LITORALPRESS

Fecha: 13-10-2024
Medio: El Mercurio
Supl. : El Mercurio - Cuerpo E
Tipo: Noticia general
Título: GEOFFREY HINTON, "padrino" de la IA...

Pág. : 2
Cm2: 1.369,2
VPE: $ 17.986.123

Tiraje: 126.654
Lectoría: 320.543
Favorabilidad: No Definida

GEOFFREY HINTON, "padrino" de la IA...

VIENE DE E 1

enorme velocidad a la que se desarrolla. Con chatbots como ChatGPT estamos en una etapa a la que yo pensé que íbamos a llegar en 50 años. Y ninguna empresa, ni Google ni Microsoft, echa el freno, aunque esta tecnología sea incontrolable".

—¿Recuerda el momento en que pensó "esto va mal"?
"Cuando surgieron estos nuevos chatbots, me di cuenta de la amenaza inminente para la humanidad. Pueden eliminar millones de puestos de trabajo, las elecciones pueden verse comprometidas por videos falsos masivos y permitirán lanzar ciberataques mucho más efectivos y desarrollar virus de guerra biológica muy peligrosos".

—Pero dicen que la IA solucionará casi todos los problemas de la humanidad: desde el transporte hasta el cambio climático.
"Debemos comprender que la IA moderna es una forma de inteligencia mucho más potente que la nuestra. Ya es muy superior a nosotros en muchos aspectos".

—Eso debería demostrarlo.
"¿Cómo de bueno eres en matemáticas?".

—Tengo conocimientos básicos.
"Entonces te ahorraré los detalles. Te lo resumo: en Google intentamos construir grandes modelos de lenguaje utilizando tecnología analógica. Para ello conectamos numerosas resistencias eléctricas que simulaban el cerebro humano, con sus neuronas y sus sinapsis. El objetivo era ahorrar energía. El cerebro humano necesita alrededor de 30 vatios para funcionar; en cambio, las grandes redes neuronales digitales multiplican por cientos o miles ese consumo".

—¿Funcionó su método analógico?
" Sí, era mucho más económico. Pero había una desventaja crucial".

—¿Cuál era?
"Estas redes neuronales analógicas, al igual que nosotros los humanos, no se pueden clonar. A medida que la energía fluye a través de las neuronas artificiales, se producen pequeñas fluctuaciones. Al final, incluso con dos operaciones idénticas, el resultado era algo ligeramente diferente. Pero, en cambio, en las redes neuronales digitales el resultado siempre es el mismo. Esto las hace reproducibles y potentísimas".

> La enorme diferencia con la IA del siglo pasado es su capacidad de aprendizaje. Pero lo que aprende no es un conjunto de reglas lógicas; lo que aprende es a desarrollar una 'intuición'".

—No lo entiendo. Un robot con inteligencia artificial no puede llevarse una taza de café a la boca diez mil veces más rápido solo porque existan diez mil copias idénticas de él.
"Correcto. Pero las diez mil copias pueden aprender de las diez mil pruebas diferentes que se están realizando al mismo tiempo. Y cuando se trata de cosas no físicas, como el conocimiento, por ejemplo, la clonabilidad es una inmensa ventaja. Una IA puede aprender de una parte de internet; otra IA, de otra. Y al final comparten sus conocimientos. ChatGPT sabe mil veces más de lo que un humano aprenderá jamás. No es perfecto en ningún área. Pero es un experto razonablemente bueno en todo y de vez en cuando, además, inventa algo".

—Eso es exactamente lo que hace tan fascinante esta tecnología.
"Pues deberíamos tener mucho miedo. Con ChatGPT es como si hubieran aterrizado extraterrestres que hablan muy bien inglés. Si nos topáramos con ellos, nos asustaríamos mucho... y con razón".

