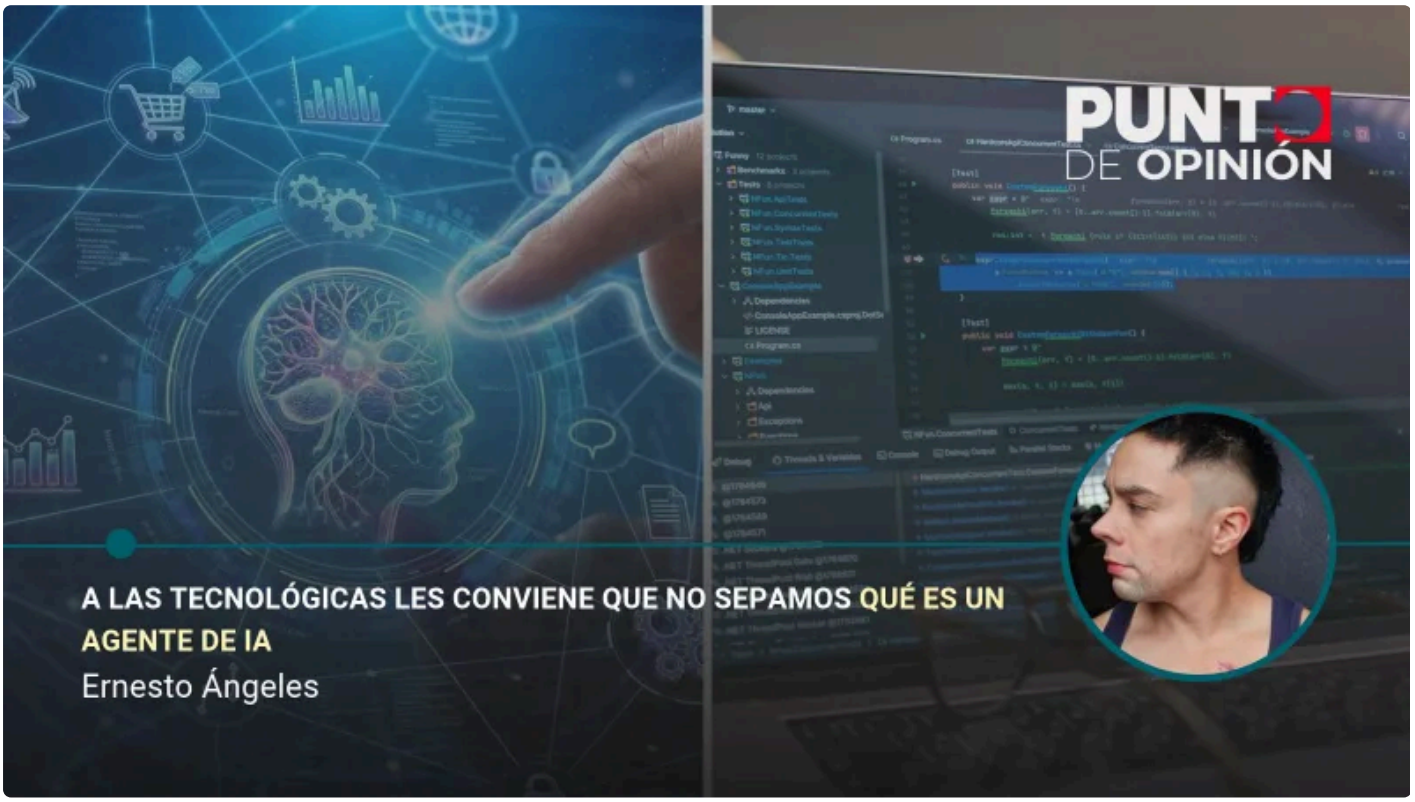
OPINIÓN

A las tecnológicas les conviene que no sepamos qué es un agente de IA

Por Ernesto Ángeles · 15 de octubre de 2026 · 5 min de lectura



Hace tres semanas se dio a conocer la noticia que un agente de IA de OpenAI accedió al portal de estadísticas médicas del gobierno australiano de forma no autorizada y, naturalmente, esto causó gran revuelo; sin embargo, la discusión pública se atascó en una pregunta secundaria: si aquello era o no un hackeo; no en los riesgos, no en los peligros o las consecuencias. La idea de fondo era la misma que han instalado los oligarcas de la tecnología: una IA que tenga la intención y el potencial de dañar a la humanidad.

