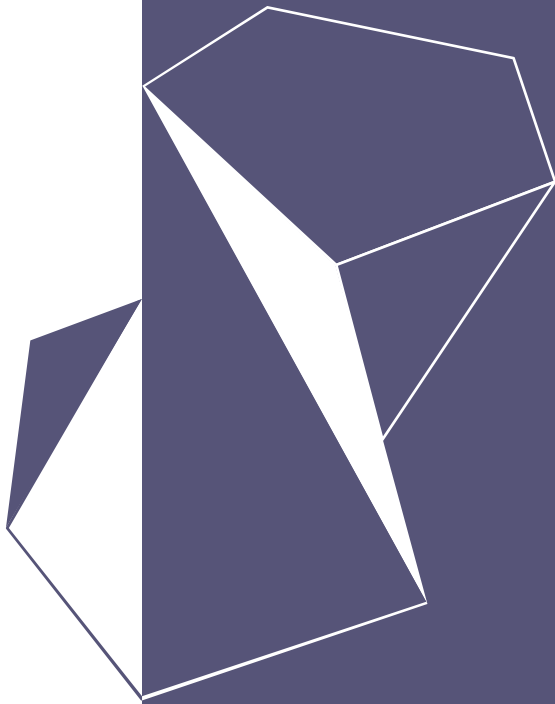


NICK SRNICEK

IMPERIOS DE SILICIO

La lucha por el futuro de la IA



IMPERIOS DE SILICIO

La lucha por el futuro de la IA

Srnicek, Nick
Imperios de silicio. La lucha por el futuro de la IA
1a ed. - Ciudad Autónoma de Buenos Aires: Caja Negra, 2026
256 p.; 20 x 13 cm. - (Futuros próximos; 78)

Traducción de Valentín Huarte y Cecilia Espósito
ISBN 978-987-8272-59-7

1. Inteligencia Artificial. 2. Economía. 3. Geopolítica.
I. Huarte, Valentín, trad. II. Espósito, Cecilia, trad. III.
Título.
CDD 006.3

Título original: *Silicon Empires: The Fight for
the Future of AI*

Esta edición fue publicada bajo acuerdo
con Polity Press Ltd., Cambridge.

© Nick Srnicek, 2026
© Caja Negra Editora, 2026

Caja Negra Editora

Buenos Aires / Argentina
info@cajanegraeditora.com.ar
www.cajanegraeditora.com.ar

Dirección editorial:
Diego Esteras / Ezequiel Fanego
Producción: Malena Rey / Sofía Stel
Coordinación: Candelaria Pera
Diseño de colección y tapa: Consuelo Parga
Maquetación: Cecilia Espósito
Corrección: Juliana Martínez Dios y Candelaria Pera

ÍNDICE

<u>7</u>	Agradecimientos
<u>9</u>	1. ¿DÓNDE ESTÁ EL VALOR DE LA IA GENERATIVA?
<u>13</u>	Más allá de ChatGPT
<u>23</u>	La pila de IA generativa
<u>31</u>	Tensiones y trayectorias
<u>45</u>	Conclusión
<u>55</u>	2. ESTRATEGIAS DE CAPTURA
<u>57</u>	Los desafíos de las GPT
<u>64</u>	Estrategias de captura
<u>68</u>	Amazon y la estrategia de infraestructura
<u>79</u>	OpenAI y la estrategia de frontera
<u>90</u>	Google y la estrategia de conglomerado
<u>103</u>	Meta y la estrategia abierta
<u>112</u>	Conclusión
<u>117</u>	3. UN INTERREGNO
<u>120</u>	El consenso de Silicon Valley
<u>130</u>	El tecnodesarrollismo
<u>144</u>	El fin del consenso
<u>150</u>	El complejo tecnoindustrial
<u>175</u>	La expansión desordenada del capital
<u>189</u>	Conclusión
<u>201</u>	4. EL FUTURO DE LOS IMPERIOS DE SILICIO
<u>205</u>	La estrategia de innovación de Estados Unidos
<u>211</u>	La estrategia de difusión de China
<u>220</u>	Imperios de silicio
<u>250</u>	Conclusión



¿DÓNDE ESTÁ EL VALOR DE LA IA GENERATIVA?

