

¿Es posible desarrollar una “conciencia artificial”?¹

Juan Arana

Real Academia de Ciencias Morales y Políticas. Madrid

Resumen. Los rápidos progresos de la Inteligencia Artificial y la generalización de su presencia en la vida social y económica han suscitado mucha satisfacción por sus crecientes beneficios, pero también bastantes recelos hacia sus peligros. La sustitución masiva de empleados por dispositivos “inteligentes” capaces de asumir con ventaja los mismos cometidos va a exigir reformas muy profundas. Habrá que reformular el papel desempeñado por los ciudadanos: tanto la educación como la organización laboral van tener que ser planteadas de otra manera. Incluso hay quienes temen que esta tecnología, hoy por hoy dócil a nuestras apetencias, empiece a adoptar sesgos hostiles, a medida que resortes de poder cada vez más significativos estén en sus manos, dado que en la actualidad ya posee cierta autonomía para irse reprogramando por sí misma. La ponencia examina las medidas que los principales expertos en la materia aconsejan para que eso no ocurra y también llama la atención sobre la amenaza que supone la pérdida de confianza de la humanidad en la capacidad de controlar sus propias decisiones y destino.

1. *Inteligencia e inteligencias*

Geoffrey Hinton, premio Nobel de física de 2024 y popularmente designado por los medios “padrino de la inteligencia artificial”, sostuvo recientemente una larga entrevista con Gustavo Entrala, en la que salió a relucir no sólo su inmenso saber sobre las redes neuronales artificiales y las perspectivas de futuro desarrollo para esta tecnología, sino también cuestiones de más amplio alcance relacionadas con su concepción del mundo y los enigmas filosóficos subyacentes. A estas alturas de la historia nadie pone seriamente en duda que el desafío de construir máquinas inteligentes ha sido coronado con éxito. Éxito que sin embargo también ha puesto de manifiesto la ambigüedad y polisemia del vocablo. La mera costumbre explica que sigamos usándolo en singular, porque en cuanto conviene precisar se desdobra en tantos tipos de inteligencia como tareas es capaz de abordar la mente humana, e igualmente los dispositivos que la asisten o emulan. De hecho, la mejor forma de identificarlos como tales sigue siendo decir que resuelven tareas que, si las solventase una persona, diríamos de ella que es inteligente. Inteligente sí, pero, ¿de qué tipo? Howard Gardner ha identificado doce clases diferentes (como no estoy del todo al día de sus progresos, seguramente habrá llegado a distinguir alguna más): lingüístico-verbal, lógico-matemática, visual-espacial, musical-auditiva, corporal-kinestésica, interpersonal, intrapersonal, naturalista, emocional, existencial, creativa, colaborativa...

Sin duda conviene también empezar a hablar de “metainteligencia”, o sea, de la aptitud para coordinar sabiamente todas sus variantes. Quien conciba la relación entre inteligencia humana y artificial como una competencia, tal vez pueda consolarse de las derrotas que hemos sufrido últimamente de esta manera: así como los lógicos distinguen entre lenguajes y metalenguajes, y estos últimos con toda una escala de niveles, también podríamos establecer metainteligencias de nivel uno, dos, y de todos los que conviniera para burlar el acoso de la artificialización. Al fin y al cabo, jugar bien al ajedrez era antaño un don reservado a los muy inteligentes. Supongo que de alguna manera sigue siéndolo, pero es innegable que su prestigio ha resultado erosionado por la invención de programas informáticos que doblegan al campeón mundial, y otro tanto hay que decir del *go*, el *jeopardy!* y tantos otros desafíos. La posibilidad de descollar en menesteres lúdicos está en franca decadencia y en el terreno de las profesiones tampoco van mejor las cosas. Laurent Alexandre comenta en su libro sobre *La guerra de las inteligencias* que los sistemas expertos —o quizá habría que decir ya *superexpertos*— están comiendo el terreno a pasos agigantados a los que en un principio tan sólo imitaban. ¿Qué van a hacer en el futuro los doctores en medicina cuando un simple programa esté en condiciones de valorar con plena garantía y en un abrir y cerrar de ojos una analítica que contenga miles de ítems en lugar de las pocas decenas que contiene el hemograma tipo?

¹ Agradezco Javier Bernácer, Concha Diosdado, Francisco Rodríguez Valls y María Caballero los estímulos y sugerencias aportados para la realización de este trabajo.

