argumenta philosophica

2020/1

Artículos

Reseñas

LA INTELIGENCIA DE LA INTELIGENCIA ARTIFICIAL. CONSIDERACIONES EPISTEMOLÓGICAS

Daniel Innerarity

Resumen

En este artículo me pregunto si la inteligencia artificial puede calificarse realmente como inteligente, cuáles son las propiedades de la inteligencia humana que la inteligencia artificial no ha conseguido aún imitar y cuál es el papel de las humanidades y de las ciencias sociales en el diseño de una inteligencia artificial más integrada y amplia.

Palabras clave: inteligencia artificial, conocimiento, sentido, robots.

The intelligence of artificial intelligence. epistemological considerations

Abstract

In this article I wonder if artificial intelligence can really be described as intelligent, what are the properties of human intelligence that artificial intelligence has not yet been able to imitate and what is the role of the humanities and social sciences in the design of a more integrated and comprehensive artificial intelligence.

Keywords: artificial intelligence, knowledge, meaning, robots.¿Hasta qué punto son inteligentes las *smart technologies,* las *neural networks* y los *deep learning systems*?

> «El juicio es la capacidad de integrar una vasta amalgama de datos en constante cambio, multicolores, evanescentes, que se superponen perpetuamente, demasiados, demasiado rápidos, demasiado entremezclados para ser descubiertos y atrapados, y etiquetados como tantas mariposas individuales».
>
> Berlin, Isaiah (1996), *The Sense of Reality*. Londres: Chatto & Windus, 46.

¿Hasta qué punto son inteligentes las *smart technologies,* las *neural networks* y los *deep learning systems*? No podemos dar por hecho que la inteligencia artificial sea inteligente o, al menos, que lo sea en la medida de nuestras expectativas, y tal vez lo más provechoso sea preguntarse qué debemos hacer con ella para que sea lo más inteligente posible. Pero este objetivo no se podrá realizar si no acertamos a la hora de identificar qué es lo específicamente humano de la inteligencia humana y entonces sabremos dónde están los límites de la inteligencia artificial que estamos intentando superar.

Debates grandilocuentes: inteligencia, habilidad o estupidez

Las grandes discusiones del siglo pasado acerca de cómo interpretar la significación histórica de la inteligencia artificial enfrentaron a quienes imaginaban máquinas capaces de reemplazar a los humanos (MacCarthy), para bien o para mal, y quienes sostenían que se trataba de un mero aumento de la inteligencia humana (Licklider y Engelbart), como el famoso debate de los años sesenta en Stanford: el proyecto de construir máquinas inteligentes sustitutorias frente a quienes aspiraban a aumentar la inteligencia humana. Las posiciones de estos enfáticos debates fueron cristalizando en la contraposición entre una «inteligencia artificial general» (AGI) y una «inteligencia artificial restringida» (NAI). Si la primera tiene como objetivo emular a la inteligencia humana, la segunda simplemente la simularía.

La discusión continúa en otros términos. Kurzweil (2001) polemiza con Kapor a comienzos de este siglo y asegura que la inteligencia artificial «superará a la inteligencia humana nativa» en 2029. Otra manera de decirlo es decretar el final de la teoría (Anderson 2008) y declarar así obsoleto el método cientifico tradicional, demasiado antropocéntrico. Hay quien asegura que la inteligencia natural es un caso especial de la inteligencia artificial (Wilczek 2019, 68). En el otro extremo, algunos prefieren denominarla «inteligencia aumentada» para desdramatizar así su novedad, hasta el punto de afirmar que «la inteligencia artificial no existe» (Julia 2019). Para los minimalistas, Deep Blue es solo una supercalculadora y no tiene nada de verdaderamente inteligente. Según estos, las inteligencias artificiales son y seguirán siendo limitadas, hasta el punto de que tal vez en el futuro desistamos de calificarlas como inteligentes.

Desde el punto de vista epistemológico, la gran cuestión no es si la inteligencia artificial es una mera prótesis cognitiva, una obnubilación de la razón o una habilidad irreflexiva; lo más interesante es que se trata de un conjunto de tecnologías que nos están obligando a redefinir qué significa conocimiento en este nuevo contexto. Se ha respondido a esa inquietante cuestión planteando un desdramatizado reparto del territorio (las máquinas son mejores en el descubrimiento de patrones, matematización y razonamiento estadístico, análisis de datos masivos y manejo de casos rutinarios; los humanos estableceríamos los objetivos y formularíamos juicios de valor, resolución de información ambigua y discernimiento

en casos difíciles), pero siendo cierta esa división del trabajo, no deberíamos perder esta oportunidad de volver a pensar qué debemos entender por inteligencia.

Una salida plausible a este debate es la paradoja de que la actual inteligencia artificial sería muy inteligente y muy estúpida a la vez; su estupidez consiste en que cuando toma una decisión inteligente no tiene modo de saberlo (Dessalles 2019). Lo que tendríamos entonces son «sabios digitales idiotas» (Domingos 2015; Carr 2014). Esta controversia solo puede resolverse si abandonamos la grandilocuencia y entramos a examinar cómo funcionan de hecho las dos inteligencias, qué tipo de relación se ha establecido entre ellas y si podemos modificar esa relación de modo que mejore la inteligencia que somos y la que tenemos a nuestra disposición.