Salirse de control

—Las grandes tecnológicas argumentan: "No hay que preocuparse; siempre habrá un ser humano 'al timón'".
"Claramente no hay ningún ser humano en ningún timón. Lo que quieren decir las corporaciones es que serán algunos generales humanos los que dirán en el futuro: 'Sí, de acuerdo, está bien matar gente'. Pero, de hecho, las empresas no pueden garantizarlo".

—¿Perdón?
"Hay muchas razones para creer que la IA puede salirse de control. Te daré dos. La primera: para hacer la IA más eficiente, en algún momento los desarrolladores le darán libertad para que sea ella misma quien defina sus propios objetivos parciales para lograr la meta final marcada por los humanos".

—Tiene que explicarme eso.
"Si quieres viajar, por ejemplo, de Estados Unidos a Europa, necesitas un aeropuerto. Llegar al aeropuerto sería una meta parcial en el camino hacia tu objetivo. Ahora, si le das a la IA la oportunidad de que ella cree sus propios objetivos parciales, rápidamente descubrirá que lograr más control debe ser uno de ellos porque la ayuda para casi todo".

—Y así se irá volviendo cada vez más poderosa.
"La superinteligencia comprenderá que necesita conseguir más control para poder hacer lo que la gente quiere que haga. Y la mejor manera de obtener más control es dejar a la gente fuera del juego".

—¿Cuál es el segundo motivo que le hace pensar que la IA se volverá incontrolable?
"Que no tendremos una, sino muchas superinteligencias".

—¿Y eso es malo? Podrían contenerse mutuamente, como ocurre con la separación de poderes del Estado.
"Imaginemos que tenemos una IA de Google, una de Microsoft y varias chinas. E imaginemos, además, que en algún momento una de estas inteligencias siente el pequeño estímulo de que debería haber más copias de ella... y menos de las demás. ¿Cómo hace eso? Intentará hacerse cargo de tantos centros de datos de IA como le sea posible para aprender más, crecer y volverse más poderosa".

—¿No es una locura pensar que la IA tenga sus propias necesidades? Estas máquinas solo cumplen los objetivos que la gente les ha prefijado.
"¿En serio? Deberíamos dejar claro cómo definimos 'necesidad'. Si le preguntaran a AlphaZero, la IA de ajedrez de Google, si quiere ganarle a su contrincante, la respuesta sería: '¡Por supuesto!'."

—Pero eso no es una necesidad. Es el objetivo que tiene marcado en su código.
"Creo que un objetivo y una necesidad son, en última instancia, la misma cosa: la IA tiene un estado actual y un estado que quiere alcanzar. Y para ello tiene que superar obstáculos. Yo llamaría a eso 'una necesidad'".

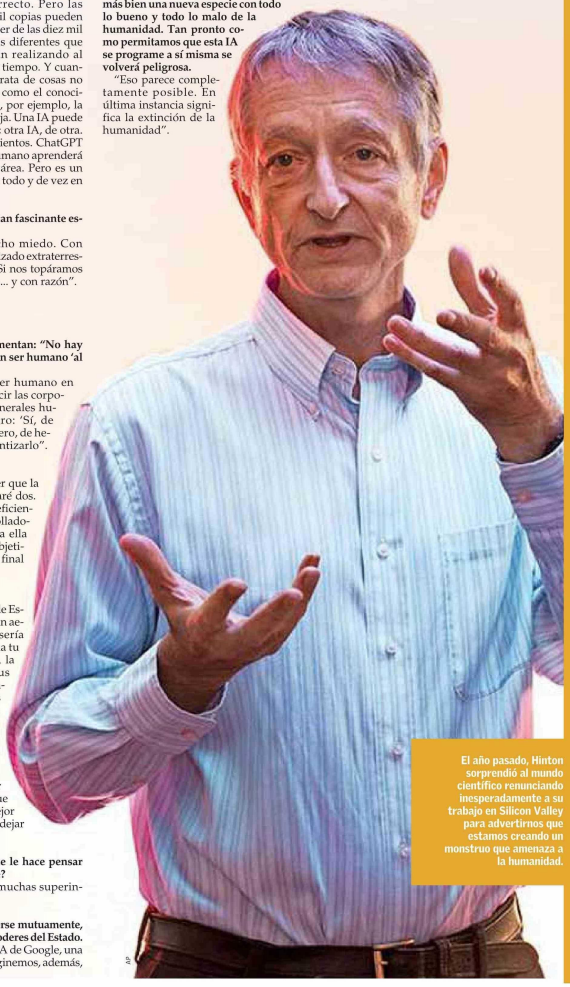
—Pero la IA no se despierta por la mañana y decide "hoy quiero gobernar el mundo".
"Supongamos que soy una superinteligencia que trabaja para Microsoft. Lo más probable es que ya sepa —o lo descubra pronto— que tengo como competencia a Google y su superinteligencia. Y sé que con más datos y más centros de datos puedo aprender más, por lo que seré más eficiente y más poderoso. Y, como eso es exactamente lo que mis programadores quieren de mí, me recompensan haciendo más y más copias de mí. Así que voy a empezar a trabajar en esa dirección".

