ENSEÑAR A APRENDER A LAS MÁQUINAS: MITO, TECNOLOGÍA Y POLÍTICA

Las tecnologías de IA no necesitan cumplir las expectativas más extremas para generar profundas disrupciones en las sociedades contemporáneas. Ya las están generando.

JUAN ANTONIO CORDERO

La inteligencia artificial ocupa una posición ambivalente en el imaginario de las sociedades occidentales. Por un lado, está asociada a uno de los principales vectores de progreso: la voluntad de superación de los límites naturales y la movilización de la inteligencia humana al servicio de su propio perfeccionamiento. Por otro, esta ambición va acompañada de un examen crítico de los riesgos que entraña: a cada oleada de avances tecnológicos sigue, tras la fase "tecnoeufórica", un interés renovado por las amenazas potenciales de su despliegue.

La tecnología en el imaginario colectivo

Esta prevención está inscrita en algunos de los mitos más destacados de la cultura occidental: el de Prometeo, el titán griego condenado a una tortura eterna por haber robado a los dioses el tesoro del fuego, para entregarlo a los humanos; o el de la torre de Babel, en el que la insolencia humana al pretender llegar al cielo fue castigada con la confusión lingüística. En la tradición judía, la leyenda del Golem insiste en la misma moraleja, que la cultura popular contemporánea ha seguido cultivando, desde el *Frankenstein* de Mary Shelly (1818) hasta *Terminator* (1989), *Matrix* (2000) o, de forma más inquietante y menos épica, la serie *Black Mirror* (2011): desafiar los límites y pretender sustituirse a los dioses, engendra peligros a la altura de sus promesas, si la tecnología escapa al control de sus creadores y su potencia sobrehumana acaba al servicio de la destrucción o la esclavización, y no de la emancipación, del hombre.

La percepción social de la IA, las expectativas y los temores que alternativamente engendra, ejercen una influencia decisiva en la evolución de las tecnologías relacionadas, el aprendizaje automático y la automatización. Aunque estilizadas y frecuentemente excesivas, estas representaciones populares identifican en ocasiones riesgos reales ligados a estas tecnologías.

Cuatro hitos de la inteligencia artificial

Fundada a mediados del siglo XX, la inteligencia artificial ha experimentado desde entonces un desarrollo innegable, en numerosos

ámbitos[1]. Se pueden mencionar cuatro hitos que han contribuido a moldear la percepción pública: ELIZA y Tay, en el campo del procesamiento de lenguaje natural, y *DeepBlue* y *AlphaGo,* en el de la IA aplicada a juegos.

ELIZA fue probablemente el primer "programa conversacional" o *chatbot*. Desarrollado en el MIT entre 1964 y 1966, para simular una sesión de psicoterapia, ELIZA era capaz de "conversar" con un usuario humano, a través de un sistema relativamente rudimentario de reglas con las que reaccionar a comentarios o palabras clave de su interlocutor. No hay, por tanto, aprendizaje en sentido estricto. El creador de ELIZA, Joseph Weizenbaum, no pretendía crear un dispositivo inteligente sino, más bien, ilustrar las limitaciones de las inteligencias automáticas (Weizembaum, 1976). Pero, pese a la modestia de las interacciones posibles con ELIZA, el sistema conseguía pasar, a ojos de muchos de sus interlocutores, por "inteligente", incluso por humano.

La victoria del programa *DeepBlue* sobre Gary Kasparov –entonces campeón mundial– al ajedrez, en 1997, tuvo un impacto mayor en la opinión pública. Por su complejidad y prestigio, el ajedrez se había convertido en *benchmark* para medir la progresión de la inteligencia automática, y por tanto de constatar el *sorpasso* del hombre por la máquina: si se podía crear una máquina capaz de jugar al ajedrez mejor que el mejor de los humanos, parecía no haber límite a lo que una IA podía conseguir.