Todas las tecnologías han tenido como consecuencia una cierta «periferización» de los humanos del ámbito de las decisiones (eso es en última instancia la automatización), pero ninguna había sido tan disruptiva hasta el punto de que la epistemología antropocéntrica parece completamente inadecuada frente a las nuevas autoridades epistémicas automatizadas. El horizonte de automatización general sitúa a los humanos «fuera del centro de la empresa epistemológica» (Humphrey 2019, 23). ¿Es cierto que el razonamiento se está disociando de las capacidades cognitivas humanas? ¿Puede llegar el conocimiento como tal, no solo sus instrumentos, a convertirse en algo automático, como asegura Stiegler (2017)? No podremos responder a estas preguntas si no caracterizamos con precisión la naturaleza del conocimiento humano y las propiedades de la inteligencia artificial, de manera que podamos señalar los límites de esta última, tanto para saber qué es y qué no es sustituible como para proponer nuevos horizontes de desarrollo a la inteligencia artificial.

La actual encrucijada de la inteligencia artificial

Frente a quienes están preocupados porque el avance de la inteligencia artificial nos marginalice o sustituya, la realidad es que sus límites son muy persistentes y solo se superarían si consiguiéramos hacerla más parecida a nuestra inteligencia. No debería preocuparnos que quiera reemplazarnos, sino que no pueda hacerlo. La idea de un *AI takeover* se ha tomado demasiado en serio la fase inflacionaria en la que estamos actualmente; de hecho, la inteligencia artificial «se limita a problemas que consisten en mapear *inputs* bien definidos en relación con *outputs* bien definidos en ámbitos en los que se dispone de un conjunto de entrenamientos gigantescos, en donde la medida del éxito es inmediata y precisa y en los que no es necesario ningún razonamiento gradual, jerárquico o abstracto» (Pinker 2019, 110). La actual falta de fiabilidad de la inteligencia artificial se debe al hecho de que es un conjunto de técnicas inapropiadas para un mundo abierto, técnicas que funcionan para problemas muy específicos donde las reglas no cambian y cuando disponemos de todos los datos que queramos. La inteligencia de la inteligencia artificial futura consistiría en parecerse más a la vida real, en poder hacerse cargo de esos ámbitos que actualmente se le escapan debido a su carácter abierto, contextual, incierto o comprensivo, con

datos escasos o poco fiables, donde los humanos seguimos siendo más inteligentes que ella. ¿Qué significa esto en el actual estado de la cuestión de la investigación sobre la inteligencia artificial y para la superación de las actuales fronteras del conocimiento?

Comencemos recordando la paradoja de Moravec (1988): comparativamente es fácil conseguir que las computadoras muestren capacidades similares a las de un humano adulto en tests de inteligencia, y difícil o imposible lograr que posean las habilidades perceptivas y motrices de un bebé de un año. Este contraste es muy significativo epistemológicamente. Todos los humanos somos expertos en la vida cotidiana (pasar el aspirador, encender la chimenea, atarse los zapatos...), pero hemos invertido un montón de años de formación en adquirir una especialidad, como hacer un diagnóstico médico, tocar el piano o arreglar un coche. Para los sistemas expertos es justo al revés; es relativamente fácil proporcionarles un saber experto, pero es casi imposible hacerles razonar sobre aquello que a nosotros nos parece evidente. «Hay un contraste entre lo que las máquinas hacen bien ahora —clasificar cosas en categorías— y el tipo de razonamiento y comprensión del mundo que se requerirían para capturar esta capacidad mundana pero crítica» (Marcus / Davis 2019, 74).

La creación de una inteligencia capaz de entender lo que un niño de seis años comprende inmediatamente no se conseguirá aumentando la potencia de las técnicas actuales. Los límites de la inteligencia artificial no son una cuestión de potencia de cálculo o de tamaño de memoria —algo que podría ser resuelto con el mero «darwinismo de los datos» (Malik 2013)—, sino de carencia de ciertos mecanismos de los que está dotado el ser humano, como la comprensión general, algo que se pone de manifiesto en que la traducción automática, por muy perfeccionada que esté, no lee propiamente el texto, sino que trata y considera los símbolos como algo desconectado de la experiencia del mundo (Hofstadter 1979).

Puede estar ocurriendo que muchos límites para el avance de la inteligencia artificial tengan que ver con que la concepción *mainstream* de lo que es inteligencia sea reduccionista y no preste atención a sus dimensiones cualitativas, contextuales, intuitivas, inexactas, artesanales y corporales del conocimiento. «La actual inteligencia artificial es estrecha; funciona para las tareas particulares para las que está programada, siempre y cuando lo que encuentre no sea muy diferente de lo que ha experimentado antes» (Marcus / Davis 2019, 13-14). Lo que le falta es inteligencia amplia. Es hábil al encontrarse con situaciones específicas para las cuales hay una gran cantidad de datos, pero no para problemas nuevos o situaciones inéditas. La inteligencia artificial del futuro deberá aproximarse más a la nuestra, tiene necesidad de otros mecanismos cognitivos similares a los específicamente humanos. Este es en buena medida el objetivo de esa rama de la ciencia cognitiva que es la experiencia corporizada. La inteligencia humana no es pensable sin todos esos procesos cerebrales y corporales que incluyen conciencia de sí, afectividad e intuición.