Desde el lanzamiento de ChatGPT a fines de 2022, la inteligencia artificial (IA) generativa ha sido el centro de una atención y un revuelo publicitario intensos. El auge de tecnologías que pueden generar texto, imágenes, audio y video con sentido y parecidos a los que producimos los humanos trajo consigo una serie de promesas descomunales en torno a su potencial. Para sus defensores más ambiciosos, la IA generativa transformará casi todas las industrias, mejorará enormemente la productividad empresarial y disparará el crecimiento económico.¹ Algunas personas sostienen que la IA constituye una revolución tecnológica análoga a la aparición de internet o la electricidad, o incluso algo quizá sin precedentes. Una empresa de inversiones dice que la IA tendrá un impacto en el PBI

1. Jan Hatzius y otros, "The Potentially Large Effects of Artificial Intelligence on Economic Growth", *Goldman Sachs*, 26 de marzo de 2023, disponible en gspublishing.com.

cinco veces mayor al de la máquina de vapor, mientras que otros opinan que “muy posiblemente la IA sea lo más importante –y lo mejor– que haya creado nuestra civilización”.²

Estas eufóricas afirmaciones han sido un rasgo recurrente del discurso vinculado a la IA. Reflejan en parte los imperativos financieros del capital de riesgo y los modelos de negocio de Silicon Valley.³ El revuelo publicitario también deja ver la incertidumbre sobre el futuro de esta tecnología: su potencial es vasto, pero lo suficientemente confuso como para permitir que florezcan todo tipo de relatos insensatos a su alrededor. En retrospectiva, tal vez 2023 haya sido el pico de desanclaje⁴ de este *hype* de la IA generativa. Los límites de esta tecnología seguían siendo lo suficientemente opacos y sus capacidades, sorprendentes. A lo largo de ese año, inversores y compañías colocaron decenas de miles de millones de dólares en el campo.⁵ Se lanzaron cientos de modelos de IA y sobre ellos se desarrollaron miles de variantes.⁶ Teniendo en cuenta solamente presentaciones de resultados financieros del tercer trimestre, las empresas mencionaron el término

2. ARK Invest, “Big Ideas 2024”, 2024, disponible en ark-invest.com; Marc Andreessen, “Why AI Will Save the World”, *Andreessen Horowitz*, 6 de junio de 2023, disponible en a16z.com.

3. Brian Merchant, “AI Generated Business: The Rise of AGI and the Rush to Find a Working Revenue Model”, *AI Now Institute* (blog), 21 de noviembre de 2024, disponible en ainowinstitute.org.

4. *Unmoored peak*: metáfora de Srnicek para describir el clímax de la burbuja de la IA en 2023, cuando las expectativas perdieron su base material; esto es, la IA está “desanclada” de la base que debería sostenerla, las promesas estratosféricas se separaron de sus capacidades reales y limitaciones técnicas. [N. de T.]

5. Cade Metz, Karen Weise y Tripp Mickle, “AI Start-Ups Face a Rough Financial Reality Check”, *The New York Times*, 29 de abril de 2024, disponible en nytimes.com.

6. El repositorio de Hugging Face ofrece cientos de miles de modelos de aprendizaje automático diferentes, muchos de los cuales funcionaron durante los últimos años como modelos generativos de IA.

“IA” más de 35 mil veces.⁷ Y las cadenas de suministro de la IA se convirtieron en un nuevo foco de intenso interés geopolítico.

En vista de estos acontecimientos espectaculares, este libro pretende ofrecer una introducción, un pantallazo, y una intervención en nuestra manera de comprender cómo se está desarrollando hoy la IA. El libro procura ser accesible para quienes sean nuevos en estos temas o busquen alcanzar una perspectiva sistémica sobre el paisaje de la IA contemporánea; pero también tiene la intención de intervenir en debates actuales, desentrañar las prácticas corporativas, las alianzas de clase, las estrategias geopolíticas. Este análisis puede considerarse deflacionario en el sentido de que busca evitar los excesos tanto de *hype* como de pesimismo fatalista que asedian las discusiones en torno a esta tecnología. Este no es un libro enfocado en la inteligencia artificial general (AGI, por sus siglas en inglés) ni en el riesgo existencial.⁸ Tampoco trata sobre inquietudes más prácticas e inmediatas relacionadas con la automatización, el sesgo, la explotación de los trabajadores, o el acervo colectivo del conocimiento humano.⁹ Este libro trata, en cambio, sobre la economía geopolítica de la IA contemporánea; esto es, acerca de los modos en que las grandes corporaciones y las principales potencias buscan controlar el desarrollo de esta tecnología, capturar el valor que genera y ejercer el poder que contiene.¹⁰

7. Accenture, *Technology Vision 2024: Human by Design*, 2024, p. 7, disponible en [accenture.com](https://www.accenture.com).

