

La IA es la máquina más grande jamás construida

Arnau Ramió | Learning Heroes

Transcripción a texto del vídeo del mismo título con ayuda de la IA de Google.

<https://www.youtube.com/watch?v=sKfO1xWOOX0>

Cada vez que le escribes algo a Chat GPT o a Claude, pasan unos 3 segundos hasta que aparece la respuesta. Tan solo 3 segundos. Edo, ¿te imaginas que esa respuesta pues viene de la nube? Viene de algo que flota por internet sin peso, como si fuera magia. Pues mira, te voy a contar lo que acabas de activar de verdad con esa pequeña pregunta que le haces al madero. Acabas de activar una máquina que empieza en una mina en el desierto de Chile, que pasa por un pueblo de Holanda donde se fabrica el aparato más preciso que jamás ha construido nuestra sociedad, que cruza una isla en Asia donde se imprimen casi todos los cerebros digitales del planeta, diseñados justamente por la empresa más grande del mundo, que además se apila en unas torres microscópicas en Corea y que termina en una nave industrial gigante en Texas. consumiendo la electricidad que consumirían ciudades enteras. La inteligencia artificial no es una app, es la máquina más grande que se ha construido en toda la humanidad y hoy te la voy a enseñar entera, pieza a pieza. Vas a entender de una vez por todas qué es la IA por dentro, por qué se está gastando en ellas, pues más dinero que ninguna otra obra de la historia y por qué esa máquina ahora mismo se está empezando a quedar sin piezas.

Empecemos por el principio, porque aquí está el error que comete casi todo el mundo. Cuando la gente dice «inteligencia artificial», piensa básicamente en Chat GPT, en Gemini o en Claude; piensa en el modelo. Sin embargo, el modelo es solo la punta del iceberg. Lo que hay debajo es lo que en la industria se conoce como el stack, es decir, todas las capas de tecnología que tienen que existir para que ese modelo te pueda responder. Te prometo que, cuando lo veas entero, no vas a volver a mirar a la IA de la misma manera. Vamos a bajar por la máquina piso a piso, empezando desde lo más alto.

Piso número 1: El modelo

Es básicamente el cerebro digital: ChatGPT, Gemini, Claude o Grok. Es lo único que tú ves, pero un modelo, en el fondo, es un archivo gigante lleno de números llamados «pesos» que, mágicamente, producen inteligencia. Es algo similar a lo que hacen las neuronas en nuestro cerebro. Pero, claro, un archivo no piensa por sí mismo; necesita que algo lo ejecute.

Piso número 2: El chip

Las famosas GPU de Nvidia. Aunque la realidad es que Nvidia no fabrica los chips: simplemente los diseña, hace los planos. Es el arquitecto, no el constructor. Lo mismo ocurre con los principales competidores del mercado, como AMD, Huawei o Cerebras, e incluso con los grandes como Google o Amazon; todos ellos se encargan principalmente del diseño de sus propios chips.

Piso número 3: La fábrica

Los planos de Nvidia se fabrican casi todos en el mismo sitio: Taiwán. Una empresa llamada TSMC produce en esa isla alrededor del 90% de los chips avanzados del mundo. Es decir, los cerebros de la revolución de la IA salen casi todos de una isla del tamaño de Cataluña. De ahí provienen las fuertes tensiones geopolíticas.

Piso número 4: La máquina que fabrica las máquinas

Aquí es donde la cosa se pone bastante interesante. Para imprimir esos chips, TSMC necesita un aparato que solo sabe fabricar una empresa en todo el planeta: ASML, situada en los Países Bajos. Es una máquina del tamaño de un autobús que cuesta unos 400 millones de dólares la unidad. Dibuja los circuitos con una luz ultravioleta extrema tan precisa que se genera disparando un láser contra gotas de estaño fundido 50 000 veces por segundo. Sus espejos son el objeto más liso jamás fabricado por el ser humano: si uno de ellos fuera del tamaño de Alemania, su mayor imperfección mediría

menos de un milímetro. Es una auténtica locura. Sin esa máquina holandesa no hay chips, y sin chips no hay modelo. Toda la inteligencia artificial del mundo pasa por un cuello de botella del tamaño de una sola empresa.