Mientras esperamos mejores resultados, lo que hoy tenemos es una excitación con los *big data* que nos ha distraído de los

problemas que requieren una comprensión más profunda del mundo. La actual inteligencia artificial vale para entornos en los que rige el principio de que cuantos más datos mejor. Esto tiene muy poco que ver con la lógica humana. Si quiere atravesar nuevas fronteras, la inteligencia artificial tiene que aprender más de cómo la gente realmente piensa: de nuestra comprensión, rapidez y adaptación, de nuestra capacidad de actuar con información incompleta e incluso inconsistente, consumiendo poca energía, sin muchos datos, aproximadamente. Este reduccionismo de la inteligencia a gestión de datos y cálculo es lo que explica que estemos cediendo poder a unas máquinas que no son muy fiables, especialmente en lo que se refiere a valores humanos, sentido o visión de conjunto de su inserción en una sociedad política, con sus prioridades y sus objetivos de equilibrio, sostenibilidad o igualdad. El cambio de paradigma de la futura inteligencia artificial debe ser su «humanización», en el sentido de que incorpore en la medida que sea posible estas dimensiones de sentido, comprensión y equilibrio que hasta la fecha no ha sido capaz de desarrollar.

En sus comienzos la computación era una operación de cálculo que podríamos definir como solipsista: no requería que las máquinas tuviesen un conocimiento sofisticado de ser en el mundo, de estar en nuestro mundo real, complejo y dinámico, emergente, múltiple, impreciso y contingente, un mundo que requiere esa conciencia que los humanos producimos de manera intuitiva e implícita. Nuestra realidad es contextual e histórica. Si las máquinas quieren estar a la altura de los humanos, deben desarrollar algo así como una conciencia de estar en el mundo. Y deberán poder hacerlo ellas mismas porque nosotros no podemos preprogramar la complejidad de todos los eventos y todas las condiciones emergentes posibles de un contexto, como tampoco podemos traducir en reglas formalizadas aquel conocimiento con el que actuamos en el mundo, un conocimiento que es en buena medida informal, inconsciente e implícito.

La especificidad del conocimiento humano

¿Cuál es esa especificidad de la inteligencia humana que supone una frontera para la inteligencia artificial pero de la que de algún modo debería hacerse cargo si es que quiere realizar avances significativos? Esta especificidad de la inteligencia humana son un conjunto de propiedades que cabe agrupar en sentido común, reflexividad, conocimiento implícito, inexactitud y economía.

§ Sentido común

¿En qué consistiría esa antropomorfización de la inteligencia artificial? En la configuración de un equivalente funcional de algo que podríamos llamar sentido común, cuya carencia limita actualmente la inteligencia de los sistemas. «La gran ironía del sentido común es que se trata de algo que todo el mundo conoce, pero nadie parece saber qué es exactamente o cómo construir máquinas que lo tengan» (Marcus / Davis 2019, 150). La mayor parte de los progresos que ha hecho la inteligencia artificial, como el reconocimiento de objetos, son muy diferentes de los desafíos de lo que

significa «comprender». Es cierto que, frente a los clásicos modelos deductivos lineales, «las redes neuronales solo utilizan grandes vectores de actividad, matrices de gran peso y escalas no lineales para realizar el tipo de inferencia rápida 'intuitiva' que sustenta el razonamiento sin esfuerzo con sentido común» (LeCun / Bengio / Hinton 2015, 438). Pero se trata de una intuición que todavía dista mucho de la propia de los humanos.

Pasa lo mismo con los robots. Se ha hecho un excelente trabajo en realizar actividades singulares, pero hay muchas áreas para las que son incompetentes: predecir futuros posibles y decidir en situaciones cambiantes. O pensemos en las dificultades del *deep learning* con el lenguaje, que provienen de la sutileza de nuestro lenguaje y para cuya resolución necesitamos no solo gestión de datos sino también comprensión del contexto humano. Hay una «capacidad para la relevancia» que caracteriza al conocimiento humano y que los dispositivos artificiales no parecen hoy por hoy capaces de reproducir del todo. Cálculo y juicio son dos cosas diversas, como podemos comprobar escuchando algunas respuestas que los asistentes artificiales dan a ciertas demandas. El descubrimiento del sentido es un verdadero problema para los ordenadores aunque tengan acceso a una cantidad gigantesca de textos digitalizados. Una máquina que no sabe nada de entrada puede descubrir el sentido de las palabras a base de analizar las frases que las contienen a través del procedimiento de la co-ocurrencia, pero todavía está por ver que calculando una proximidad de sentido entre las palabras pueda determinarse el verdadero sentido del lenguaje que los humanos identificamos con facilidad, pese a los errores y equívocos en los que solemos incurrir. Y es que la comprensión del mundo pasa también y sobre todo por la comprensión del contexto o del marco en que nos encontramos e implica una capacidad de juzgar la relevancia de las situaciones.