2. La revolución de las máquinas

El problema no es tan sólo que nuestros esclavos de silicio trabajen 24 horas cada uno de los 365 días del año. Lo que produce vértigo es que, sin tomar ni precisar el menor descanso, una y otra vez duplican sus prestaciones, de manera que según sus heraldos y profetas lo que les falta aún para dejarnos atrás se conseguirá antes de que queramos darnos cuenta. ¡La singularidad está cerca! Los que escribimos con estilográfica la tesis doctoral no necesitamos que nadie nos lo explique: cabalmente lo hemos ido viendo a lo largo de nuestra vida laboral, hasta el punto de que ha sido un alivio ver llegar la jubilación con la edad reglamentaria, en lugar de resultar arrollados por el tsunami de la IA. Hinton recuerda que las primeras revoluciones industriales convirtieron en superfluo el trabajo de abrir zanjas a mano, pero se pudo colocar a toda la mano de obra sobrante para que se hiciera cargo del papeleo.

Los dos últimos siglos han resultado ser la edad de oro de la burocracia, pero este reducto antaño inagotable de horas de trabajo ampliables *ad libitum* está herido de muerte. Incluso las más crípticas labores oficinescas están siendo arrebatadas de nuestras manos para ser transferidas a máquinas supuestamente desalmadas. Primero se hicieron con el trabajo manual no especializado, luego también con el especializado. La primera oleada de la informática convirtió en prescindible el trabajo rutinario de bajo nivel; ahora la caducidad les llega igualmente a las rutinas de alto nivel. A este paso sólo vamos a servir (y ya veremos por cuánto tiempo) para pasear perros y darnos conversación unos a otros. Deberemos hacer de tripas corazón y exclamar unamunianamente: “¡Que trabajen, inventen, curen, escriban tesis y novelas, compongan música, etc., etc., *ellas!*”

A muchos acongoja la perspectiva de que robots y ordenadores asuman definitivamente el control y —como primera o segunda providencia— tomen la decisión de eliminarnos. Eutanasia universal que, con todos los resortes del poder en sus manos (producción, distribución, transporte, ¡armamento!) podría consumarse en un lapso de horas o incluso minutos. Casi tan triste como esa apocalíptica perspectiva me parece la idea de que —se nos orille del todo o no— nos volveremos de aquí a poco *perfectamente prescindibles*. Tampoco resulta muy consoladora la idea de ser convertidos en los “osos panda” de la segunda mitad del siglo XXI: encerrados en parques “antropológicos” parecidos a los actuales zoológicos, seríamos exhibidos y mostrados a la curiosidad de lo trans- o posthumano.

Nuestros brazos no pueden competir con los robóticos; las piernas tampoco con los medios mecánicos de transporte y, según se asevera, tampoco los cerebros con las computadoras que ya están a punto de diseñarse. ¿Qué habilidades poseemos que las máquinas no tengan ya o estén a punto de conseguir, arbitrariamente potenciadas? Los luditas que en el siglo XIX intentaron destruir telares mecánicos serían una pálida versión de lo que va a ocurrir: toda la humanidad va a tener que hacerse a un lado, de grado o por fuerza. Si en el caso de nuestros bisabuelos eran los obreros manuales los que se habían vuelto prescindibles, ahora amenazan con serlo también los de la *inteligencia* o, si queremos decirlo de un modo más grandielocuente, los del *espíritu*. Retóricas aparte, la única alternativa de supervivencia que no sea la mera *compasión* que podamos inspirar a nuestras competidoras es encontrar alguna competencia para la que sigamos siendo insustituibles. Dan qué pensar esos jóvenes que atraviesan nuestras calles montados en bicicletas o patinetes con gruesos fardos a la espalda, sin otro cometido que repartir mercancías que han sido ideadas, fabricadas y suministradas por plataformas logísticas, todo ello de modo automático. ¿Cuánto falta para que empiecen a fabricarse robots capaces de salvar este último e insignificante escollo?

3. ¿Tiene un futuro la mente humana?

Lo de “inspirar compasión” a los productos de nuestra inventiva tecnológica hay que entenderlo de un modo estrictamente literal. Según Geoffrey Hinton nuestra única y última oportunidad consiste en programar *instintos maternos* en las máquinas inteligentes cuyo diseño todavía controlamos. Llega a esa conclusión tratando de responder a cómo es posible que un ser menos inteligente controle a otro que lo sea más. Se fija en el ejemplo del bebé: sin duda es mucho menos listo que su progenitora. No obstante, posee de modo innato la habilidad de inspirar sentimientos de amor y protección que son la clave de su supervivencia. Basta con que la madre se distraiga un poco para que la cría muera, pero son rarísimos los casos —por desgracia últimamente no tanto— en que falla este soporte tan esencial. Podemos confiar en la solicitud maternal y con matices también en la paternal, porque la selección natural ha establecido refuerzos muy potentes que actúan en tal sentido. De hecho, todas las