Piso número 5: La memoria

Aquí ya nos adentramos en el ecosistema que rodea al chip de la IA. El chip calcula, pero los datos tienen que vivir en algún sitio, y solamente hay tres empresas que fabrican prácticamente toda la memoria del planeta: Samsung, SK Hynix y Micron. Quédate con esta capa, porque dentro de un rato volveremos a ella y no te va a gustar lo que está pasando ahí.

Piso número 6: El centro de datos

Todo lo anterior se junta en naves industriales del tamaño de 40 campos de fútbol, con cientos de miles de chips funcionando a la vez día y noche. La IA es la nueva electricidad, y estos centros son como las centrales que generan esa energía. Piensa en ellos como refinerías donde entran electrones y sale inteligencia.

Piso número 7: El mundo físico

La electricidad que alimenta todo esto, el cobre que transporta la energía, el acero, el agua, el suelo y la mano de obra que lo construye. Ese es el stack. Ahora ya puedes ver lo que casi nadie ve: la IA no es solo software, es la cadena industrial más compleja que jamás ha existido, repartida por medio planeta y con varios pisos que dependen de una sola empresa, de una sola isla o de tres fabricantes.

Las dimensiones de la inversión

La pregunta obvia al final es: ¿cuánto cuesta construir algo así a escala mundial? Eso es justo lo que nadie ha terminado de asimilar. Los tres gigantes tecnológicos van a gastar este año unos 785 000 millones de dólares en construir esta máquina. Según Moody's, en el año 2027 la cifra rozará el billón de dólares americanos (un trillón en la nomenclatura anglosajona, lo que equivale a un millón de millones de dólares cada año).

Por su parte, Ark Invest calcula que la inversión total en centros de datos se va a triplicar hacia el año 2030, alcanzando el billón y medio de dólares al año. Para que entiendas la dimensión de esto, vamos a compararlo con las grandes obras de la historia:

- El internet de los años 90 (La burbuja de las .com): Aquella locura que nos hizo cablear el planeta entero con fibra óptica llegó a mover en su pico el 1.5% de la economía estadounidense. La IA ya lo ha superado, situándose en el 2.5% y subiendo.
- Los ferrocarriles del siglo XIX: La obra que construyó América llegó a mover el 6% de toda la economía.
- Grandes infraestructuras: La construcción de las autopistas, las redes de telecomunicaciones para el teléfono móvil o la red eléctrica han sido movimientos masivos. Por encima de eso, en la historia solo quedan hitos como el New Deal o las guerras mundiales. Se prevé que la inversión en inteligencia artificial las podría superar todas.

Si las predicciones se cumplen, vamos camino a niveles de movilización de recursos y capital que la humanidad solo ha alcanzado a ver para construir países enteros o para ir a la guerra. Solo que esta vez no es para ninguna de las dos cosas: es para construir inteligencia.

De hecho, esta obra ya está sosteniendo la economía actual. Solo el gasto en centros de datos explicó la mitad del crecimiento de Estados Unidos en la primera mitad del año 2025. La mitad. Esto, además, marca el final de toda una era. Durante veinte años, la historia de la tecnología era la de dos jóvenes montando un negocio multimillonario desde su casa. Instagram, por ejemplo, se vendió por 1000 millones de dólares teniendo solo 13 empleados. WhatsApp se vendió por 19 000 millones con poco más de 50 empleados. Era una época de código, un portátil y a correr. Esa era ha muerto.

La riqueza ha vuelto a la industria pesada. Ahora el éxito se mide en gigavatios, en toneladas y en años de obra. En la próxima década, crear una empresa tecnológica no será solo escribir código; será mover átomos y miles de millones en inversión.

Los frenos físicos: El desafío energético

Una obra tan agresiva, tan acelerada y tan gigante es imposible que no se encuentre con frenos físicos en el camino. Y esto, que era de esperar, ya está ocurriendo. Vamos a verlos uno por uno, porque ahí es donde está la clave de todo lo que viene.

