automation, 2*(1): 14–23. (Describes the subsumption architecture, a reactive agent design).

Dennett, D. C. (1971). Intentional Systems. *Journal of Philosophy, 68*(2), 87–106. (Philosophical basis for attributing beliefs and desires to systems).

Ferber, J. (1999). *Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence.* Addison-Wesley.

Finin, T., Fritzson, R., McKay, D., McEntire, R. (1994). KQML as an agent communication language. Proceedings of the third international conference on Information and knowledge management - CIKM 94. p. 456. doi:10.1145/191246.191322. ISBN 0897916743. S2CID 1129799. http://www.fipa.org/repository/aclspecs.html

FIPA. Foundation for Intelligent Physical Agents – (Standards organization for agent interoperability, est. 1996).

Gates, Bill. (2023). The Age of AI has Begun – GatesNotes Blog, Nov 9, 2023. (Discusses personal AI assistants and agents revolutionizing computing).

Genesereth, M. and Fikes, R. (June 1992). Knowledge Interchange Format Version 3.0 Reference Manual (PDF). Stanford Logic Group Report. Logic-92-1. Stanford University. Retrieved 7 August 2014. http://www.ksl.stanford.edu/software/ontolingua/

IBM – Think Blog. (2023). Agentic AI vs AI Agents: Understanding the Difference. IBM Think Topics. (Explains agentic AI as a framework vs AI agents as components).

IFAAMAS. International Foundation for Autonomous Agents and Multiagent Systems – http:// (Professional association supporting the AAMAS conference and MAS research).

Jennings, N. R., Sycara, K., & Wooldridge, M. (1998). A roadmap of agent research and development. *Autonomous Agents and Multi-Agent Systems, 1*(1), 7–38. (Overview of the state of agent R&D in late 90s; covers architectures, applications, challenges).

Marro, S., La Malfa, E., Wright, J., Li, G., Shadbolt, N., Wooldridge, M., Torr, P. (2024). *A Scalable Communication Protocol for Networks of Large Language Models. arXiv:2410.11905.*

Ossowski, S., Sierra, C., & Botti, V. (2012, November 28). *Agreement Technologies: A Computing Perspective. Law*, Governance and Technology Series. Springer Netherlands. http://doi.org/10.1007/978-94-007-5583-3_1. https://dev.to/tomer_barak/self-reference-in-ai-agents-2nlh

Rao, A. S., and Georgeff, M. P. (1991). Modeling Rational Agents within a BDI-Architecture. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, (473–484).

Rao, A. S., and Georgeff, M. P. (1995). BDI-agents: From Theory to Practice Archived 2011-06-04 at the Wayback Machine. In *Proceedings of the First International Conference on Multiagent Systems (ICMAS"95), San Francisco, 1995*.

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th Edition). Pearson. (AI textbook defining agents and environments; Chapter 2 on intelligent agents).

Searle, J. R. (1969). Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press (Verlag). ISBN 978-0-521-09626-3.

Shoham, y. (1993). Agent-oriented programming. *Artificial Intelligence, Volume 60,* Issue 1, 51-92. ISSN 0004-3702.

Sierra Garcia, C., Botti Navarro, V. J. Ossowski, D. S. (2011). Agreement Computing. *KI - Künstliche Intelligenz, 25*(1), 57-61. https://doi.org/10.1007/s13218-010-0070

Wooldridge, M. (2002). *An Introduction to Multi-Agent Systems.* John Wiley & Sons. (Textbook covering MAS definitions, architectures, and examples).

Wooldridge, M. (2009). *An Introduction to Multiagent Systems* (2nd ed.). Wiley.

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review, 10*(2), 115–152. (Classical paper defining intelligent agents and their properties).