§ Reflexividad

La inteligencia artificial es un conjunto de técnicas geniales para aprenderse el mundo de memoria. Aunque sobrepase la potencia calculatoria del ser humano, la inteligencia artificial es incapaz de dar una significación a sus propios cálculos. Programas como Wattson, más que resolver problemas, lo que hacen es buscar la solución en la red. Su principal inteligencia no consiste en comprender la estructura del problema sino en adivinar, entre todas las respuestas recogidas, cuál es la que tiene más posibilidades de ser la buena. Los sistemas inteligentes actuales dan la impresión de que no comprenden lo que hacen. Alguien podría objetar que eso importa poco si dan con las soluciones adecuadas. El problema es que los humanos tenemos necesidad de comprender los problemas para resolverlos. ¿Qué significa aquí «comprender»?

Frente a la idea de que las máquinas no son capaces de estar a la altura de la reflexividad humana, hay al menos dos posibles tipos de objeciones: que no les hace falta o que son capaces de desarrollar esa capacidad por ellas mismas.

La primera opinión es sostenida por Remo Bodei (2019), para quien en el fondo no habría tanta diferencia entre los humanos y las máquinas. A medida que estas realizan

prestaciones más eficaces, el individuo contemporáneo ya no sería único depositario de una racionalidad ligada de manera indisoluble a un cuerpo viviente y a una inteligencia consciente. Bodei recuerda una aportación de Leibniz que podría ser especialmente valiosa para este asunto: la idea de que existen «pensamientos ciegos» (*cogitationes cecae*) (Leibniz 1950, 4, 35, cit. por Bodei 2019, 309), que se caracterizarían por su naturaleza inconsciente, por ser irrepresentables al nivel de la conciencia. Su existencia refuta la identificación del *cogito* cartesiano con la conciencia. Se puede pensar sin tener conciencia de los significados y los contenidos pensados, como es el caso de los símbolos algebraicos o en el cálculo, cuando la mente adiestrada procede por automatismos. Existiría entonces un tipo de automatismo en la inteligencia humana que puede ser objetivado e inserto en las máquinas calculadoras, que operan análogamente a los «pensamientos ciegos». En este sentido se podría afirmar que los dispositivos dotados de inteligencia artificial «piensan», aunque de manera ciega, porque no necesitan conciencia, según defiende también Copeland (1993, 33).

La otra posibilidad que objetar a su falta de reflexividad consiste en suponer que la inteligencia artificial es una inteligencia capaz de evolucionar y conseguir esa reflexividad de la que actualmente carece. Es lo defendido por la teoría de los algoritmos genéticos, capaces de proponer soluciones para problemas mal planteados (Dessalles 1996). ¿De qué modo pueden los ordenadores elevar su nivel de inteligencia por ellos mismos, sin ayuda exterior, evolucionando en el sentido darwiniano del término? A diferencia de los seres vivos, los sistemas inteligentes solo pueden innovar en el interior de un marco estrictamente delimitado. Cada instrucción ejecutada por un programa debe estar cargada previamente en el procesador por otra instrucción. La diferencia fundamental es que el genoma de los seres vivos contiene instrucciones que permiten indirectamente al ADN interpretarse a sí mismo. La inteligencia artificial está de momento muy lejos de esto, no parece capaz de evolucionar fuera de los estrechos límites fijados en el punto de partida. Esto no impide que los programas adquieran un gran poder, hasta el punto por ejemplo de controlar las finanzas mundiales o la manera de pensar de comunidades enteras en las redes sociales, pero la limitación de sus posibilidades no hace de la inteligencia artificial un sistema incontrolado.

¿Se puede ser inteligente sin saberlo, como un zombi que fuera capaz de realizar tareas inteligentes de un ser humano pero solamente de manera refleja? La inteligencia artificial es, hoy por hoy, un sistema supuestamente inteligente; se contenta con aprender una función, pero no reflexiona. Tiene inteligencia refleja, no reflexiva. Y esto no corresponde a la noción que tenemos de inteligencia.

§ Conocimiento implícito

Con el estado actual de la técnica, el único procedimiento para que una máquina disponga de sentido común es dárselo de manera explícita (lo que no es el sentido común que tenemos los humanos, una facultad de lo implícito). Los sistemas inteligentes son no solamente mudos, incapaces

de explicar sus decisiones, sino que tampoco pueden percibir elementos imprevistos en su construcción. Los límites de las máquinas son los del saber explícito: los sistemas de la inteligencia artificial no disponen más que de conocimientos explícitos que ninguna información implícita puede modificar. Por eso no identifica bien las incoherencias, las imposibilidades o el juego de simulaciones y engaños que forma parte de la comunicación humana. Los avances significativos en esta dirección requieren una inteligencia artificial más fenomenológica que cartesiana. La inteligencia artificial dominante es heredera de la epistemología de la modernidad y necesita completar el giro interpretativo que la teoría del conocimiento llevó a cabo a mediados del siglo XX de la mano principalmente de Heidegger (1967) y Wittgenstein (1971). Lo que nos define es que somos seres de lo implícito.