El primer freno es el más evidente: la **energía**. Los centros de datos del mundo consumen a día de hoy unos 485 TWh (teravatios-hora), y para el 2030 serán 950 TWh, según la Agencia Internacional de la Energía. Es decir, en cinco años vamos a enchufar a la red el equivalente al consumo de Japón entero. La parte de la IA dentro de ese consumo no es que se vaya a duplicar, se va a triplicar. En Estados Unidos, los centros de datos van a representar casi la mitad de todo el crecimiento de la demanda eléctrica de aquí a final de la década. Van a consumir más electricidad que la industria del aluminio, la del acero, la del cemento y la química todas juntas.

El problema es que la red eléctrica no se construye a la misma velocidad que el software. La propia agencia calcula que uno de cada cinco proyectos de centros de datos en el mundo está ahora mismo en riesgo de retraso porque la red no da más de sí. Conseguir un transformador o cables de alta tensión tarda hoy el doble de tiempo que hace tres años; y las turbinas de gas, que son la supuesta solución rápida, tienen años de lista de espera. El resultado es que las empresas están haciendo cosas que hace cinco años sonarían a ciencia ficción. Microsoft, por ejemplo, ha firmado un contrato de 20 años para resucitar un reactor nuclear; ni más ni menos que Three Mile Island, el del famoso accidente de 1979. El día que se anunció el acuerdo, la empresa eléctrica dueña del reactor subió un 22% en bolsa en un solo día. Cuando los gigantes del software se pelean por reactores de los años 70, es que la energía se ha convertido, literalmente, en oro.

Pero ojo, porque ni siquiera este es **el freno más urgente**. El más inmediato está dentro de tu ordenador y lo pagas tú. Tu ordenador tiene dos tipos de **almacenamiento**. En primer lugar, está el **disco duro**, que es como el archivador: es donde lo guardas absolutamente todo, pero acceder a esta memoria es muy lento. Por eso también tenemos la **memoria RAM**, que es como la mesa de trabajo: solo pones lo que estás utilizando ahora mismo y el acceso es instantáneo. Cuanto más grande es tu disco duro, más puedes guardar; cuanta más memoria RAM tienes, más cosas puedes manejar a la vez.

Ahora imagina el cerebro de la inteligencia artificial ejecutado por una GPU que hace billones de operaciones por segundo. Ese cerebro necesita que le sirvan los datos a una velocidad brutal y, por lo tanto, la memoria RAM normal (la de tu portátil) no da abasto. Da igual lo rápido que piense; se pasa el día esperando a que lleguen los datos. Es como poner un Fórmula 1 en una carretera comarcal: no le sacas todo el jugo.

La solución se llama **HBM** (High Bandwidth Memory o memoria de alto ancho de banda), y la idea es preciosa. En vez de poner la memoria lejos del chip, la apilan en torres de forma vertical y la pegan directamente al procesador, conectándola mediante miles de cables a la vez. Han cambiado la carretera comarcal por una autopista de 1000 carriles construida justo al lado del cerebro.

¿Y cuál es el problema? Que fabricar esa memoria apilada es una auténtica pesadilla. Cuando se fabrica memoria normal, de cada 10 chips producidos, más de nueve funcionan bien porque es un proceso maduro. Sin embargo, al apilar 12 o 16 capas para hacer memorias HBM, si una sola capa falla, tienes que tirar toda la torre. El rendimiento se desploma: de cada 10 solo salen bien la mitad o poco más. Además, la HBM consume tres o cuatro veces más obleas de

silicio que la memoria normal, y encima la mitad se desecha por el camino.