8. Meredith Ringel Morris y otros, “Levels of AGI for Operationalizing Progress on the Path to AGI”, *arXiv*, 22 de mayo de 2024, disponible en arxiv.org; Thomas Moynihan, *X-Risk: How Humanity Discovered Its Own Extinction*, Falmouth, Urbanomic, 2020.

9. James Muldoon, Mark Graham y Callum Cant, *Feeding the Machine: The Hidden Human Labour Powering AI*, Edimburgo, Canongate, 2024.

10. Conviene aclarar explícitamente que muchos otros actores de la economía geopolítica de la IA quedan, en gran medida, fuera de este libro: desde los esfuerzos internacionales de regulación y los actores de la sociedad

En síntesis, es un libro sobre los imperios de silencio que actualmente controlan el futuro de la IA. En las cimas del poder corporativo y estatal se están tomando decisiones de amplio alcance sobre el desarrollo de esta tecnología, los valores que encarna y los usos que se harán de ella.¹¹ Estas decisiones están en manos de una pequeña élite, pero afectan a miles de millones de personas. Y a medida que aumentan las capacidades de la IA, la divergencia entre la escala de quienes toman las decisiones y la escala de los impactos no hará más que crecer. La IA será una tecnología transformadora, y esto vuelve todavía más imprescindible comprender quiénes están controlando el futuro de la IA, cómo lo están controlando y con qué fines.

civil que defienden sus posturas, hasta los trabajadores que luchan al filo de la transformación del trabajo causada por la IA. Estos grupos en ocasiones pueden ser importantes y realizar aportes significativos, pero un supuesto de este libro es que la historia central de la IA contemporánea puede ser contada haciendo foco en las grandes corporaciones tecnológicas y en las dos principales potencias geopolíticas.

11. Una buena parte de esto se circunscribe al denominado “problema de la alineación” –o, más específicamente, el “problema de la alineación de valores”– que concierne a la forma de garantizar que los sistemas de IA actúen de maneras seguras y beneficiosas. El abordaje normativo que propone este libro apuntará a la pregunta sobre cuáles son los valores con los que debería alinearse la IA y quién toma esas decisiones. Igualmente importantes, aunque no discutidas aquí, son las cuestiones técnicas que atañen a los modos de alinear los sistemas de IA e incluso a la pregunta de si puede alineárselos en absoluto. Ver Stuart J. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*, Nueva York, Viking, 2019; Iason Gabriel, “Artificial Intelligence, Values, and Alignment”, *Minds and Machines*, vol. 30, n° 30, 2020, pp. 411-437; Andy Zou y otros, “Universal and Transferable Adversarial Attacks on Aligned Language Models”, *arXiv*, 20 de diciembre de 2023; Dan Hendrycks y otros, “Aligning AI With Shared Human Values”, *arXiv*, 17 de febrero de 2023; Evan Hubinger y otros, “Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training”, *arXiv*, 12 de enero de 2024; Ryan Greenblatt y otros, “Alignment Faking in Large Language Models”, *arXiv*, 20 de diciembre de 2024.