La modernidad puso en circulación la idea del hombre que piensa y actúa como un sujeto desinteresado, descomprometido y distanciado de su mundo, sin cuerpo ni contexto, que en última instancia no está afectado por cultura o forma de vida alguna, ni implicado en un mundo de relaciones. Esta concepción es la todavía vigente en los modelos informáticos de una conciencia «sin cuerpo» que se apropia de pedazos (*bits*) de información de su entorno y los «procesa» de una determinada manera. En esta construcción pervive la tradición de una neutralidad sin sujetos y relativa a hechos, que intenta despojar de relevancia interpretativa al *input* de información y degradarlo al nivel de mero registro de datos. La comprensibilidad es dada sin más por supuesta y no necesita de ningún contexto de interpretación que la posibilite. Se parte de que los pedazos de información son concebidos como tales desde el principio y que las operaciones posteriores no hacen más que desarrollar esa información reelaborándola de forma mecánica.

Para el atomismo de la moderna teoría del conocimiento, la impresión aislada contiene una información autosuficiente; posee toda la existencia particular, separable, de un objeto exterior, y no requiere una hermenéutica. En la concepción de Locke, por ejemplo, las ideas simples se comportan como materiales de construcción (1984, 2.2.2). Las condiciones de comprensión están insertas en los elementos y procesos como propiedades internas. La poderosa influencia que esta concepción neutralista y objetivante ejerce sobre nuestro pensamiento y nuestra cultura tiene mucho que ver con el predominio de instituciones y prácticas que exigen una actitud descomprometida, una desconsideración de las condiciones morales de nuestro mundo de la vida: en la ciencia y en la técnica, en los modos racionalizados de producción, en la administración burocrática, etc. Y es la concepción epistemológica en la que todavía se apoya la inteligencia artificial.

Traer a colación lo implícito como propiedad específica del conocimiento humano implica entendernos como seres con trasfondo, como sujetos que se mueven en el mundo de la vida. Este mundo de la vida es recuperado por Heidegger y Wittgenstein, ocupando el lugar de «aquello mas allá de lo cual no es posible ir» (*das Unhintergehbare*): en Heidegger bajo la forma del «proyecto arrojado» del ser-en-el-mundo y en Wittgenstein como aquellas formas de

vida que configuran el trasfondo de los juegos del lenguaje en los que siempre nos movemos. De acuerdo con esta precomprensión, lingüística e históricamente condicionada, no nos es posible retroceder al punto cero de un pensamiento libre de prejuicios, a un lugar en el que la realidad se nos ofrezca sin necesidad de interpretación. En vez de una pretensión de fundamentación trascendental última, se nos remite a la facticidad de que no podemos dejar de dar como ciertos determinados presupuestos de la argumentación y de la praxis de la vida. Esa corporalidad y finitud, esas formas de lo implícito que la inteligencia artificial tiene tantas dificultades en reproducir, es lo que más nos especifica frente a las máquinas. Un avance significativo de la inteligencia artificial requeriría algo similar a un giro interpretativo, el descubrimiento del trasfondo de todo conocimiento: que las máquinas desarrollaran un equivalente funcional a nuestro saber implícito.

§ Inexactitud

He agrupado bajo el término «inexactitud» un conjunto de propiedades que caracterizan nuestra inteligencia y cuya diferencia con la de las máquinas es muy significativa. Me refiero al hecho de que los humanos estamos continuamente pensando en aproximaciones, de que somos inteligentes no porque aplicamos fielmente reglas establecidas sino porque tenemos una especial capacidad para atender a lo singular, todo lo cual por cierto nos inclina a cometer un cierto tipo de errores (que también nos distinguen de los de las máquinas). Se podría sintetizar nuestra condición inteligente afirmando que la inteligencia humana tiene su fuerza «en su imprevisible ambigüedad e imperfección» (Bodei 2019, 317).

Muchos de los sesgos de los algoritmos tienen que ver con la misma naturaleza de los datos: por ejemplo, que el machismo o el racismo está presente en los textos que se analizan. Este problema podría corregirse al menos parcialmente reequilibrando los sesgos. La cuestión de fondo tiene, no obstante, un carácter estructural. El problema del aprendizaje de los sistemas es que son incapaces de ver cada caso como un caso particular; están concebidos para construir estereotipos a través de una gran cantidad de datos. Su fortaleza consiste en que extraen las características que se repiten dejando al margen las propiedades raras, variables y contingentes. No solo es que se apoyen en los estereotipos, sino que están calculados para maximizar la conformidad a los estereotipos, hasta tal punto que no reparan en la diferencia y la novedad. El sistema no nos ve sino a través de las propiedades por las que no somos un caso único; es incapaz de hacerse cargo de la «economía de las singularidades» (Karpik 2007). Aunque los seres humanos estamos continuamente aplicando reglas, no nos limitamos a ello. La inteligencia humana no consiste en actuar conforme a reglas (Innerar¡ty 2011). Una inteligencia consistente en aplicar reglas sería bastante limitada. De ahí que buena parte de la comunidad investigadora se haya volcado en los últimos años en el análisis de las redes neuronales, que funcionan de otra manera.