La crisis de la memoria y el impacto en tu bolsillo

¿Y qué pasa con esto? Pues que cada oblea que se dedica a la IA es una oblea que no va a tu portátil, a tu móvil o a tu coche. Es un juego de suma cero y solo hay tres empresas en el mundo repartiéndose el 95% del mercado: Samsung, SK Hynix y Micron. Ha pasado lo que tenía que pasar: la obra más grande de la historia ha vaciado la despensa. SK Hynix anunció en octubre que toda su producción de 2026 estaba vendida antes de empezar el año. Micron se encuentra en la misma situación. Goldman Sachs afirma que esta es la mayor escasez de memoria en 15 años. La memoria normal para ordenador se ha encarecido aproximadamente cuatro veces y media desde el pasado verano. Los analistas de TrendForce lo describen como algo sin precedentes. ¿Y quién crees que paga todo esto? Lo pagas tú con un portátil más caro o con un móvil con menos memoria. Ese es el problema.

Pero sobre todo lo pagan los que construyen las máquinas. Mira este dato: en la generación anterior de los servidores de Nvidia, la memoria representaba el 9% del coste total; en la nueva, según Morgan Stanley, la memoria ya es el 26%. Se ha comido más de un cuarto del presupuesto de los chips y le ha quitado terreno hasta a la propia GPU, que ha bajado del 63% al 51%. La inteligencia artificial dejó de ser abstracta el día en que encareció tu portátil o el portátil de tu hijo. Lo mejor viene ahora, porque el propio jefe de Micron en la última presentación de resultados dijo que espera que el mercado siga apretado hasta al menos el 2027 y que ahora mismo no ve el momento en el que la oferta pueda alcanzar a la demanda. Dicho de otra manera: los que fabrican la memoria, que son los que mejor conocen y entienden el sector, no saben cuándo se va a arreglar esto. Y claro, tú te preguntarás: «Oye, ¿y por qué no fabrican más y ya está?». Pues porque una fábrica de memoria cuesta 20.000 millones de dólares y tarda de 2 a 3 años en construirse. No habrá volumen nuevo de verdad hasta el periodo 2027-2028. Apunta bien ese desfase, porque esa es la clave de todo: la demanda corre en meses y la oferta responde en años. Si te parece grave que una pieza tarde 3 años, espera a ver la siguiente.

El sótano de la infraestructura: El cobre

Bajemos al sótano de la máquina, a los materiales; y sobre todo a uno: **el cobre**, que es el sistema circulatorio de toda esta revolución. Es el encargado de los cables que traen la electricidad, los transformadores y la refrigeración. Solo el mazo de conexiones de un rack moderno de Nvidia lleva más de 5000 cables. Aquí te tengo que contar algo que casi nadie te va a contar. A principios de año circuló por todos los titulares que un centro de datos gigante necesitaba ni más ni menos que 500.000 toneladas de cobre. Hubo proyecciones apocalípticas y el pánico estaba servido. Sin embargo, el dato era un error de Nvidia: básicamente un error de unidades en un documento técnico que corrigieron en silencio. La cifra real de ese equipamiento eran unas 200 toneladas; es decir, se equivocaron por un factor de 2500. Te lo cuento porque aquí hay que ver las cosas de forma objetiva. La realidad es que la IA, a día de hoy, apenas representa el 2% de la demanda mundial de cobre. El problema no es la IA en sí; el problema es que llega al mismo tiempo que todo lo demás, porque estamos en la era exponencial. Piénsalo: coches eléctricos que llevan varias veces más cobre que uno de gasolina; paneles solares, aerogeneradores y redes eléctricas totalmente nuevas; el auge de la robótica y aires condicionados para medio planeta. Y ahora, encima de todo eso, sumamos los centros de datos. **S&P Global** ya mete en sus cálculos a los robots humanoides como el siguiente gran consumidor de cobre. Todas las plataformas de la era exponencial están compitiendo por los mismos átomos, por los mismos materiales. El resultado es que, según el estudio de referencia de S&P, la demanda mundial de cobre va a crecer en torno a un 50% de aquí al año 2040. ¿Y qué le pasa a la oferta? Que la oferta hace pico en el 2030 y luego empieza a caer. Tendremos una demanda disparada, una oferta cayendo y un agujero de 10 millones de toneladas en medio. De nuevo te preguntarás: «¿Por qué no abren más minas?». Porque estamos en la misma situación de desfase temporal, pero multiplicada. Un centro de datos se construye en año y medio o dos años.