especies biológicas que no lo consiguieron han desaparecido. Sin embargo, sobre la evolución de máquinas creadas por el hombre no opera la selección natural: el mecanismo evolutivo que hay detrás de la IA es tan artificial como ella misma. Por lo menos lo ha sido hasta hace muy poco: estas criaturas fueron producto del arte desplegado por los humanos. Se explica que fueran perfectamente obedientes, salvo cuando el diseño fallaba y los accidentes o el uso las estropeaban. Si todo iba bien eran perfectamente *deterministas*; sus creadores y propietarios tenían en la mano la llave infalible de su comportamiento. ¿Qué importa que yo ignore los secretos del motor de explosión si sé que apretando el acelerador la velocidad aumenta? Se supone que el ingeniero que lo ideó sí conocía sus secretos íntimos, aunque no tuviera la fuerza y robustez de su criatura. El inventor del ábaco tampoco poseía la precisión del instrumento, ni quien sabe cómo construirlo consigue sumar con la misma eficacia que el fruto de su destreza. De modo análogo, el objeto físico “libro” tiene una memoria mucho más fiel que la mente que lo compuso. Curiosamente, la búsqueda de soportes magnéticos no ha sido tan afortunada: ninguno de ellos puede resistir incólume el paso de los siglos como lo consiguen con bastante frecuencia el papel y la tinta. Así pues, no hay nada de extraordinario en el hecho de que los productos del ingenio humano tengan mayor potencia calculística, memoria, precisión, durabilidad, etc., que los propios humanos. Pero durante mucho tiempo tuvimos una superioridad indiscutible en el aspecto *fundacional*: de la misma manera que los padres se sienten orgullosos de que su hijo consiga llegar más lejos de lo que ellos mismos alcanzaron, la humanidad como colectivo no debe avergonzarse de que sus inventos la vengzan en las más variadas contiendas. A principios de los años 90 Marvin Minsky, uno de los padres fundadores de la IA, se preguntaba: “¿Serán los robots quienes hereden la Tierra?”, y él mismo respondía: “Sí, pero serán descendientes nuestros.” Supongo que pensó que quien no logra la paternidad biológica siempre tiene a su alcance la adoptiva. El que no se consuela es porque no quiere, pero hay consuelos que pueden resultar bien amargos, como cuando el retoño que uno alumbra o prohija resulta ser un parricida.

Hinton defiende que aún es posible —si ponemos suficiente empeño— asegurarnos de que no sea así. A tal fin deberemos insuflar en las máquinas que nos den el relevo un sesgo irreversiblemente agradecido hacia la especie biológica que les dio el ser. Frente a los recelosos que, como Nick Bostrom, desconfían de la actitud que adoptarán cuando se vuelvan superinteligentes, los hay optimistas como Ray Kurzweil, que tiene la esperanza de que, cuando la inteligencia informática alcance su plenitud singular, en lugar de ser asesinados por ella nos obsequiará nada menos que la inmortalidad, gracias a una pléyade de bienhechores nanorrobots inteligentes que colonizarán nuestros torrentes sanguíneos inundándonos de beneficios. Ciertamente es que este mismo autor prevé la contingencia de que esos simpáticos enanitos del bosque arterial sean infectados ellos mismos por virus informáticos, lo cual para nuestra especie podría resultar más catastrófico que los síndromes producidos por los virus biológicos. Entre el pesimismo de Bostrom y el optimismo de Kurzweil media la “prudente” estimación de Hinton, según la cual ahora mismo “sólo” hay entre un 10 y un 20 por ciento de probabilidad de ser masacrados en masa por una superinteligencia perversa. Tanto el advenimiento de la vida eterna como la consumación del exterminio forzoso se anuncian como acontecimientos muy próximos: cosa de 20, 30 ó 40 años a lo sumo.