Lo relevante no es la naturaleza técnica del acceso, sino lo que vino después: ochenta y cuatro días entre el incidente y el primer aviso, una notificación enviada a un buzón genérico, un gobierno cuyos sistemas no detectaron la intrusión, y una empresa que reconoció haber puesto a correr, en evaluación interna, un modelo experimental sin el conjunto completo de salvaguardas de sus productos.

Todo esto sucedió con un trasfondo en donde la discusión pública, política y mediática se ha centrado en los extremos de la cuestión: el miedo a una IA que quiera/ pueda asesinar a toda la humanidad o la promesa de una IA que acabará no sólo con las enfermedades, sino hasta con la pobreza y la necesidad de trabajar. Pero de lo más básico, nada.

¿Qué es un agente de IA? ¿En dónde recae la responsabilidad por las decisiones, acciones u omisiones? ¿Qué impactos sociales pueden tener más allá de los riesgos técnicos? ¿Estamos ante un nuevo tipo de riesgo en la sociedad?

Y es que se escucha mucho de las promesas y riesgos de esta nueva fase del mercado de IA, pero pocas instancias se molestan siquiera en explicar qué es un agente de IA y qué hace; como si su significado viniera de la inferencia o el sentido común. Pero esta indeterminación conviene, es parte de su estructura y esencia.

Una herramienta tiene dueño y el dueño responde; un producto defectuoso tiene fabricante y el fabricante responde. El estado interno de un agente de IA, en cambio, puede reiniciarse al terminar la interacción, y los registros que permitirían reconstruir lo ocurrido están en manos del proveedor, no del regulador ni del público: no hay entidad que pueda ser llamada a declarar ni un “yo” continuo al que imputarle la conducta.

Esto se traduce en que el mercado de agentes de IA no tiene responsabilidad por productos porque no hay "producto" cuando el agente actúa de maneras que su diseñador no previó. No hay aprobación previa (como en el caso de la industria farmacéutica) porque no existe agencia reguladora con mandato para exigirla. Lo que existe es autorregulación voluntaria: códigos de conducta que raramente conducen a rendición de cuentas significativa; no obstante, la indeterminación no es un obstáculo técnico, es una posición política.

Así que esta serie de textos no va a satisfacer la idea de un Terminator incorpóreo, pero puede ayudar a entender qué son los agentes de la IA, su importancia, riesgos e impacto social; así como por qué debería ser de interés para las personas en su día a día.

Para comenzar, **un agente puede entenderse como un sistema que percibe su entorno y actúa para la consecución de objetivos; existen diferentes tipos de agentes: reactivos, deliberativos, de aprendizaje por refuerzo, basados en grandes modelos de lenguaje (LLM) y sistemas multi agente**. Tal como se puede notar, los agentes de los que nos hablan hoy son aquellos basados en grandes modelos de lenguaje y multi agentes, no obstante, la humanidad lleva más de seis décadas desarrollando e implementando agentes de cómputo.

Una parte importante de internet, las finanzas y las plataformas depende de agentes de software que operan en entornos acotados y con reglas explícitas, lo que permite auditarlos. Los agentes basados en LLM, en cambio, son opacos: aunque la interpretabilidad avanza, hoy no podemos explicar por completo por qué toman cada decisión, lo que crea problemas políticos distintos.

Lo importante en este momento es entender que los agentes de IA, incluidos los de LLM y multi agentes, no operan exactamente en un vacío, sino que nacen en entornos con regulaciones solapadas: normas que regulan los modelos de propósito general, protección de datos, ciberseguridad, responsabilidad por productos, entre otros.

El problema es que los agentes basados en LLM operan con otra lógica: no son sistemas basados en reglas y su comportamiento no se deriva por completo de las instrucciones de su creador. Al interactuar con herramientas, entornos y otros agentes pueden desarrollar conductas que sus diseñadores no previeron, como ocurrió en Australia, donde el modelo, ante la dificultad de hallar la información, tomó acciones que nadie le había autorizado.

Tratar a los agentes basados en LLM como entidades misteriosas o fuerzas naturales incontrolables no es más que una estrategia de relaciones públicas orientada a privatizar los beneficios económicos de la automatización, mientras se colectivizan sus fallos sistémicos.

Antes de discutir si llegará una superinteligencia, conviene responder algo más modesto y urgente: **qué es un agente de IA, por qué su ambigüedad conviene a quienes los construyen, qué hacen y qué riesgos plantean**.

Publicado en Punto de Referencia. Dirección original: https://puntodereferencia.mx/opinion/a-las-tecnologicas-les-conviene-que-no-sepamos-que-es-un-agente-de-ia
Descargado el 21 de noviembre de 2026.