Una mina de cobre tarda, de media mundial, 17 años desde que se descubre hasta que produce (en Estados Unidos la media es de 29 años). Y encima, el grifo de los descubrimientos se está cerrando: en la última década solo se han encontrado 14 grandes yacimientos, mientras que en los 23 años anteriores se habían encontrado 225. La IA escala en tiempos de software, pero los materiales escalan en tiempo geológico. Por primera vez en la historia, los dos relojes marcan la misma hora: memoria a 3 años, minas a 30 años.

El freno humano: Mano de obra

Pero esto no acaba aquí. Aún queda un freno más, y es el más inesperado de todos porque no se arregla ni con dinero ni con tiempo. Toda esta máquina la construyen y la mantienen personas, y el factor humano se ha convertido en el freno que nadie vio venir. Primero, las manos: la mano de obra. En Estados Unidos se necesitan 300.000 electricistas nuevos esta década y, encima, toca reemplazar a 200.000 que se jubilan. El trabajo eléctrico representa hasta el 70% del coste de construir un centro de datos. El sindicato del sector lo llama, literalmente, una «situación de vida o muerte» para las tecnológicas. Fíjate cómo han cambiado las cosas: después de décadas diciéndoles a los jóvenes que estudiaran en la universidad para acabar en una oficina, hoy los electricistas de veintipocos años están cobrando 250 000 dólares en Texas sin título universitario.

Y, en segundo lugar, tenemos a los vecinos, porque esta obra gigante se construye al lado del entorno de gente real, y esa gente ha empezado a decir que no, que no lo quiere. En 2025, 48 proyectos de centros de datos por un valor de 56.000 millones de dólares fueron bloqueados o paralizados, básicamente por la oposición local en Estados Unidos. Las cancelaciones se han multiplicado por cuatro en un año, y una encuesta de Gallup lo resume todo con mucha claridad: el 71% de los estadounidenses no quiere vivir al lado de un centro de datos. Es un rechazo que incluso supera al que generan las centrales nucleares. Y no es un capricho; al final es pura lógica. La factura de la luz sube en las zonas donde hay centros de datos, y los recursos de agua y suelo también se tensan. Son preocupaciones legítimas. La infraestructura invisible de internet se ha vuelto visible, y la confianza de las comunidades no se compra: se construye. Lo mismo ocurre con la formación de una generación entera de técnicos y electricistas. Puedes imprimir dinero, pero no puedes imprimir electricistas. Energía, memoria, materiales y personas: **cuatro frenos**, cada uno con un reloj más lento que el anterior. Y llegados aquí, la pregunta inevitable es: ¿frena esto la revolución de la inteligencia artificial? Puede sonar extraño, pero no, no la frena (o, al menos, no obligatoriamente). Y hay **tres razones** que lo explican muy bien:

- 1. La eficiencia:** La Agencia Internacional de la Energía afirma que el consumo por tarea de inteligencia artificial está cayendo a un ritmo —y cito textualmente— «sin precedentes en la historia de la energía». La IA es cada día más eficiente y lo vemos todos. Además, las empresas ya están aprendiendo a autoabastecerse: generación en la propia planta, baterías avanzadas y reactores nucleares pequeños (SMR), cuya cartera de acuerdos casi se ha duplicado en tan solo un año.
- 2. El reciclaje:** Casi todo el cobre minado en la historia sigue en circulación y ya cubre un 30% de la demanda global.
- 3. Los incentivos extraordinarios:** Los incentivos económicos y de competitividad para crear la tecnología más importante e impactante de toda la historia son lo suficientemente altos como para que encontremos soluciones extraordinarias a problemas extraordinarios.

Lo importante aquí es entender que en la carrera de la IA existen cuellos de botella que no podemos solventar de un día para otro, generando oportunidades enormes para aquellos que se adelantan y entienden el mapa de lo que está pasando. Y cuidado, porque la eficiencia tiene una trampa. Cuanto más barata es la inteligencia, más se consume. Esto se conoce como la paradoja de Jevons. Cuanto más cae el coste de la IA, esta no se vuelve necesariamente más barata

globalmente, porque ahora simplemente consumimos más. Tenemos más usuarios, más agentes y más demanda total. Así que los frenos no paran la máquina, simplemente ponen precio a cada pieza y, en general, ese precio tiende a aumentar; y cuando hay aumento de precios, hay oportunidades.