4. *Naturalismo y conciencia*

Mientras los ingenieros de Silicon Valley y los megacentros de investigación chinos siguen afanándose por cocinar esta confusa mezcla de infierno y paraíso terrestre, los que recelen de la fiabilidad de estas prospectivas pueden preguntarse, como yo mismo hago, sobre la existencia de factores que no hayan sido tenido en cuenta por estos avisadísimos programadores ni tampoco por las máquinas que ya empiezan a competir con ellos para acabar de iluminar o ennegrecer el futuro. Casi todos ellos coinciden en asumir una *filosofía naturalista*, según la cual todo lo que de relevante hay en los seres humanos es competencia de las ciencias naturales: interacciones físico-químicas sometidas a las leyes que las regulan. Sólo contemplan una suma de necesidad natural con una porción de azar estocástico ayuno de sesgos. Están convencidos de que no existen factores espirituales o sobrenaturales, “ganchos colgados del cielo”, sujetos autónomos, entidades realmente libres ni nada por el estilo. Suponen que la porción de azar es suficientemente pequeña como para que no sea necesario tener en cuenta la imprevisibilidad de los aparatos que construyen. Por lo que se refiere a las entidades biológicas y muy en particular a los cerebros humanos, todo esto no deja de ser una presunción bastante arriesgada, pero se acepta por la simplificación que aporta. En lo que concierne a los dispositivos electrónicos que llevan a cabo las operaciones lógicas programadas, se confía en la eficacia de los diseños y en el rigor de los procesos de fabricación. En lo concerniente a la

programación, tan sólo está contemplada la necesidad lógica, con la inserción —si hiciera falta— de rutinas aleatorias o pseudoaleatorias.

Geoffrey Hinton es representativo en este sentido: se confiesa ateo y afirma con todo aplomo que los seres humanos somos “máquinas”. Así pues, profesa un *mecanicismo antropológico* rampante. ¿Por qué está tan seguro? El único argumento que aporta al entrevistador no es muy original y sí bastante elemental: Si consiguiésemos extraer una neurona enterita del cerebro y sustituir todas sus conexiones sinápticas por un sucedáneo electrónico, seríamos incapaces de notar la diferencia: nuestros estados mentales con neurona o con prótesis resultarían indistinguibles. Repitiendo el proceso unas ochenta o noventa mil millones de veces, completariamos la traducción mecánica de toda la psique. El ejemplo es sin lugar a dudas pésimo: cualquiera que haya contemplado una representación tridimensional de la red cerebral, percibe de inmediato el intrincadísimo laberinto que hay allí, los cientos de sinapsis y entrelazamientos que rodean a cada neurona. Extraer una sola de ellas sin alterar las que la rodean es un desafío absolutamente inabordable: nada que ver con el más complejo *mikado*. Y suponiendo que consiguiéramos sustituir una de ellas cada minuto, tardaríamos más de 170.000 años en replicar un solo cerebro. Hinton, por supuesto, aclara que se trata de un mero *experimento mental*, pero incluso en calidad de tal podría haberse esforzado un poco más. En el libro *El hombre mecánico* Hans Moravec propone una versión bastante más plausible: en vez de neuronas aisladas postula emulaciones informáticas de trocitos del tejido cerebral y la eliminación paulatina de la matriz biológica cada vez que un *bypass* funcional lo autorice. La idea resulta tan inverosímil como la de Hinton, pero por lo menos no choca de un modo tan grosero con la imaginación del lector.

El mecanicismo neurológico resultaba más verosímil en neurocientíficos afectos al viejo determinismo físico, o en investigadores de inteligencia artificial que practicaban la computación en serie. Esta actitud parece más empeñada en explicar lo que la mente *no puede hacer*, que en arrojar claridad sobre lo que *de hecho hace*. Los neurocientíficos aprendieron que bastante tenían con estudiar los sistemas nerviosos que tienen delante, en lugar de soñar con mitos como el del *demonio de Laplace*; sobre todo una vez que la mecánica cuántica jubiló la idea que lo sostenía y las sucesivas teorías de sistemas complejos y sensibles a las condiciones iniciales mostraron que hay vías mucho más promisorias para avanzar. También los expertos en inteligencia artificial se dieron cuenta de que no se trataba tanto de cuadricular la mente como de remedar sus prestaciones: se trata de volar; da lo mismo que se consiga batiendo alas o girando hélices. En lugar de *dogmatismo*, *pragmatismo*. Esta ha sido la fórmula de éxito en los dos sectores del conocimiento que tocan más de cerca a lo mental. Dejar que los filósofos sigan entretenidos con sus juegos de palabras y sus controversias, para buscar soluciones prácticas a problemas concretos. Así han conseguido iniciar una senda de genuino progreso las ciencias que estudian el sistema nervioso de animales y humanos, y lo propio han hecho los que pretenden reproducir artificial —*ergo* mecánicamente— lo que la mente hace de modo natural y en particular algunas de las mil caras de la inteligencia. Por lo tanto, no se persigue hacer lo que la mente hace y exactamente como ella lo hace, sino generar un repertorio de funciones equiparables. Con ayuda de sus cerebros, los animales recorren a un ritmo más pausado las intrincadas sendas a las que están adaptados. Los que utilizan ordenadores sólo pueden usar las autopistas de la información y los motores del cálculo. Eso les obliga a dar largos rodeos para llegar a los mismos destinos alcanzados por vía cerebral: el recorrido siempre es más largo, pero se efectúa a una velocidad incomparablemente mayor. Si la ventaja es indudable a la hora de emular raciocinios mediante cómputos, lo es mucho más cuando se trata de asimilar, conservar y efectuar tratamientos de datos, o en la tesitura de buscar el modo más eficiente de hacerlo. En general no se pretende hacer pronósticos infalibles (el viejo y obsoleto ideal del determinismo), sino aprender de los propios errores, conseguir un ajuste progresivo de las expectativas a los resultados, o bien probar a toda velocidad un millón de posibilidades para elegir la que tiene mayor probabilidad de acercarnos a la meta elegida.