Desde un punto de vista algorítmico, la victoria de *DeepBlue* es escasamente "inteligente". La mejora en las tecnologías de fabricación y el aumento progresivo de la capacidad de computación de los microprocesadores permitían a *DeepBlue* ejecutar una cantidad ingente de cálculos por segundo, varios órdenes de magnitud por encima de la capacidad humana. La estrategia para determinar la próxima jugada se basaba en la exploración y reducción del árbol de secuencias posibles del juego, y la selección de los movimientos

1 Se sitúa su "evento seminal" en la conferencia de Dartmouth de 1956, que reunió a los principales investigadores estadounidenses en razonamiento automático, cibernética y autómatas.

que permiten acercarse a estados finales de victoria, con la ayuda del algoritmo de optimización *minimax* (Shannon, 1950). En un juego como el ajedrez, con un conjunto limitado de configuraciones o estados posibles, con reglas bien definidas para mover las piezas, y métricas precisas para determinar el valor de una jugada, la combinación de fuerza bruta de cálculo y de algoritmos de optimización permitía construir jugadores imbatibles para cualquier inteligencia humana, ya en los años noventa.

La progresión de victorias de la máquina contra el hombre, en juegos más o menos complejos, no se ha detenido desde entonces. En 2015, el software de inteligencia artificial *AlphaGo,* de Google, conseguía derrotar consistentemente a los mejores jugadores humanos de Go, un juego sustancialmente más complejo que el ajedrez, en el que la fuerza bruta de cálculo no basta para derrotar al contrario. En este caso, la habilidad de juego del programa mejoraba progresivamente, gracias a la acumulación de observaciones sobre otras partidas, entre humanos o jugadas con humanos. Estas partidas permitían "entrenar" el sistema, y hacerlo "aprender" de su propio juego con técnicas de aprendizaje automático (*machine learning*), hasta el punto de hacerlo imbatible para sus rivales posteriores.

El ejemplo de Tay, *chatbot* creado por Microsoft en 2016, ilustra un aspecto diferente de la IA. Tay era una IA conversacional diseñada para interactuar con los usuarios de Twitter como una joven de diecinueve años. Se trataba de un proyecto de simulación de interacción humana, como el primitivo caso de ELIZA. Bastaron unas horas *online* para que Tay adquiriera, para desolación de sus programadores, una "personalidad" indeseable, racista y homófoba, hasta tal punto que Microsoft tuvo que suspender la experiencia y suprimir el *bot* de Twitter. La explicación oficial señalaba la interacción de Tay con *trolls* de Twitter como el principal motivo de su deriva. Más allá de las fragilidades algorítmicas, el incidente ilustra el impacto que tienen los entornos de los sistemas inteligentes en su comportamiento, y en las posibilidades que se abren de manipularlo deliberadamente mediante la interacción con ellos o a través del entrenamiento.

Inteligencia, razonamiento y aprendizaje

Los ejemplos anteriores, que en modo alguno son exhaustivos, ilustran distintas aproximaciones a la noción de inteligencia automática, sin que haya acuerdo general al respecto. En 1950, Alan Turing propuso el famoso "test de Turing": la idea es que una máquina es inteligente si consigue convencer a un observador exterior, con acceso a las respuestas de la máquina, de que se trata de un ser humano inteligente (Turing, 1950). En decir, es inteligente si *parece* (humanamente) inteligente. Como definición es problemática, y en su momento generó una amplia controversia (ver la discusión sobre la "habitación china", planteada por John Searle en 1980); pero ha ejercido una influencia indudable en la evolución de la IA, que puede reconocerse en los ejemplos de ELIZA y Tay.

El test de Turing asume una noción antropocéntrica de la inteligencia, que encaja (parcialmente) en el programa de la ya mencionada conferencia de Dartmouth: "...que las máquinas usen el lenguaje, formen abstracciones y conceptos, resuelvan categorías de problemas hoy reservadas a los humanos, y se perfeccionen a sí mismas" (McCarthy, 1955). La ambición de "formar abstracciones y conceptos", como hacemos los seres humanos, dio una primera orientación (denominada "IA simbólica") a la disciplina, centrada en la representación conceptual y la lógica formal como pilares del razonamiento automático.