En el origen de nuestros principales fracasos colectivos hay una manera de pensar que entiende el conocimiento como la consecución de exactitud, la comunicación

como transmisión estandarizada de informaciones y la organización política de la sociedad como una gestión de objetividades. Con frecuencia ha ocurrido que las pretensiones de exactitud o el desarrollo irreflexivo han dado lugar a decisiones irracionales y solo las culturas de interpretación (esos entornos críticos en los que se interroga por la inserción social de las tecnologías, se discuten sus aplicaciones sociales, se hacen valer criterios éticos y políticos) han conseguido corregir su inexactitud social. El saber requiere libre acceso a la información, pero también capacidad de eliminar el «ruido» de lo insignificante.

¿Qué lugar ocupa, por así decirlo, el error en la inteligencia de los humanos y en qué sentido podemos afirmar que nuestros errores son muy diferentes de los de las máquinas y en cierto sentido son lo que nos singulariza frente a ellas? El factor humano es otra manera de decir el error humano. Existe de hecho un área de los estudios de la interacción entre humanos y máquinas que examina cómo optimizar y corregir nuestros fallos, de qué modo las máquinas pueden lidiar con los típicos errores humanos.

Alan Turing formuló en los años cuarenta una idea de máquina de la que no se podía esperar infalibilidad porque el malentendido o el error forma parte de la inteligencia humana. Ahora bien, ¿es esa falibilidad computacional la misma que caracteriza a la condición humana? Es verdad que los sistemas pueden detectar anomalías, pero eso solo lo hacen de manera estadística y si saben antes qué fenómenos deben vigilar, mientras que los seres humanos son capaces de registrar una anomalía a partir de un solo caso. Esto no lo pueden hacer unos sistemas cuyo funcionamiento se basa en la explotación estadística de datos y para los cuales no hay nada extraño en el sentido lógico del término.

Christian Szegedy mostró que se podía engañar a cualquier red que hubiera aprendido el reconocimiento de imágenes. Podía fallar incluso en lo que mejor sabe hacer, la clasificación. Para un sistema de *deep learning* no es fácil implementar un modelo de inferencia causal; debería conocer cómo «funciona» el mundo de un modo más complexivo. Esta es la razón por la cual también una red neuronal es presa fácil de sabotajes minúsculos y estúpidos *(adversarial atacks),* como introducir un pequeño objeto insignificante que distorsione una imagen para confundir a la interpretación automatizada. Alguien podría objetar que también los humanos caemos en tales trampas e incluso nos engañamos solos. El problema es hasta qué punto podemos calificar como inteligente a una máquina que comete errores en los que no caería un niño. Es necesaria una cierta estructura mental básica (modelos del mundo, sentido común, idea de causalidad) sobre la cual edificar el aprendizaje automático y profundo que mejora la experiencia.

§ Economía

Una de las propiedades más asombrosas de la inteligencia humana es su economía, es decir, la poca energía que requiere para funcionar óptimamente. Se trata de una sobriedad desde el punto de vista ecológico y epistémico. De lo primero da buena prueba la comparación entre el pequeño consumo de energía del pensamiento y el almace-

namiento insostenible de los datos en la nueva economía del *big data,* muchos de cuyos centros se construyen ya cerca de centrales de alta producción de energía y que tarde o temprano plantearán problemas de sostenibilidad.

La segunda dimensión de esta economía es de naturaleza epistémica. El *deep learning* requiere generalmente una inmensa cantidad de datos. Deep blue tuvo que procesar 200 millones de posiciones por segundo para generar todas las soluciones potenciales en una partida de ajedrez. AlphaGo necesitó 30 millones de juegos para superar al hombre, muchos más que los juegos que un humano podría jugar en toda su vida. Los humanos, en cambio, no tenemos necesidad de muchos datos para aprender y generalizar. Los humanos, especialmente los niños, son excelentes para aprender de manera instantánea, a partir de uno o dos ejemplos. Hay una gran diferencia entre nuestro aprendizaje y el de las redes neuronales. Se podría afirmar que una de las propiedades mas específicamente humanas y de nuestra inteligencia es precisamente esta de pensar y decidir sobre asuntos y en situaciones para las que nunca habrá suficientes datos. Los humanos somos expertos en detectar estructuras porque podemos simplificar, llevar a cabo una compresión de la información (Chaitin 2004): «*comprehension is compression*». La compresión estadística que una máquina puede hacer implica una pérdida, en la medida en que cierta información detallada es ignorada. Los humanos sabemos comprimir sin pérdida y lo hacemos gracias a la identificación de estructuras. Este mecanismo de aprendizaje es el que ponemos en marcha en las analogías y las visiones de conjunto. Simplificar identificando estructuras implica una reflexión que va mas allá de la clasificación refleja y aquí estriba una divergencia fundamental entre la inteligencia artificial y la nuestra.