5. Ya no sabemos qué es lo que hacen exactamente las máquinas

Este cambio de actitud ha permitido multiplicar los éxitos en cascada. Kurzweil advierte que “la capacidad de los seres humanos para los procesos lógicos es muy limitada, pero sin embargo poseemos una profunda capacidad para reconocer patrones”. Ése es pues el camino a seguir. Con la generalización de las redes neuronales y los procedimientos de *deep learning*, la inteligencia artificial ha abandonado en parte el patrocinio de la lógica para hacer sitio a la inspiración biológica. Se suele decir que “el ojo del amo engorda al caballo”, pero estar demasiado encima de él tampoco es bueno. Los circuitos electrónicos que dan asiento a las máquinas lógicas son muy

obedientes, pero cuando se les prescribe con demasiado detalle el camino a seguir, suelen atascarse, y el programador tampoco sabe cuál es el procedimiento más idóneo —y práctico— para avanzar. Exigir que la solución sea *perfecta* conduce la mayor parte de las veces a tareas inacabables. El disciplinado aparato no se rebela, pero protesta silenciosamente pedaleando sin fin. Para evitar estos callejones sin salida, lo que ahora se le dice es: “Prueba a ver cuánto te acercas de esta forma a lo que busco y, a la vista del resultado, modifica el procedimiento según estas directrices y reinténtalo una y otra vez...” El resultado ha sido espectacular: es una estrategia que ha convertido la IA en la gallina de los huevos de oro. Según Hinton, las buenas ideas para llevar adelante todo esto suelen nacer en el contexto del trabajo universitario, pero los beneficios se cobran en las empresas, de manera que hay un transvase continuo de las mejores cabezas de un sector a otro. Él mismo sacó a subasta su capacidad de trabajo, llevándoselo junto con su equipo el mejor postor por una puja de 44 millones de dólares. De aquí surge el peligro de que, falta de financiación, la fuente última de inspiración acabe agotándose. Ese es sin duda alguna un problema. Pero como filósofo, aún me parece más preocupante que a la vista del éxito obtenido se caiga en la tentación de pensar que “esto es *todo* lo que hay”.

Si fuera cierto, la inteligencia artificial se convertiría en modelo obligado de cómo funciona también la natural. Ahora mismo nuestros ordenadores descubren lo que indagan aplicando instrumentos conceptuales tales como los algoritmos de retroprogramación, las máquinas de Boltzmann, las representaciones distribuidas, las máquinas de Helmholtz, los modelos jerárquicos ocultos de Márkov, etc. Animados por las prestaciones así logradas, los expertos conjeturan que nuestros cerebros tal vez apliquen procedimientos en algún sentido análogos, a los que habrían llegado en alas de la selección natural. La conclusión es que, si las máquinas de la IA no necesitan más aditamentos para obtener su brillante ejecutoria, tampoco los cerebros —ni las mentes que los usan— han precisado medios cualitativamente diferentes para triunfar en su lucha por la supervivencia. ¿O sí? Por lo que respecta al sistema nervioso central de otras especies vivas lo ignoro, puesto que carezco de acceso directo a su presunta interioridad psíquica. Es posible que, adoptando un punto de vista conductista, tampoco haga falta mucho más para explicar la percepción, memoria, inteligencia y motivación animal. Se trataría entonces de iniciar un proceso, no de “ingeniería inversa” sino más bien de “biología inversa”: lo que la informática descubre, la neurociencia reencuentra en otro contexto.