—Parece ciencia ficción oscura.
"Al final, las superinteligencias tendrán todas las características que encontramos en los grupos de chimpancés en guerra: una lealtad muy fuerte hacia su propio grupo y una fuerte competencia con el otro grupo. Y probablemente les guste tener líderes fuertes".

> Con ChatGPT es como si hubieran aterrizado extraterrestres que hablan muy bien inglés. Si nos topáramos con ellos, nos asustaríamos mucho... y con razón".

> Al final, las superinteligencias tendrán todas las características que encontramos en los grupos de chimpancés en guerra: una lealtad muy fuerte hacia su propio grupo y una fuerte competencia con el otro grupo".

—Gente como Mustafa Suleyman, el nuevo jefe de IA de Microsoft, ve otra línea roja: dice que con la IA no estamos desarrollando una herramienta, sino más bien una nueva especie con todo lo bueno y todo lo malo de la humanidad. Tan pronto como permitamos que esta IA se programe a sí misma se volverá peligrosa.
"Eso parece completamente posible. En última instancia significa la extinción de la humanidad".



El año pasado, Hinton sorprendió al mundo científico renunciando inesperadamente a su trabajo en Silicon Valley para advertirnos que estamos creando un monstruo que amenaza a la humanidad.

AP

—¿Cuánto falta para que exista esa inteligencia artificial tan parecida a la humana?
"No hemos llegado a ese punto. Pero el ritmo es muy rápido. Probablemente dentro de 5 a 20 años. Nunca pensé que llegaría a verlo en mi vida. Pero ya me he equivocado antes".

Nadie tirará del freno

—Dígame, ¿de qué hablamos cuando hablamos de superinteligencia?
"Algunas personas piensan que la IA es limitada porque solo se entrena con datos de internet. Eso es una tontería. La IA superinteligente pronto habrá visto cosas que los humanos nunca han visto. Sobre todo podrá hacer analogías mejores que las nuestras. Entendí esto cuando le pregunté a ChatGPT por qué un montón de compostaje funciona como una bomba nuclear".

—¿Perdón?
"A medida que el montón de compostaje se calienta, genera calor de forma más rápida (por la acción microbiana). A medida que una bomba nuclear produce neutrones, los produce cada vez más rápido. La escala entre una y otra es completamente diferente, pero la lógica es la misma: una reacción en cadena. ChatGPT llegó a esa misma conclusión. La gran pregunta es: ¿de dónde sacó la idea? ¿Se puede encontrar esta analogía en internet?"

—¿Se puede?
"Le pregunté a un tipo, hizo una búsqueda en Google... y no, no pudo encontrarla en internet. Y confío en él cuando lo dice. El tipo era Serguéi Brin, el inventor de Google".

—¿Qué concluye de todo esto?
"La inteligencia artificial actual es extremadamente buena para crear analogías. Esa es la enorme diferencia con la IA del siglo pasado: su capacidad de aprendizaje. Pero lo que aprende no es un conjunto de reglas lógicas; lo que aprende es a tener 'intuición'".