El peaje de los embudos financieros

Vuelve a mirar el stack un segundo: el mayor río de dinero de la historia empresarial, camino al billón de dólares al año, pasa entero por **cuatro embudos**:

- Una empresa que diseña prácticamente todos los chips y se ha convertido en la más grande del mundo (Nvidia).
- Una isla que los fabrica (Taiwán).
- Una empresa holandesa que fabrica la máquina que permite imprimirlos (ASML).
- Tres fabricantes que controlan toda la memoria (Samsung, SK Hynix y Micron).

Quien es dueño del embudo cobra un peaje sobre toda la obra entera. Los ingresos de SK Hynix, sin ir más lejos, casi se han triplicado en un año con más margen que la propia Nvidia. La pregunta importante entonces es: ¿quién acabará controlando estos embudos? ¿Será China, que domina la red eléctrica y el refinamiento de los materiales, o Estados Unidos, que domina los chips y el capital?

Conclusión: La verdadera inteligencia colectiva

Para acabar, déjame cerrar con la reflexión que me llevo de todo esto. Hay un ensayo famoso de economía que dice que nadie en el mundo sabe fabricar un lápiz de principio a fin. La madera que viene del bosque, el grafito de la mina, la goma de la plantación; miles de personas que no se conocen, coordinadas sin que nadie las dirija para hacer algo tan simple como un lápiz. Pues ahora multiplica eso por mil millones. El viaje de tu pregunta a ChatGPT toca una mina en Chile, un pueblo en los Países Bajos, unas oficinas en Estados Unidos, una isla en Taiwán, unas torres de memoria en Corea y una nave en Texas. Millones de personas en cinco continentes que jamás se verán las caras, construyendo juntas la misma máquina: la máquina más grande de toda la historia. Llevamos a todo el mundo hablando de replicar la inteligencia humana, pero para construir esta inteligencia artificial ha hecho falta desplegar otra inteligencia que ya teníamos y que casi nunca celebramos: la **inteligencia colectiva**. La capacidad que tenemos de coordinar empresas, países, capital y talento hacia un objetivo común, a una escala que ninguna otra especie puede ni soñar. Queremos construir una máquina que piense como nosotros y, en el intento, estamos demostrando lo extraordinario que es cuando pensamos juntos. Te dijeron que la inteligencia artificial vivía en una nube. Ahora ya sabes que esto no es así. Vive en una máquina de siete pisos repartida por medio planeta, construida por millones de manos con un coste de cientos de miles de millones. Y esa máquina se está montando ahora mismo delante de ti.

Muchas gracias por vuestro apoyo.

COMENTARIO VALORATIVO DEL AUTOR DE ESTA TRANSCRIPCIÓN

A nivel técnico es un magnífico análisis de lo que supone la IA y sus principales retos. Por eso hemos propuesto su publicación, aunque para nada entra en el cuestionamiento del sistema en el que se está desarrollando esta loca carrera cortoplacista. Es una mirada intramuros de un sistema que no se cuestiona. Subraya el mozo su tecnooptimismo sobre todo en la despedida. “Estamos en la ERA EXPONENCIAL”, cierto, pero no se cuestiona en absoluto, todo lo contrario, la imposibilidad de ese crecimiento exponencial indefinido propia del sistema capitalista en un momento de extralimitación a todos los niveles, y para nada entra en la policrisis que define este tiempo que vivimos, marcado por la crisis energética y de materiales (algo se aproxima a esto, pero solo cuestionando los ritmos), la deuda pública desbocada y el desastre medioambiental, que nos ponen a todos al borde del precipicio, con el riesgo cierto de acabar con toda la civilización. No hay peor ciego que el que no quiere ver cuando él mismo nos dice al final que “nos anticipamos a lo que viene”.

Ángel Encinas.