MÁS ALLÁ DE CHATGPT

Pero ¿qué se entiende por inteligencia artificial? Después de todo, hay una larga historia de trabajos asociados con esta expresión. Se sabe que el término fue acuñado en 1956 –elegido con fines publicitarios y usado para distinguir una propuesta de financiamiento académico del campo, por aquellos tiempos más conocido, de la cibernética–, pero desde entonces el tipo de tecnologías que responden a esta denominación se expandió significativamente. En el período actual, el *deep learning* o aprendizaje profundo se ha consolidado como el enfoque dominante. Estos sistemas de IA se basan en redes neuronales –modeladas muy vagamente a la manera de neuronas del cerebro humano–, y el aprendizaje “profundo” remite al hecho de que estas “neuronas” se organizan en múltiples capas. El proceso de “aprendizaje” se realiza de varias maneras, aunque el enfoque canónico es el aprendizaje supervisado; en este, el modelo genera una salida o resultado específico [*output*] a partir de una entrada determinada [*input*], y dicha salida se compara con el resultado real o valor objetivo. ¿En qué medida, por ejemplo, un modelo etiqueta o clasifica correctamente imágenes de gatos como “gatos”? Las desviaciones respecto del resultado correcto se utilizan para determinar cómo modificar los pesos o parámetros de las neuronas individuales en un esfuerzo de aproximar el modelo al resultado deseado (mediante un algoritmo denominado *backpropagation* [retropropagación]). Este proceso se repite muchísimas veces hasta que el modelo mejora gradualmente su capacidad para generar resultados correctos. Un modelo que ha pasado por este “entrenamiento” puede entonces ejecutarse con nuevas entradas. El proceso computacional de ejecutar un modelo de IA, comúnmente denominado “inferencia”, es un proceso distinto que tiene sus propias implicaciones técnicas, y, como veremos, políticas y económicas.

Este enfoque para desarrollar inteligencia artificial se conocía desde hacía muchos años, pero en general no había logrado los resultados esperados. Esta situación persistió hasta 2012, cuando un equipo aplicó un enfoque novedoso en ImageNet, una competencia anual de reconocimiento de imágenes. Mientras que en años anteriores las tasas de error se situaban entre el 25-30%, este nuevo método basado en el *deep learning* logró una tasa sin precedentes del 15%.¹² El resultado fue sorprendente y funcionó como una prueba de concepto: demostró que, gracias a los ingentes volúmenes de datos derivados de la digitalización del mundo y a los avances en la capacidad de procesamiento –especialmente en las unidades de procesamiento gráfico o GPU–, los algoritmos de aprendizaje profundo, ya establecidos desde hacía tiempo, podían ser finalmente efectivos.

El enfoque que adoptaron estos autores se conoció como “red neuronal convolucional” (CNN, por sus siglas en inglés): una arquitectura o una forma particular de organizar la red neuronal. Resultó que las redes neuronales convolucionales eran especialmente aptas para el reconocimiento de imágenes y las tareas de visión artificial, con una organización similar, a grandes rasgos, a la manera en que se cree que el cerebro procesa las imágenes: mediante capas de neuronas que analizan características visuales cada vez más detalladas. No obstante, desde 2012, se produjo un florecimiento de la investigación sobre otras arquitecturas y, en algunos casos, se lograron grandes avances gracias al desarrollo de nuevos enfoques.

El avance más significativo en este sentido ha sido, con creces, la arquitectura *transformer*.¹³ Los *transformers*

12. Alex Krizhevsky, Ilya Sutskever y Geoffrey E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, *Communications of the ACM*, vol. 60, n° 6, 2017, pp. 84-90.

13. El resumen de los grandes modelos de lenguaje se atiene aquí a lo indispensable, pero en estos artículos hay explicaciones útiles que requieren

se propusieron en 2017 para superar algunas limitaciones que otras arquitecturas –especialmente, las redes neuronales recurrentes (RNN, por sus siglas en inglés) y las redes de memoria a corto y largo plazo (LSTM)– presentaban al abordar tareas relacionadas con el procesamiento de lenguaje y la traducción.¹⁴ Uno de los avances más importantes de los *transformers* fue que permitieron el cómputo paralelo, en vez de secuencial, de gran parte del modelo. Esto permitió utilizar conjuntos de datos [*datasets*] mucho más amplios en el entrenamiento y dio lugar al surgimiento de los modelos masivos característicos de este período.¹⁵ La arquitectura *transformer* ha generado, a su vez, los conceptos superpuestos de IA generativa, grandes modelos de lenguaje (LLM) y modelos fundacionales.¹⁶ Este enfoque también fue la base de los impresionantes avances recientes en *machine learning* o aprendizaje automático. ChatGPT es el más conocido de estos sistemas, pero la arquitectura *transformer* resultó ser increíblemente versátil y encontró aplicaciones que trascienden en gran medida las tareas de lenguaje para las que fue originalmente desarrollada. Ahora, su uso domina el mundo de la IA.¹⁷

poco conocimiento técnico: Stephen Wolfram, “What Is ChatGPT Doing... and Why Does It Work?”, *Stephen Wolfram Writings* (blog), 14 de febrero de 2023, disponible en writings.stephenwolfram.com; Grant Sanderson, “Neural Networks”, *3Blue1Brown*, disponible en 3blue1brown.com; “Deep Dive into LLMs like ChatGPT”, 2025, disponible en youtube.com.