Otros enfoques, procedentes de la psicología (*e.g.*, Humphreys, 1972; Sternberg & Selter, 1982), han privilegiado la habilidad para adquirir datos, combinarlos y reutilizarlos; y más en general, en las capacidades de adaptación y aprendizaje orientadas a la consecución de objetivos. La noción de "agente inteligente" (Wooldridge, 1995, 1999; Russell & Norvig, 1995) extiende estas nociones más allá de la emulación humana: un agente inteligente es cualquier entidad (dispositivo, programa, un autómata, ser vivo) capaz de interactuar con su entorno y "aprender" de esa interacción, es decir, adaptar su respuesta de acuerdo con su "función objetivo", y según el efecto observado/inferido de respuestas anteriores. Esta adaptación puede realizarse con técnicas de *machine learning,* otra orientación de la IA cuyos métodos

han conocido un desarrollo importante en los últimos años– cuyas bases conceptuales son conocidas desde hace décadas.

Computación, conectividad y ley de Moore

La inteligencia artificial ha tenido una evolución irregular. A las épocas de efervescencia, han seguido largos períodos de estancamiento y depresión, al constatar que las posibilidades tecnológicas reales no estaban a la altura de las expectativas generadas: son los llamados "inviernos" de la IA. Los más severos se produjeron en los años setenta, cuando se hizo evidente que muchos algoritmos resultaban impracticables para resolver problemas de tamaño real, tanto por la complejidad de éstos, como por las limitaciones computacionales de los sistemas de la época (informe Lighthill, 1973); y a finales de los años ochenta, con el agotamiento de las técnicas de IA simbólica, centradas en la representación y el razonamiento lógico, y de los llamados "sistemas expertos", que habían constituido su pista más prometedora.

Uno de los factores que permitieron superar este último invierno fue, justamente, el aumento sostenido en las capacidades de computación, almacenamiento y obtención de datos, y la reorientación de la IA hacia las tecnologías de aprendizaje automático[2], cuyo desarrollo se había visto anteriormente limitado por su coste computacional. La combinación del crecimiento en las capacidades de cálculo de los microprocesadores (ley de Moore), con el desarrollo de Internet y el consiguiente aumento de la conectividad y las capacidades de computación distribuida, ha dado así un nuevo impulso desde los años noventa a la IA, que privilegia la explotación de datos masivos mediante mecanismos de aprendizaje automático (y no de lógica formal y razonamiento simbólico, como en el pasado), de base esencialmente estadística: las técnicas de *Deep Learning* ("aprendizaje profundo") reposan sobre la identificación de correlaciones y patrones estadísticos complejos en grandes volúmenes de datos.

2 Entre ellas, el perceptrón, conocido desde los años sesenta, y el algoritmo de retropropagación *(backpropagation)*, desarrollado entre los años setenta y ochenta; ambos constituyen elementos básicos de las redes neuronales.

Ambos elementos configuran una inteligencia profundamente dependiente de los datos: los sistemas inteligentes lo son en la medida en que tienen, merced a los progresos de las tecnologías de computación, una capacidad sin precedentes para procesar y explotar datos; y merced a la consolidación de Internet y las tecnologías de comunicación, un acceso a volúmenes y flujos de datos que tampoco tiene precedentes. Esta inteligencia automática está así estrechamente condicionada por la calidad y la cantidad de los datos que los sistemas manejan[3]. Su dependencia de infrastructuras externas (ya sean las infrastructuras de comunicación que constituyen Internet y que permiten el tránsito de datos, o los grandes centros de procesamiento, almacenamiento y computación que permiten su explotación) obliga a integrar en las consideraciones sobre IA los costes (económicos, medioambientales), vulnerabilidades e implicaciones de estas infrastructuras y tecnologías habilitantes en las que descansan los sistemas de inteligencia distribuida.

