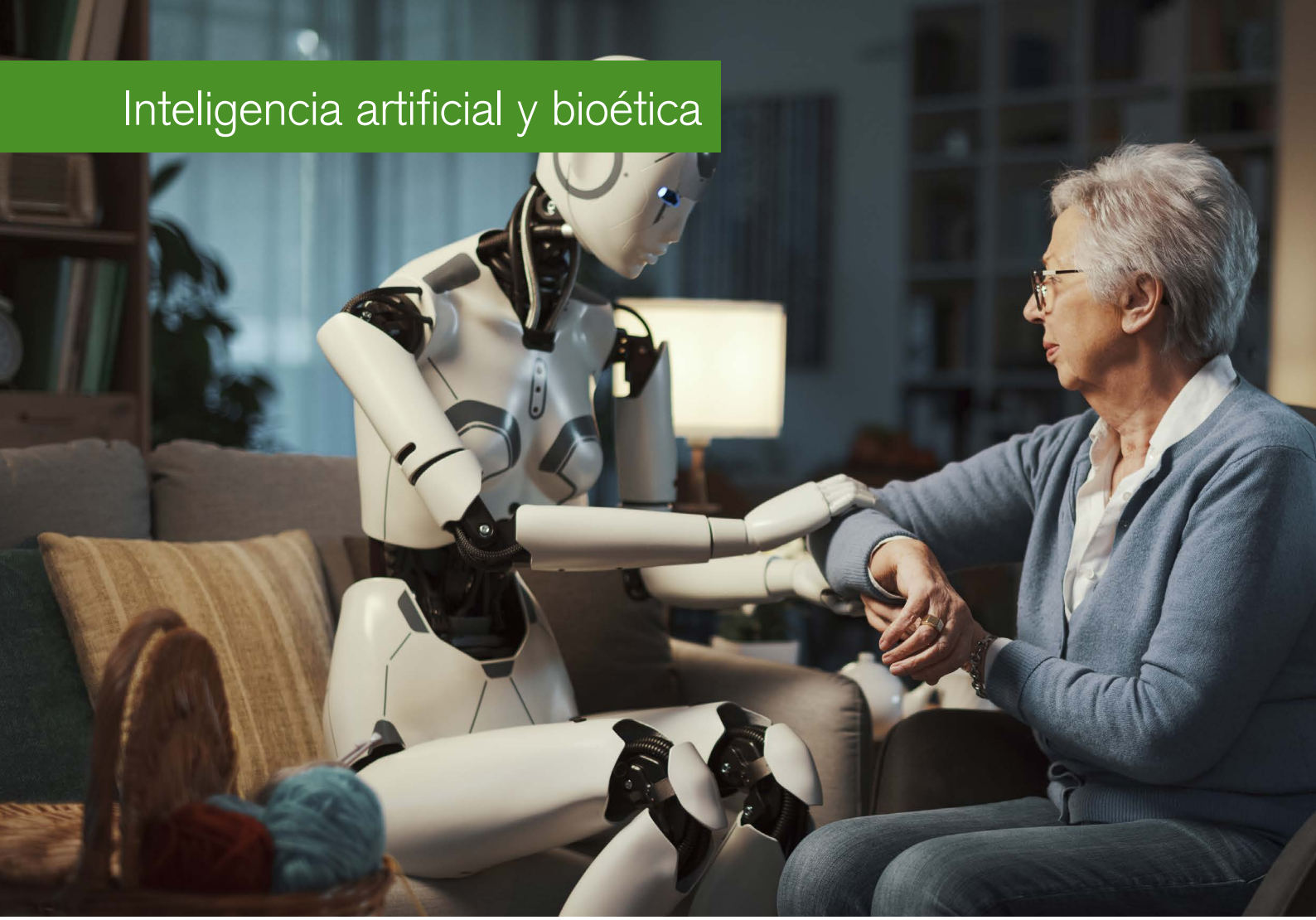

Inteligencia artificial emocional en el reverso del test de Turing. Al borde de la singularidad tecnológica son precisas cuatro nuevas leyes para la robótica

Emotional Artificial Intelligence on the Flip Side of the Turing Test. At the Edge of Technological Singularity, Four New Laws for Robotics are Needed

Autor

Luis Enrique Echarte Alonso

Unidad de Humanidades y Ética Médica
Facultad de Medicina
Universidad de Navarra
Email: lecharte@unav.es
https://orcid.org/0000-0002-8059-1992

Resumen

En este artículo defiendo que el creciente desarrollo de la inteligencia artificial específica (IAE) y, en particular, de la IAE para la interacción emocional (IAE-E), puede ir desdibujando las fronteras existentes entre los problemas éticos asociados a la IA débil (IAD), pequeños y presentes, ahora o a corto plazo, y los asociados a la IA fuerte (IAF), de gran calado y solo teóricos o potencialmente presentes a largo plazo. En mi argumentación muestro la importancia que adquiere el test de Turing, por su aproximación objetivo-subjetiva, al análisis y evaluación del segundo tipo de riesgos. Porque no se trata únicamente de cuánto la IAE-E es capaz de simular la inteligencia humana, sino también de cuánto más susceptibles somos los seres humanos de ser persuadidos por dichas simulaciones. La conclusión principal a la que llego es que para que el desarrollo tecnológico depare en progreso humano, necesitamos ser conscientes de los efectos que tiene para la IA-E, primero, el paroxismo social en las estrategias de autoengaño, segundo, la barbarie del especialismo y, tercero, el primitivismo tecnológico y docente.