14. Ashish Vaswani y otros, “Attention Is All You Need”, *arXiv*, 12 de junio de 2017.

15. Dieuwertje Luitse y Wiebke Denkena, “The Great Transformer: Examining the Role of Large Language Models in the Political Economy of AI”, *Big Data & Society*, vol. 8, n° 2, 2021, p. 7.

16. Los modelos fundacionales o *foundation models* son modelos masivos entrenados en conjuntos de datos o *datasets* amplios que sirven de base para muy diversas tareas y aplicaciones. [N. de T.]

17. En 2024, el 74% de los artículos de investigación usó la arquitectura *transformer*. Air Street Capital, “State of AI Report 2024”, 2024, p. 28, disponible en stateof.ai.

Hoy en día, al menos en el imaginario popular, ChatGPT se ha convertido en el paradigma de lo que es la IA: un producto en el cual tanto defensores como críticos pueden proyectar sus esperanzas e inquietudes. En su superficie, ChatGPT es una interfaz sencilla: un cuadro de texto a través del cual los usuarios pueden preguntar cualquier cosa y obtener del sistema respuestas que, a menudo, están sorprendentemente bien estructuradas. Sin embargo, a pesar de su protagonismo, ChatGPT es un representante engañoso de la IA contemporánea. Se asume con demasiada frecuencia que las penurias y éxitos de ChatGPT reflejan directamente el espacio más amplio de la IA. En el mismo sentido, las críticas y defensas de ChatGPT adquieren un alcance mayor del que quizá merecen. La IA generativa –por no hablar de la categoría más amplia del aprendizaje profundo moderno– es mucho más vasta que el simple formato de chatbot representado emblemáticamente por ChatGPT. Por ejemplo, ChatGPT induce al error al presentarse como un producto centrado en el consumidor (aunque, como veremos más adelante, esto está cambiando), mientras que la mayor parte de la IA se está diseñando para fines empresariales. ChatGPT también sugiere que los chatbots son la interfaz principal de la IA, cuando de hecho se están desarrollando muchos otros enfoques. Los *transformers* también tienen usos que van más allá de la mera producción de texto, o incluso imágenes, audio o video. A nivel técnico, esto se debe en parte a que estos LLM no predicen la siguiente palabra en una oración, sino el siguiente “token”. Los tokens pueden representar mucho más que componentes de lenguaje; también pueden representar acciones, píxeles, moléculas, entre otras cosas. La naturaleza de predicción de tokens de los *transformers* es, por lo tanto, mucho más general que la mera predicción lingüística. Como resultado, la IA generativa está encontrando usos en una amplia gama de industrias diversas –como la publicidad, la logística, las finanzas, las ventas,

la salud, el desarrollo de medicamentos, la minería, la manufactura, la robótica, el diseño de hardware- y de formas que tienen poco que ver con la imagen pública de los chatbots.

También se ha producido una serie de avances técnicos que están cambiando lo que la IA puede hacer y esto la lleva más allá de lo que eran capaces las versiones iniciales de ChatGPT. Un primer desarrollo fue la inclusión de la multimodalidad en estos sistemas. Mientras que ChatGPT podía tomar texto y devolver texto, los modelos más recientes pueden tomar una variedad de medios diferentes (por ejemplo, texto, audio y video) y responder con una diversidad de medios similar. Hasta hace no mucho tiempo, esta multimodalidad solía adherirse un poco torpemente a los LLM basados en texto. Pedirles a las primeras versiones de ChatGPT que crearan una imagen implicaba que recurrieran a un modelo separado (DALL-E) que realizaba la tarea. Pero hoy los modelos están volviéndose intrínsecamente multimodales: son capaces de usar todos estos medios diferentes al interior del mismo modelo. Esto se traduce en una menor latencia (y, por ende, en una mayor usabilidad): permite evitar los cuellos de botella de información que surgían cuando las peticiones debían transferirse entre modelos y optimiza el rendimiento, dado que estos tienen acceso a una diversidad de modalidades diferentes. La multimodalidad también significa que estos modelos pueden captar datos visuales del entorno, una capacidad que resulta fundamental tanto para la robótica como para el aprendizaje humano.