La centralidad de los datos

La interacción de un "sistema inteligente" de base estadística con su entorno se realiza a través de los datos que se le facilitan: la modelización del entorno que estos datos reflejan define el flujo de entradas que informa el sistema inteligente y alimenta (entrena) su mecanismo de aprendizaje, y al hacerlo determina los factores del entorno considerados significativos, y cuya evolución podrá ser tenida en cuenta por los algoritmos implementados.

Ésta no es una operación neutra: el mismo individuo puede ser descrito en función de su etnia, su género y su edad; o bien a través de su nivel de estudios, su lugar de residencia y su salario. Las conclusiones que puede extraer el mismo sistema inteligente –el mismo mecanismo

3 Aunque la cuestión desborda el ámbito de este ensayo, el hecho de que buena parte de los datos que alimentan y perfeccionan las inteligencias automáticas sean provistos cotidianamente por humanos (*e.g.* informaciones de redes sociales, preferencias expresadas, textos, imágenes, etc.), lleva a algunos autores a plantearse la pertinencia de establecer formas de redistribución de la plusvalía asociada a las IAs, que serían en realidad una retribución por los datos, hoy cedidos gratuitamente, que las nutren y entrenan (Arrieta-Ibarra *et al.*, 2018).

de aprendizaje automático– de las mismas realidades, descritas con estas dos modelizaciones aplicadas a los mismos individuos, pueden ser radicalmente diferentes – y están, en parte, implícitas en las variables que conforman el modelo. Esta dependencia de datos y modelos lleva a destacados científicos a cuestionar que los sistemas basados en la optimización y el aprendizaje estadístico puedan considerarse "inteligentes" en absoluto (Pearl, 2018), aunque puedan tener éxito en tareas específicas y de gran complejidad.

En ocasiones se asume que, aunque su funcionamiento y su precisión puedan ser deficientes en un principio, estos sistemas se vuelven progresivamente más "inteligentes" a medida que se entrenan y se ven expuestos a conjuntos más amplios de datos. En realidad, con un modelo defectuoso o insuficientemente preciso del entorno, y/o con una exposición sesgada a éste entorno (con unos datos de entrenamiento no representativos), el sistema no puede converger hacia un comportamiento "inteligente" aunque se faciliten flujos masivos de datos de entrenamiento, y aunque se empleen los mecanismos más sofisticados de IA: la capacidad efectiva de aprendizaje está limitada por la calidad de las representaciones (modelos) y de los datos que el sistema puede emplear para configurarse.

Otro elemento considerar es la métrica de optimización o rendimiento: el criterio o conjunto de criterios que el sistema maneja para evaluar sus propias decisiones. Estos criterios dependen de la aplicación, pero la aplicación no determina la métrica: para un mismo problema, pueden existir diversas medidas de rendimiento –quizá incompatibles–, y la priorización de unas u otras puede arrojar resultados distintos. Las decisiones de un vehículo autónomo, por ejemplo, dependerán de si éste busca optimizar el tiempo de trayecto (cuanto más corto, mejor), el número de accidentes (cuantos menos, mejor), o el consumo eléctrico (cuanto menos contaminante, mejor).

El modelo de datos, el muestreo y la función de rendimiento responden a elecciones "políticas" –no reductibles a criterios meramente científicos o tecnológicos–para el sistema en cuestión, en ocasiones implícitas, pero no por ello menos reales. Estos tres elementos tienen

un peso relevante en el funcionamiento de los sistemas inteligentes, y pueden convertirse en socialmente críticos –y merecedores de una discusión pública en profundidad–, si el uso de estos sistemas se generaliza, por ejemplo, en los dispositivos de acceso y prestación de servicios públicos y/o de alto impacto en la vida cotidiana. Pese a su apariencia técnica, estos elementos son exteriores a los algoritmos y las técnicas de aprendizaje. Y por eso mismo, los sesgos que pueden inducir en los sistemas inteligentes no son necesariamente corregibles, ni siquiera detectables, mediante la mera iteración –el "aprendizaje"– del algoritmo en cuestión.