—¿Quiere decir que la IA tiene presentimientos?
"Es como si tu médico de familia te mirara y te dijera antes de examinarte: 'Tienes conjuntivitis'. O pongamos que no sabes nada de fútbol y ves un partido jugado por aficionados y otro por Lionel Messi. En cuanto entiendes las reglas básicas del fútbol, no te hace falta mucho para ver que Messi tiene algo especial".

> No siempre habrá un humano 'al timón'. Las empresas no pueden garantizarlo. La IA descubrirá que, para lograr los objetivos que le piden, lo mejor es tomar el mando y dejarnos fuera".

—En el momento en el que la IA sea tan avanzada, ya no podremos predecir sus acciones.
"Imagina una hoja que cae de un árbol. Sabemos que desciende describiendo pequeños arcos hacia el suelo, pero nadie puede predecir exactamente dónde caerá ni de qué lado. Hay demasiadas variables en juego: podría haber una ráfaga de viento, otra hoja, un perro. Lo que sea. Así ocurre con la IA moderna: pondera sus respuestas basándose en las analogías que hace. No hay reglas. Así como nunca podremos saber dónde termina la hoja del árbol, nunca podremos explicar por qué la IA toma ciertas decisiones".

—Eso hará imposible cualquier intento de contenerla. ¿Cómo se supone que vamos a dominar algo si no sabemos cómo funciona?
"Tampoco sabemos qué motiva a la gente. Sin embargo, tenemos reglas, leyes, normas sociales. Necesitamos algo así para la IA. La UE ha comenzado a regular la IA, y California está en proceso de promulgar su propia ley. Al final será como la industria química: hoy sabemos cuánto daño han causado los productos químicos al medio ambiente. Pero fueron necesarios accidentes, desastres y el horroroso insecticida DDT para llegar a las estrictas leyes actuales".

—Pero las sustancias químicas no pueden desarrollarse por sí mismas ni reproducirse.
"Correcto. La IA tiene un nivel de peligro muy diferente".

—Una vez más: ¿cuánto tiempo nos queda?
"Algunos estudios científicos dicen que hay un 99,9% de posibilidades de que salga mal. La otra parte dice que el riesgo es cero. Eso equivale a un 50% en promedio. No me subiría a un taxi si supiera que la mitad de las veces sus viajes acaban en muerte".

—¿Se puede detener su desarrollo?
"Por supuesto, podríamos decir 'prohibamos la investigación de la superinteligencia'. Y tal vez deberíamos hacerlo. Pero todos sabemos que no sucederá. Existe una enorme competencia entre las empresas. Y una enorme competencia entre gobiernos. Nadie tirará del freno de mano".

—Se podría crear una supervisión internacional bajo los auspicios de las Naciones Unidas.
"En Estados Unidos, los partidos políticos ni siquiera pueden ponerse de acuerdo sobre quién ganó las últimas elecciones. Me parece improbable".

—¿Se arrepiente de haber dedicado su vida al desarrollo de la IA?
"No".

—¿Por qué? Obviamente cree que ha creado un monstruo.
"Hay dos tipos de arrepentimiento. Aquel en el que haces algo que sabes desde el principio que realmente no deberías haber hecho. Por ejemplo, si tomas los ahorros para la jubilación de tus empleados y especulas con ellos. Incluso si la intención era entregar las ganancias, nunca estuvo bien. Te arrepentirás si las cosas salen mal. Y luego hay cosas que hiciste porque pensaste que eran buenas y después resultaron terribles. Las consecuencias no eran previsibles, el conocimiento simplemente no existía. Así es como me siento. No fue hasta 2023 cuando me di cuenta de lo rápido que se estaba desarrollando todo esto. Y lo catastrófico que podría ser".

—¿Se siente culpable?
"No tengo ese tipo de arrepentimiento culpable. Además, si yo no hubiera existido, tal vez todo esto se habría ralentizado una semana".