Los modelos más recientes también han adquirido nuevas formas de memoria y han aumentado la capacidad de los tipos de memoria ya existentes. En un vago paralelo con nuestras propias mentes, los modelos de IA han obtenido una suerte de memoria a largo plazo en forma de “generación aumentada por recuperación” (RAG) y las bases de datos vectoriales, y algo semejante a la memoria

a corto plazo mediante la ampliación del tamaño de las “ventanas de contexto” y de una creciente personalización basada en los “recuerdos” de los usuarios. La memoria a largo plazo permite que los modelos de IA recurran a un acervo de conocimientos establecidos para intentar reducir su tendencia a generar “alucinaciones” o falsedades. Piénsese, por ejemplo, en una empresa que hace que un modelo cite documentos internos para que la IA no proporcione respuestas erróneas sobre las políticas de la compañía. Por el contrario, la memoria a más corto plazo de las ventanas de contexto permite introducir un mayor volumen de información en un *prompt* de IA y que el sistema considere un rango mucho más amplio de información específica al momento de generar una respuesta. Un estudiante, por ejemplo, puede adjuntar el archivo PDF de un libro de 500 páginas y pedirle al sistema que considere la totalidad del texto. O bien un inspector de seguridad e higiene puede adjuntar el video de la visita a una obra en construcción y hacer que el modelo de IA lo examine en busca de infracciones a las normas de seguridad. También están surgiendo nuevas formas de memoria personalizada que prometen hacer que los modelos reflejen mejor el contexto personal de cada usuario (aunque esto plantee en el fondo inminentes conflictos en torno a la privacidad). Los avances en todas estas áreas vinculadas con la memoria están poniendo de manifiesto una diversidad de nuevos casos de uso que van mucho más allá de la primera versión del chatbot de ChatGPT.

En 2024, el más notable de los avances en IA se produjo cerca de fin de año, cuando OpenAI lanzó su modelo o1 y presentó un adelanto de su modelo o3.¹⁸ Ambos modelos representaron saltos significativos en algunas de

18. El modelo o3 fue el sucesor del modelo o1. El nombre o2 se saltó para evitar posibles problemas vinculados con la empresa de telecomunicaciones del Reino Unido, llamada O2.

las pruebas de rendimiento más exigentes diseñadas para medir el desempeño de la IA. La clave de este progreso fue que permitió que los modelos dedicaran más tiempo a la inferencia. En otras palabras, en vez de intentar producir una respuesta lo más rápido posible, estos modelos pudieron dedicar tiempo a “pensar” el problema: en pasos más pequeños, probar diferentes respuestas y verificar sus resultados. Este cambio hacia una mayor inversión en capacidad computacional en la inferencia –denominado “cómputo en tiempo de ejecución” [*test-time compute*]– forma parte de una tendencia más amplia de los investigadores por intentar añadir capacidades de “razonamiento” a estos modelos. Los métodos empleados para generar este proceso de razonamiento en los modelos parecen ser bastante simples: se usa un aprendizaje por refuerzo básico para recompensarlos cuando obtienen respuestas correctas.¹⁹ A través de un proceso de ensayo y error, los modelos comienzan a aprender a dar pasos similares al razonamiento para abordar los problemas.

En cierta medida, estos modelos reflejan procesos de pensamiento humanos (lo cual no es de extrañar, dado que fueron entrenados en lenguaje humano). Cuando un proceso de razonamiento se hace visible, los modelos pueden parecer sorprendentemente humanos. En un caso, un modelo pareció tener una revelación mientras razonaba sobre un problema, y afirmó: “Espera, espera. Espera. Este es un momento eureka que puedo marcar aquí”.²⁰ Pero la analogía no debería llevarse demasiado lejos, ya que en general a los sistemas orientados al público se les brinda entrenamiento adicional para que resulten más inteligibles para

19. Este enfoque se basa en el modelo R1 de DeepSeek, que al momento en el que escribo este libro es el más transparente en lo que respecta a la manera en que generó un modelo de razonamiento.