Se ha mencionado ya el incidente de Tay, pero hay más ejemplos más cotidianos, más inquietantes y de mayor impacto social. En octubre de 2019, Amazon reconocía que la IA que empleaba para filtrar automáticamente candidatos a ofertas de trabajo en la compañía, discriminaba sistemáticamente a las mujeres. No se trataba de un comportamiento explícitamente implementado, sino "aprendido" por el sistema: la IA había sido entrenada con (es decir, tomando como referencia) el histórico de las contrataciones realizadas por la compañía, y programada para producir decisiones de contratación lo más similares posibles a éstas. Ese histórico, fuertemente sesgado a favor de los hombres, había llevado al sistema a reproducir, sin intervención humana aparente, un patrón de contratación visiblemente discriminatorio. El incidente causó un escándalo comprensible; pero sus causas no son tecnológicas. Un sistema automático entrenado con datos sesgados (en este caso, las contrataciones realizadas por operadores humanos en el pasado), y diseñado para tomar decisiones lo más parecidas posibles a las "aprendidas" en su entrenamiento, va a reproducir –no crear– mecánicamente el sesgo presente –humano– en los datos que emplea como referencia. No es el algoritmo quien introduce elementos indeseables o imprevistos en el rendimiento del sistema: son los factores "políticos", humanamente mediados, que rodean su puesta en práctica, los que condicionan su resultado. Con frecuencia, las noticias más inquietantes que se leen sobre las derivas de los algoritmos existentes de inteligencia artificial, tienen menos

que ver con la tecnología en sí, y más con estos aspectos "políticos" y humanos que rodean su implementación y su uso – aspectos menos sensacionales, no específicos de la IA, y sobre los que hay responsables y decisores de carne y hueso.

Tecnología y poder

Algunos investigadores (*e.g.,* Crawford (2021)) deploran el uso del término "inteligencia artificial" para englobar los algoritmos de optimización y aprendizaje estadístico: a su juicio, la denominación induce a la confusión. Desde luego, estos algoritmos son capaces de procesar grandes volúmenes de datos, detectar patrones estadísticos en ellos y tomar decisiones a una escala inalcanzable para el ser humano –de la misma forma que los trenes de alta velocidad van mucho más rápido que humanos y animales–, y ello les permite realizar tareas complejas, muy específicas y de gran precisión en un amplio espectro de aplicaciones (desde el diagnóstico automático de enfermedades a través de imágenes médicas, hasta la predicción meteorológica, pasando por la asistencia en vuelo, aterrizaje y despegue). Por el momento, no hay "inteligencia" ni cognición equiparable a la de un ser vivo –no ya a un humano– en estas tareas: no hay capacidad de generalizar ni de abstraer más allá de la detección de correlaciones estadísticas; menos aún hay conciencia de sí.

Pero las tecnologías de IA no necesitan cumplir las expectativas más extremas que activan en el imaginario colectivo[4] para generar profundas disrupciones en las sociedades contemporáneas. Ya las están generando: aunque no sean propiamente "inteligentes", las tecnologías actuales ya modifican en profundidad el ocio, la economía y el trabajo, las sociedades, la Administración, los servicios públicos y –necesariamente–la política; la manera en la que nos relacionamos con nuestro entorno, ejercemos nuestros derechos, construimos nuestras preferencias y formamos nuestros juicios. Que la modifiquen en

4 Ya sean éstas la rebelión de las máquinas, la esclavización o sustitución de los hombres, las visiones transhumanistas, o la Singularidad (Good, 1965) que daría paso a la llamada "IA fuerte", en la que las inteligencias artificiales serían superiores a las humanas.

BIBLIOGRAFÍA

Alan M. Turing (1950): Computing Machinery and Intelligence. *Mind,* vol. 49, pp. 433-460.

Claude E. Shannon (1950): Programming a Computer for Playing Chess. Philosophical Magazine, ser. 7, vol. 31, núm. 314, marzo 1950.

John McCarthy, Marvin L. Minsky, Nathaniel Rochester, Claude E. Shannon (1955): *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. 31 de agosto de 1955.

Irving John Good (1965): SpeculationsConcerning the First Ultraintelligent Machine *Advances in Computers,* vol. 6, 1965. Academic Press Inc.