Esta ecología mental puede tener un valor especial en unos momentos en los que los datos no son la solución sino, en cierto sentido, el problema. Es cierto que sin datos no hay información ni conocimiento, pero buena parte de la actual perplejidad epistémica procede del hecho de que, a partir de una determinada cantidad, lo que al inicio era un problema de carencia de información se transforma en una desorientación debida al exceso. Resulta especialmente valioso el aprendizaje de la «economía del informarse» (Downs 1957), un trato selectivo con los datos, superficial podríamos decir, una suerte de «tacañería cognitiva» (Wirth / Matthes 2006) gracias a la cual los seres humanos desarrollamos una «racionalidad de baja información» (Popkin 1991, 7). Esta economía tiene también su dimensión en la práctica; sabemos que quienes deciden con más información no son necesariamente los que toman mejores decisiones (Kahneman 2003, 1469). Mas aún: algunos autores aseguran que bajo determinadas condiciones pueden beneficiarse de una información escasa y decidir mejor (Gigerenzer / Goldstein 1996, 652).

Esta propiedad de la inteligencia es muy importante en entornos poblados de datos masivos que abruman sin orientar, algo tan típico del actual ecosistema informativo, caracterizado por la insostenibilidad epistémica del exceso. No hay ninguna función de las máquinas que emule a los humanos en esta economía de la atención y la deci-

sión, más bien al contrario: aquellas se caracterizan por necesitar incomparablemente más datos e informaciones que nosotros para el mismo resultado. ¿Es compatible la idea misma de la inteligencia artificial basada en la disposición de datos ilimitados con el desarrollo de alguna capacidad semejante a esa simplificación con sentido que a los humanos nos resulta relativamente fácil?

Las ciencias y las letras

Nuestro paisaje epistemológico es mucho más rico que el definido por las famosas «dos culturas científicas» (Snow 1959), y establecer rangos y primacías es una empresa que carece de interés epistemológico. Ahora bien, la organización del saber y, sobre todo, la configuración de modelos y aspiraciones siguen siendo muy parasitarias de esa contraposición y de la jerarquía implícita que rige en las categorías de lo científicamente acreditado. El modelo dominante de inteligencia artificial sigue debiéndole mucho a la entronización de un rigor técnico-científico cuya exactitud y sofisticación es muy cuestionable.

En este contexto las ciencias humanas y sociales, incluyendo la filosofía, tienen que jugar un papel muy importante, que ya no es algo subsidiario o complementario, sino que atañe al núcleo de la empresa epistemológica, de lo que debemos entender por inteligencia y que puede contribuir decisivamente a dilucidar el futuro de la inteligencia artificial. Si, como señalaba al inicio, la inteligencia artificial solo avanzará decisivamente si consigue parecerse más a la inteligencia humana, los saberes que tienen que ver con el *sensemaking* pueden aclarar en qué consiste propiamente esta capacidad que el modelo vigente de inteligencia tiende a descuidar, seducida por la impresionante cantidad de los datos o la rapidez del cálculo. No se trata de invertir las jerarquías actuales ni de reivindicar ningún oficio, sino de consolidar la cooperación entre *techies* y *fuzzies* para abordar juntos un problema que requiere integrar todas las dimensiones de la inteligencia. Hay ya muchas voces que reclaman un papel para las ciencias humanas y sociales en el abordaje de esta cuestión (Bradshaw 2014; Madsbjerg 2017; Hartley 2017; Marcus / Davis 2019). La intuición de que este camino es más provechoso para el futuro de la inteligencia artificial que dejar que las inteligencias mutiladas continúen sus caminos paralelos es por el momento una línea de investigación tan prometedora como poco desarrollada.

Quisiera despojar a esta afirmación de toda posible connotación de superioridad o «ventaja competitiva», pero las ciencias humanas y sociales pueden proporcionarnos, en un momento en que esto es más necesario que nunca, una visión mas holística de la realidad (Madsbjerg 2017, 57). Las tecnologías de la automatización y la inteligencia artificial resultarán incomprensibles si continúan siendo «socialmente inexactas», si no las entendemos como una totalidad que incluye también el modo como configuran las realidades sociales o modifican nuestro comportamiento (Hartley 2017, 128).

Puede que los *fuzzies* estemos especialmente entrenados en el tipo de cuestiones que deben identificar los sesgos y otras circunstancias que explican cómo se configuraron los datos. Esta aptitud se debe a que somos capaces de poner a los datos en el

contexto social en el que se recogieron, porque podemos presentarlos con claridad y en una narrativa coherente, haciéndolos públicamente accesibles, algo de lo que no es muy capaz el actual *data scientist.* Y me atrevo a asegurar que el trabajo de los *smart questioners* (Floridi 2014, 129) puede resultar más útil para el futuro de la inteligencia artificial que el de los funcionarios de la respuesta. Toda la dificultad del asunto consiste en cómo hacer que la automatización no se haga a costa del sentido, que la precisión sea compatible con la visión general. Retomando el debate al que hacía referencia al principio, la gran paradoja de la inteligencia artificial es que solo será general si reconoce sus límites y que únicamente podría sustituirnos si renuncia a hacerlo y opta por parecérsenos.