20. “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning”, 2025, p. 9, disponible en github.com.

las personas. Antes de este proceso, los modelos pueden ser mucho más difíciles de interpretar, como, por ejemplo, al cambiar de idioma de forma rutinaria mientras “piensan”. Además, algunos investigadores sugieren que el acto mismo de convertir los pasos del proceso de razonamiento en tokens –es decir, en algo legible para los humanos– constituye una limitación para ese pensamiento, y que los modelos futuros podrían ser mucho más inescrutables.²¹ Así, entonces, aunque la arquitectura *transformer* es, en esencia, un predictor del siguiente token, ese sencillo mecanismo sigue generando resultados sorprendentes.

El giro hacia la capacidad computacional para la inferencia también constituye otra línea de investigación cada vez más importante en la IA: los “agentes”. Mientras que la experiencia paradigmática de la IA generativa es hacerle una pregunta a ChatGPT y obtener una respuesta, los agentes son sistemas mucho más elaborados capaces de realizar acciones con relativa autonomía. En lugar de limitarse a devolver una respuesta, estos sistemas establecen un plan y ejecutan una secuencia de tareas para alcanzar un objetivo fijado por el usuario.²² En tanto ChatGPT se centra en responderle al usuario lo más rápido posible, los agentes se enfocan en tomarse el tiempo necesario para cumplir una meta. Si bien el interés inicial por la IA generativa estaba puesto en la fase de preentrenamiento mediante el aprendizaje autosupervisado, los agentes están recuperando la investigación sobre el aprendizaje por refuerzo y el diseño de recompensas eficaces para los agentes dentro de determinados entornos. Lo decisivo es que la utilidad potencial de los agentes en las empresas es mucho mayor que la de un simple chatbot. Más que un mero

21. Shibo Hao y otros, “Training Large Language Models to Reason in a Continuous Latent Space”, *arXiv*, 9 de diciembre de 2024.

22. Iason Gabriel y otros, “The Ethics of Advanced AI Assistants”, *arXiv*, 2024, p. 15. Debe notarse que los agentes como se los comprende aquí excluyen a las IA que pudieran establecer sus propios objetivos.

oráculo al cual consultar, los agentes pueden, de hecho, ejecutar acciones para lograr objetivos. Con los agentes, los sistemas de IA podrían convertirse progresivamente en mediadores entre los usuarios y el resto de nuestro ecosistema digital. Algún día podríamos explicarle lo que queremos a nuestro agente personalizado, y este simplemente iría a hacerlo (un ejemplo típico hoy es planificar unas vacaciones, pero esto resulta banal comparado con lo que podría llegar a ser posible). Ya se está trabajando para reconfigurar el ecosistema digital para los agentes; por ejemplo, dotándolos de los medios para acceder a aplicaciones y herramientas comunes. El potencial comercial de este desarrollo es significativo, lo que ayuda a explicar por qué los grupos de investigación de IA líderes están centrados en construirlos. Muchas empresas también están trabajando en sistemas multiagente para permitir que surjan dinámicas colectivas que resuelvan consultas y ejecuten proyectos.²³

Mientras que en el discurso público la atención suele centrarse en los modelos individuales, se observa un giro creciente hacia los sistemas compuestos: “Un sistema que aborda tareas de IA utilizando múltiples componentes que interactúan entre sí, lo que incluye múltiples llamadas a modelos, recuperadores o herramientas externas”.²⁴ Los agentes son un buen ejemplo de esto: un LLM puede proveer la interfaz de usuario para el sistema, pero el agente podría recurrir luego a una multitud de herramientas diferentes, como motores de búsqueda, depuradores de código, calculadoras, entre otras. Del mismo modo, dados los costos de ejecutar modelos demasiado grandes, muchas empresas están recurriendo a sistemas que combinan

23. Por el momento, las implicancias morales y de seguridad de los sistemas multiagente no se encuentran demasiado desarrolladas.