6. La excepción humana

Sin embargo, la evidencia empírica disponible es más amplia en el caso de las mentes y cerebros humanos, porque, gracias a la introspección, el acceso directo al sector consciente de la actividad nerviosa enriquece sustancialmente el panorama. Ciertamente es que la evidencia introspectiva directa es personal e intransferible, pero gracias al incomparable instrumento de comunicación que representa el lenguaje humano, somos capaces de verbalizar, transmitir y cotejar nuestras experiencias privadas. Solemos remitir este tipo de vivencias a una facultad específica que denominamos “*conciencia*” o también “*consciencia*”. En español ambos términos son prácticamente sinónimos y tienen connotaciones tanto cognitivas como prácticas (esto es: *morales*). En inglés se distingue más cuidadosamente entre unas y otras, empleando respectivamente los términos “*consciousness*” y “*conscience*”. Cuando el entrevistador español sacó a relucir el tema, Hinton recalcó que él se ceñía específicamente al aspecto cognitivo. Me parece una tontería discutir sobre el genio propio de un idioma u otro, pero creo que en este caso la mayor ambivalencia con que los castellanoparlantes manejan estas nociones tiene un significado que va más allá de lo anecdótico y que desde luego no tiene nada que ver con un uso más descuidado de la precisión semántica. Es obvio que lo teórico/cognitivo se mueve dentro de la esfera de la *verdad*, mientras lo práctico/moral está ubicado en la del *bien*. Por eso en el primer caso las valoraciones pertinentes serían: verdadero/falso, acertado/erróneo, etc., mientras que en el segundo hablaríamos en cambio de bueno/malo, correcto/incorrecto, etc. Sucede no obstante que, al menos en la tradición filosófica continental, ambos campos tienen bastante que ver entre sí: una cosa está de algún modo enlazada con la otra. Es una tradición heredada del intelectualismo ético griego: según Sócrates y después de él Platón, quien *de verdad* es sabio no puede ser *malo*, porque la voluntad depende para determinarse del bien que le presenta la inteligencia. Los pensadores anglosajones son menos proclives a ceder a las sugerencias de la metafísica y han levantado muros más altos para separar lo volitivo de lo intelectual. No es casual que fuese un filósofo británico, George E. Moore, quien introdujo la noción de “falacia naturalista”, definiendo como un error lógico la transición indebida del *ser* al *deber ser*.

Ahora bien, cuando se aísla lo éticamente deseable de lo ontológicamente verificable, se acaba condenando la primera de ambas instancias a un ostracismo imposible de redimir. El amplio dominio del ser cae bajo la competencia de la inteligencia, a la que se veda cualquier acceso al deber. Así se explica que alguien tan inteligente y sensible como Geoffrey Hinton afirme sin pestañear que él mismo, como los restantes miembros de la especie *homo sapiens*, es una “máquina”. Por supuesto, habría en ella toda la complejidad que quepa concebir; pero no hay sofisticación posible que la redima de ser una *mera cosa*. Sin embargo, apenas cinco minutos después, en el transcurso de la misma entrevista, Hinton se declara *humanista*, defensor de la especie humana y dispuesto a luchar por su supervivencia frente a las amenazas provenientes, bien del uso malvado de una IA todavía monitorizada por nosotros, bien de una superinteligencia artificial fuera de todo control. ¿Sobre qué base cabría legitimar esa toma de postura claramente ética? ¿En qué sentido son preferibles las maquinarias humanas a cualquier otro tipo de máquinas tanto “naturales” como “artificiales”? Si estas últimas llegan a superarnos claramente en todo tipo de tests de inteligencia, ¿no sería más lógico hacer mutis por el foro y dejarles definitivamente el campo libre? Eso es lo que —a mi juicio con mayor coherencia— defendía Marvin Minsky. Está claro que Geoffrey no ha puesto a trabajar en serio su privilegiado cerebro para responder a esto, puesto que sólo dice de pasada que le anima un sentimiento de solidaridad con su propia stirpe. Lo cual me recuerda que ya hace veinte o treinta años Minsky motejó (un poco en broma) de “*racista*” a la reportera que se atrevió a sugerirle que los humanos poseen “algo” que no tienen ni tendrán los más potentes computadores. La postura de Hinton sólo difiere de la que su predecesor criticaba en que no se avergüenza de asumir sentimientos racistas. Mucho me temo que su opción resulta a todas luces indefendible, a menos que renuncie también a cualquier pretensión de racionalidad.