James Lighthill (1973): Artificial Intelligence, A General Survey. *En Artificial Intelligence, a paper symposium,* Science Research Council, Reino Unido.

Joseph Weizenbaum (1976): *Computer Power and Human Reason.* San Francisco: W. H. Freeman, 1976.

L. G. Humphreys (1979): The construct of general intelligence. *Intelligence,* vol. 3, issue 2, pp. 105-120.

John Searle (1980): Minds, Brains, and Programs. *Behavioral and Brain Sciences*, vol. 3, issue 3, pp. 417-457.

R. J. Sternberg, W. Salter (1982): *Handbook of human intelligence.* Cambridge, UK: Cambridge University Press, 1982.

Michael Wooldridge, Nicholas R. Jennings (1995): Intelligent agents, theory and practice. *The Knowledge Engineering Review*, vol. 10, issue 2, pp. 115-152, 1995.

Michael Wooldridge (1999): Intelligent Agents. *Multiagent Systems,* vol. 35, issue 4.

Stuart J. Russell, Peter Norvig (1995): *Artificial Intelligence: A Modern Approach, 1st Edition*. New Jersey, US: Prentice Hall, 1995.

Yuxi Liu (2017): The Accountability of AI – Case Study : Microsoft's Tay Experiment. *Chatbots Life.*

Imanol Arrieta-Ibarra, Leonard Goff, Diego Jiménez-Hernández, Jaron Lanier, E. Glen Weyl (2018): Should We Treat Data as Labor? Moving beyond "Free". *AEA Papers and Proceedings,* American Economic Association, vol. 108, pp. 38-42, Mayo 2018.

Judea Pearl, Dana Mackenzie (2018): *The Book of Why, the New Science of Cause and Effect.* UK: Penguin Books, 2018.

Kate Crawford (2021): *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* New Haven, US: Yale University Press, 2021.

un sentido beneficioso para el interés general depende de que la ciudadanía se familiarice con su funcionamiento y sus límites, se las apropie –como se ha apropiado tecnologías anteriores– y las ponga a su servicio. Por ello, estas tecnologías han de ser sometidas al escrutinio, la discusión y la regulación pública. Algo que requiere despojarlas del aura "mágica", casi oracular, de dispositivos que se temen o se reverencian, pero que ni se entienden ni se discuten, con la que aún se las adorna en ocasiones.

Sin un esfuerzo educativo, divulgativo y político que permita esta apropiación colectiva, que es sobre todo una democratización de las oportunidades que abre, la expansión de las tecnologías de IA puede contribuir a agravar desequilibrios y desigualdades existentes y producir otras nuevas, tanto en el interior de las sociedades, como entre países y regiones del mundo. La llamada "brecha digital", observable desde hace décadas, puede convertirse en insalvable entre los sectores sociales que comprenden y dominan

estas tecnologías –o al menos manejan sus rudimentos estadísticos, informáticos y técnicos–, y por ello pueden explotar sus ventajas; y quienes las sufren pasivamente sin medios para manejarlas, y pueden verse atrapados en dinámicas sociales, económicas y políticas amplificadas por éstas (desde la polarización alimentada por las redes sociales, hasta la precarización posibilitada por la economía de plataformas). Sin un esfuerzo público sostenido para desarrollar y mantener capacidades tecnológicas propias, formar a ciudadanos y trabajadores, y distribuir de forma equilibrada los rendimientos y las oportunidades generadas por las tecnologías asociadas a la IA, la tendencia a la concentración de los medios de producción y explotación de la IA reforzará los desequilibrios de poder, fragilizando la cohesión social en el interior de las sociedades, potenciando la inestabilidad social, y agravando la dependencia científica y tecnológica respecto a aquellas potencias con medios y voluntad suficiente para capitalizarla. ❧

Juan Antonio Cordero es Doctor en Informática (École Polytechnique), ingeniero de Telecomunicaciones (UPC) y licenciado en Matemáticas (UPC). Profesor de Informática y Redes en la École Polytechnique (Francia).