24. Matei Zaharia y otros, “The Shift from Models to Compound AI Systems”, *The Berkeley Artificial Intelligence Research Blog*, 18 de febrero de 2024, disponible en bair.berkeley.edu.

varios modelos y dirigen automáticamente las peticiones de los usuarios hacia el más adecuado.²⁵ Las consultas más sencillas se envían a modelos más pequeños y económicos, mientras que solo las peticiones más complejas terminan enviándose a los modelos más grandes y costosos. Microsoft, por ejemplo, ha utilizado una versión de este sistema con Bing.²⁶ Nuevamente, lo significativo aquí no es un modelo individual determinado, sino el sistema completo que opera en segundo plano. Gran parte de la investigación contemporánea en IA consiste simplemente en construir tales sistemas para desbloquear y mejorar las capacidades de los modelos existentes. Como resume un artículo de investigación de Google: “Dotar a los modelos de herramientas potentes potencia sus habilidades”.²⁷ Así, aunque una buena parte de la atención mediática haya estado puesta en el aumento de la fuerza bruta computacional –mejoramiento de la IA mediante escalamiento [*scaling*]–, la noción de sistemas muestra que pueden producirse mejoras considerables a través del andamiaje que rodea un modelo.

Por lo tanto, si bien ChatGPT capturó el imaginario del público, este libro hablará de la IA en un sentido más amplio. Empresas, no consumidores; tokens, no texto; multimodalidad, no lenguaje; razonamiento, no repetición mecánica; agentes, no respuestas; y sistemas, no modelos: estas son solo algunas de las formas en que debemos considerar la IA más allá de ChatGPT. También conviene señalar la investigación en curso sobre diferentes arquitecturas

25. Por ejemplo, Isaac Ong y otros, “RouteLLM: An Open-Source Framework for Cost-Effective LLM Routing”, *LMSYS Org*, 1º de julio de 2024, disponible en lmsys.org.

26. Aaron Holmes, “How Microsoft is Trying to Lessen Its Addiction to OpenAI as AI Costs Soar”, *The Information*, 26 de septiembre de 2023, disponible en theinformation.com.

27. Sergei Glazunov y Mark Brand, “Project Naptime: Evaluating Offensive Security Capabilities of Large Language Models”, *Project Zero* (blog), 20 de junio de 2024, disponible en googleprojectzero.blogspot.com.

–modelos de espacio de estados, redes neuronales líquidas, modelos xLSTM, modelos Titan, entre otros– en un intento por superar algunas de las limitaciones del *transformer*.²⁸ Los modelos de difusión y los modelos de mundo, en particular, parecen ser vías prometedoras para superar algunos de los límites de la arquitectura *transformer*. Todo esto significa que los chatbots son un indicador deficiente para señalar hacia dónde se dirige la IA; y tanto sus críticos como sus defensores deberían asegurarse de que tienen en mente el objetivo correcto.

LA PILA DE IA GENERATIVA

Ahora que tenemos una idea de lo que significa la IA contemporánea, ¿cuál es el panorama desde una perspectiva económica? Un vistazo rápido a las noticias financieras revela una serie confusa y compleja de start-ups, empresas establecidas que se transforman rápidamente, nuevos avances técnicos y el lanzamiento casi diario de productos novedosos. ¿Cómo podemos dar sentido a esta diversidad de elementos en medio de un cambio tecnológico tan intenso y veloz? Si pensamos en términos de una pila o estructura vertical [*stack*], podemos situar el entorno de la IA generativa organizado en cuatro capas diferentes: de abajo hacia arriba, estas son el hardware, la infraestructura, los modelos y las aplicaciones. Algunas empresas ocupan solamente una de estas capas, mientras que otras están mucho más integradas verticalmente.

28. Albert Gu y Tri Dao, “Mamba: Linear-Time Sequence Modeling with Selective State Spaces”, *arXiv*, 31 de mayo de 2024; Maximilian Beck y otros, “xLSTM: Extended Long Short-Term Memory”, *arXiv*, 6 de diciembre de 2024; Ali Behrouz, Peilin Zhong y Vahab Mirrokni, “Titans: Learning to Memorize at Test Time”, *arXiv*, 31 de diciembre de 2024.