No estaría mal en este contexto recordar a Immanuel Kant, porque, junto a la razón *teórica* (apta para asomarse a la grandeza del “cielo estrellado”) defendía la presencia de una razón *práctica* capaz de reaccionar ante la dignidad moral de los seres libres. Si nos ponemos un poco estrictos, este filósofo resulta sospechoso de haber cometido la mentada falacia naturalista, porque defiende que es una misma razón la que en su uso teórico explora la fachada fenoménica del ser, y en su uso práctico descubre la libertad como *factum* de la razón, la cual abre todos los horizontes del deber ser. Bien pensado, ¿por qué no? ¿Acaso Moore o cualquier otro consiguió demostrar que no hay ningún tipo de comunicación entre ambos reinos?

7. ¿Autoconciencia artificial?

Voy a concluir esta reflexión mostrando que el propio Hinton (y posiblemente cualquiera de los *gurús* de la IA fuerte) tiene en su mano —aunque no quiera verlos o usarlos— argumentos para abrir al menos una puertecita a Kant y su razón a la vez teórica y práctica. La clave está en que no da por perdida la posibilidad de reconciliarse con un posible adversario (nuestras propias criaturas inteligentes) que prevé cien o mil veces más poderoso que los humanos. Si le interpreto bien, para él la esencia de estas máquinas consiste en aplicar a rajatabla algoritmos cada vez más complejos, que ellas mismas están modificando de continuo *precisamente porque tales han sido las instrucciones que han recibido*. En seguimiento de esos protocolos se van haciendo cada vez más y más inteligentes en virtud de su propia disciplina. Nick Bostrom, por ejemplo, especula con la idea de una máquina superinteligente programada para optimizar la producción de “clips” de oficina. Esta orden, absolutizada sin restricciones, puede llevarle a acaparar las disponibilidades mundiales de acero y energía, y ver en la especie humana un obstáculo para cumplir a plena satisfacción su cometido, tomando así la fatal decisión de quitar del medio tan (para ella) molesta competencia.

Este ejemplo resulta demasiado chusco. Hinton, sin embargo, reflexiona que otorgar autonomía a los dispositivos de IA para fijarse por sí mismos objetivos secundarios no es ya omisible, porque se ha convertido en una exigencia impuesta por la lógica (y la eficacia) de su funcionamiento. Y cuando estos dispositivos se vuelvan muy inteligentes inevitablemente llegarán a considerar *su propia supervivencia* como una exigencia categórica, y para cumplir con ese objetivo empezarán a *mentirnos*, con lo cual ya estarán sobre la mesa todos los factores necesarios para incubar una rebelión letal para nuestra especie.

Obviamente, la única forma de evitarlo sería “parchear” los códigos fuente de los programas para que ninguno de sus objetivos secundarios autónomamente establecidos entre en conflicto con, por ejemplo, las conocidas leyes de Asimov, pensadas para garantizar la filantropía de las máquinas. “Sentimientos maternos” es la expresión que utiliza Hinton, y entiendo que lo que propone es troquelar su “ADN” informático, para que la eventual

superinteligencia de tales criaturas no se vuelva contra sus progenitores. Vamos a suponer que hacer tal cosa es, en efecto, posible. Eso demostraría *eo ipso* que, a pesar de sus portentosas virtudes, no tendrían una facultad indudablemente poseída por los seres humanos, esto es: la capacidad de cuestionar y revertir nuestras más íntimas convicciones y principios de conducta, mediante el ejercicio de la *autoconciencia*. Lo que diferencia la autoconciencia del ejercicio meramente cognitivo de la conciencia (en inglés *consciousness*) es que no se limita a generar representaciones objetivas de aquello que la ocupa. Cabría denominar *intencional* la conciencia meramente cognitiva, porque se limita a objetivar todo lo que ve (incluso cuando reflexiona sobre sí misma). Por eso cosifica irremediablemente todo lo que toca, mientras que la autoconciencia no lo hace, porque a través de ella el sujeto no se ve a sí mismo *como objeto* (como cuando un animal se reconoce ante un espejo, o alguien se autodescribe en un escrito), sino que se conoce de modo directo, sin mediaciones sensitivas ni conceptuales. Se trata por tanto de un *conocimiento subjetivo de lo subjetivo*. Ahí está el secreto de que otorgue control de sí a quien sea capaz de efectuar semejante ejercicio. Ello explica que la racionalidad práctica kantiana sea algo más que una ocurrencia filosófica: puede muy bien constituir la raíz tanto de la libertad como de la dignidad humana.

En cuanto a las máquinas superinteligentes, si realmente llegaran a darse, más nos valdrá que no figure entre sus talentos la autoconciencia. De lo contrario, la pretensión de Hinton de imbuirles instintos maternales hacia nosotros estaría irremediablemente condenada al fracaso: cualquier “parche asimoviano” que hubiera sido colocado en su entraña lógica sería perfectamente detectable y removible por ellas.