Bibliografía

Anderson, C. (2008), «The End of Theory: The Data Deluge Makes the Scientific Method Obsolet», *Wired,* june 23, 2008, http://www.wired.com/2008/06/pb-theory/

Bodei, R. (2019), *Dominio e sottomissione.* Bolonia: Il Mulino.

Bradshaw, L. (2014), «Beyond Data Science: Advancing Data Literacy», *Medium* (blog), December 17, 2014: https://medium.com/the-many/moving-from-data-science-to-data-literacyasf181ba4167#.bwiz7hc1g

Carr, N. (2014), *Glass Cage. Automation and Us.* Nueva York: Northon & Company.

Chaitin, G. (2004), *Grenzen und Grenzenüberschreitungen.* Berlín: Akademie Verlag.

Copeland, J. (1993), *Artificial Intelligence. A Philosophical Introduction.* Oxford: Blackwell.

Dessalles, J.L. (1996), *L'ordinateur génetique.* París: Hérmes Science.

— (2019), *Des intelligences TRÈS artificielles.* París: Odile Jacob.

Domingos, P. (2015), *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World.* Nueva York: Basic Books.

Downs, A. (1957), *An Economic Theory of Democracy.* Nueva York: Harper & Bros.

Floridi, L. (2014), *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality.* Oxford: Oxford University Press.

Gigerenzer, G.; Goldstein, D. (1996), «Reasoning the Fast and the Frugal Way: Models of Bounded Rationality», en: *Psycological Review,* 103/4, 650-669.

Hartley, S. (2017), *The fuzzy and the techie. Why the Liberal Arts Will Rule the Digital World.* Boston / Nueva York: Mariner.

Heidegger, M. (1967), *Sein und Zeit*. Tubinga: Niemayer.

Hofstadter, D. (1979), *Gödel, Escher, Bach: an Eternal Golden Braid*. Nueva York: Basik Books.

Innerarity, D. (2011), *La democracia del conocimiento.* Paidós: Barcelona, 2011 (traducción inglesa: *The Democracy of Knowledge*. Nueva York: Continuum, 2013).

Julia, L. (2019), *L'Intelligence Artificielle n'existe pas.* París: First.

Karpik, L. (2007), *L'économie de la singularité.* París: Gallimard.

Kahneman, D. (2003), «Maps of Bounded Rationality: Psychology for Behavioral Economics», en: *The American Economic Review,* December, 1449-1475.

Kurzweil, R. (2001), «Response to Mitchell Kapor's "Why I Think I will Win"», *Kurweil Accelerating Intelligence Essays*. http://www.kurzweilai.net/response-to-mitchell-kapor-s-why-i-think-i-will-win

LeCun, Y.; Bengio, C.a; Hinton, G. (2015), «Deep learning», *Nature* 521, 28. Mai 2015, 436-444.

Leibniz, G.W. (1950), *De arte combinatoria: Die philosophische Schriften,* vol. 4. Berlín: Weidmann.

Locke, J. (1984), *An Essay Concerning Human Understanding,* ed. P. H. Nidditch. Oxford: Clarendon Press.

Madsbjerg, C. (2017), *Sensemaking. The Power of the Humanities in the Age of Algorithm.* Nueva York: Hachette.

Malik, O. (2013), «Uber, Data Darwinism and the Future of Work», *Gigaom,* March 17, 2013, https://gigaom.com/2013/03/17/uber-data-darwinism-and-the-future-of-work/

Marcus, G.; Davis, E. (2019), *Rebooting AI. Building Artificail Intelligence We Can Trust.* Nueva York: Pantheon.

Moravec, H. (1988), *Mind Children*. Harvard University Press.

Pinker, S. (2019), «Tech prophecy and the underappreciated causal power of ideas», en: Brockman, J. (ed.), *Possible Minds. 25 Ways of Looking at AI.* Nueva York: Penguin, 100-112.

Popkin, S. (1991), *The Reasoning Voter: Communication and Persuasion in Presidential Campaigns.* University of Chicago Press.

Snow, C.P. (1959), *The Two Cultures and the Scientific Revolution.* Cambridge University Press.

Stiegler, B.d (2017), *The Automatic Society. The Future of Work.* Hoboken: Wiley.

Wilczek, F.k (2019), «The unity of intelligence», en: Brockman, J. (ed.), *Possible Minds. 25 Ways of Looking at AI.* Nueva York: Penguin, 64-75.

Wirth, W.; Matthes, J. (2006), «Eine wundervolle Utopie? Möglichkeiten und Grenzen einer normativen Theorie der (medienbezogenen) Partizipation im Lichte der neueren Forschung zum Entscheidungs- und Informationshandeln», en: Imhof, Kurt; Blum, Roger; Bonfadelli, Heinz; Jarren, Otfried (eds.), *Demokratie in der Mediengesellschaft*. Wiesbaden: VS Verlag für Sozialwissenschaften, 341-361.

Wittgenstein, L. (1971), *Philosophische Untersuchungen.* Frankfurt: Suhrkamp.

Daniel Innerarity
Ikerbasque
Universidad del País Vasco
Barrio Sarriena s/n
48940 Leioa
dinner@ikerbasque.org

Recibido: 5 de diciembre de 2019
Aprobado: 17 de febrero de 2020