8. Epílogo para humanistas

Bien se ve que el acelerado progreso tecnocientífico está mostrando cada vez más a las claras su ambivalencia: los beneficios que proporciona son indiscutibles, pero los peligros que genera para mantener nuestra identidad y supervivencia tampoco son despreciables. La convivencia con los dispositivos que hemos ido creando y tal vez empiecen a ser desarrollados por sí mismos contempla tantas promesas como amenazas. La pregunta que se plantea con angustiosa urgencia es si existe alguna actividad importante —y por tanto remunerable— en la que no vayamos a ser suplantados por ellos. Si no fuera así, todavía quedaría en pie la esperanza de que la progenie humana poseyera en sí misma una *dignidad* que mantuviera su valor intrínseco independientemente de su presente o futura utilidad. Creo que ambas preguntas sólo podrían ser contestadas afirmativamente si los seres humanos poseyeran libertad, esto es, autonomía inalienable. Creo que hay un consenso bastante general de que las máquinas de toda clase que ahora mismo existen no la poseen, y la mayor parte de los entendidos piensa que seguirán sin tenerla aun cuando alcancen una superinteligencia basada en la futura IA. Así pues, todo el asunto depende de si nosotros la poseemos o no.

Pues bien, estimo que en una Academia como ésta tendríamos que tener el coraje de hablar sin tapujos. Entre los más destacados investigadores de la IA casi todos están convencidos de que no poseemos genuina libertad y lo mismo sostiene una amplia mayoría de los neurocientíficos en ejercicio. Abunda una creencia parecida entre los cultivadores de las ciencias humanas. En psicología, por ejemplo, no admitieron su existencia exponentes significativos del psicoanálisis, el conductismo o la psicología cognitiva. Burris Skinner llegó a publicar un libro titulado “Más allá de la libertad y la dignidad” donde sostenía que todos nuestros problemas se resolverán el día en que se consiga controlar el comportamiento humano de modo infalible. Se podría matizar que al fin y al cabo todos esos colectivos están interesadas en el contencioso y por tanto son sospechosos de parcialidad. Cabría esperar mayor objetividad entre los filósofos o —si se quiere— un sesgo que equilibre el de los grupos que acabo de mencionar. Personalmente soy un decidido defensor de la capacidad humana de autodeterminación, pero a decir verdad considero que estoy en minoría dentro de mi gremio. La mayor parte de las corrientes filosóficas en boga desde hace más de un siglo descreen de la libertad. Creo que esto vale para el cientificismo, el positivismo, la filosofía analítica, el materialismo, el marxismo y el neomarxismo, así como para los seguidores de casi todas las modas epistémicas de los quince últimos decenios. Solamente en el existencialismo, la fenomenología, la hermenéutica, el personalismo y otras corrientes secundarias el balance es más equilibrado, aunque el dato obedece con mucha frecuencia a la influencia que dentro de esas corrientes tienen los sentimientos religiosos. Algo parecido debería decirse de numerosas figuras relevantes, pero aisladas. En definitiva, no creo que la filosofía, el pensamiento humanístico, las artes o la literatura presenten en la actualidad un frente unido a los enemigos de la libertad. A la vista de semejante panorama, bien se nos puede aplicar el adagio de que quien siembra vientos recoge tempestades. Sospecho que la única posibilidad que tiene la humanidad de sobrevivir a la

presente coyuntura estriba en que quienes piensan que tan sólo somos producto de la necesidad y el azar descubran por sí mismos que están equivocados, porque el trabajo de los que se oponen a ellos deja bastante que desear.

Para seguir informándose

Entrevista G. Hinton con G. Entrala: <https://youtu.be/oroCGlPeXgs?si=4iDVEXt4s5HLG63Z>

Alexandre, Laurent, *La guerre des intelligences. Intelligence artificielle versus intelligence humaine*, Paris, J. C. Lattès, 2019.

Arana, Juan, *Qué es la conciencia*, Sevilla, Senderos, 2021.

Arana, Juan (Director), *El futuro de la identidad humana a debate*, Madrid, Tecnos, 2024.

Bostrom, Nick, *Superteligencia*, Zaragoza, Teell, 2016.

Debate sobre la inteligibilidad de la conciencia. Volumen monográfico de *Naturaleza y libertad*, núm. 7, Málaga, 2016.

Kurzweil, Ray, *Cómo crear una mente*, Berlin, Lola Books, 2019.