Abstract

In this paper, I argue that the growing development of Specific Artificial Intelligence (SIA) and, in particular, SIA for emotional interaction (SIA-E), may blur the existing boundaries between the ethical problems associated with Weak AI (WIA), small and present, now or in the short term, and those associated with Strong AI (SIA), of great importance and only theoretical or potentially present in the long term. In my argument I show the importance of the Turing test, for its objective-subjective approach, to the analysis and evaluation of the second type of risks. For it is a question not only of how much IAE-E is able to simulate human intelligence, but also of how much more susceptible we humans are to being persuaded by such simulations. The main conclusion I draw is that for technological development to bring about human progress, we need to be aware that the social paroxysm of self-deception strategies, the barbarism of specialization, and technological and teaching primitivism have a catastrophic impact on AI-E.

Key words

Ética de la inteligencia artificial; IA emocional; test de Turing; inteligencia fuerte; inteligencia general.

Artificial intelligence ethics; emotional AI; Turing test; strong intelligence; general intelligence.

Fechas

Recibido: 13/02/2024. Aceptado: 23/05/2024

Así, sabios... no sois hombres de vuestro tiempo;
sois hombres del porvenir,
los precursores de la razón futura.

Holbach

1. Inteligencia artificial fuerte y general

John Searle establece por primera vez, en su artículo *Minds, brains, and programs*, la distinción entre inteligencia artificial débil y fuerte. Define la primera como una herramienta que la mente humana puede utilizar para explicar y probar hipótesis, mientras que la segunda no sería una herramienta para entender la realidad, sino que ella misma entendería, es decir, sería "realmente una mente, en el sentido de que se puede decir literalmente que las computadoras que reciben los programas adecuados comprenden y tienen otros estados cognitivos" (Searle, 1980). Esta distinción ha suscitado, desde hace ya más de tres décadas, diferentes debates sobre la naturaleza de lo mental. Uno de ellos es si existen diferencias objetivas, manifiestas o externas entre ambos tipos de inteligencias. La cuestión tiene importancia también a nivel práctico, porque ¿puede un programa de inteligencia artificial llegar a poseer (replicar) una inteligencia fuerte (IF)? Y no menos importante, ¿sería capaz de simular, al menos, la conducta de los agentes con IF? La segunda pregunta incluye las operaciones más elevadas como son las relacionadas con los razonamientos intuitivos, la creatividad o la toma de decisiones éticas. Tres hipótesis engloban las principales respuestas a la segunda pregunta.

¿Puede un programa de inteligencia artificial llegar a poseer (replicar) una inteligencia fuerte (IF)? Y no menos importante, ¿sería capaz de simular, al menos, la conducta de los agentes con IF?

En la primera, a la que denominaré hipótesis tecno-optimista, se afirma que la IA sí llegará a manifestar intencionalidad, mental y, por ende, replicará la inteligencia humana y las conductas asociadas a esta. En esa respuesta encontramos dos posturas diferenciadas. Por un lado, en posiciones como las de la teoría de la identidad del estado cerebral (también denominadas como *fisicalismo de tipo*, que es una de las versiones del *materialismo reductivo* o *monismo materialista*) se propone que los estados mentales son idénticos a los estados cerebrales. Esto implica, en último término, que la IAF no es sino un tipo complejo de inteligencia débil (ID). Por tanto, es cuestión de tiempo que los programadores alcancen dicha meta. La pregunta por la simulación se vuelve, así, irrelevante (Smart, 2004; Bostrom, 2016; Dennett, 2018). Con matices, en esta hipótesis podríamos también añadir a los defensores del conductismo, del neoconductismo y del funcionalismo (Moore, 2003; Fodor, 2000; Putnam, 1975).

La hipótesis *tecno-pesimista* asume la tesis contraria: las máquinas nunca podrán llegar a tener estados mentales ni reproducir las funciones superiores humanas. A esta van a suscribirse principalmente quienes mantienen las posiciones de corte dualista más extremas. Por ejemplo, la articulada por John Eccles, para quien los fenómenos no

pueden explicarse a partir de fenómenos materiales (neuronales, químicos…), ni mucho menos reproducirse con máquinas o programas que, por definición, son fruto de la pura objetivación del mundo físico (Eccles, 2015, pp. 167-183). En este marco, un computador no solo no podría replicar la IF, sino que tampoco podría reproducir funcionalmente, con mera ID, las operaciones intelectuales que dependen de la conciencia.

En tercer lugar, la *hipótesis tecno-escéptica* engloba un amplio grupo de corrientes que fluctúan entre el monismo y el dualismo psicofísico, con posiciones moderadas, y en las que se reconoce, más o menos explícitamente que, aunque un programa no pueda expresar IF, tal situación no significa necesariamente incapacidad para simular cualquier tipo de operación intelectual. No habría forma objetiva de demostrar, por tanto, que un programa con ID replica y no solo simula las operaciones más características de una IF. Esta hipótesis es coherente, entre otros, con el *fisicalismo no reduccionista* de Malcolm Jeeves (2009), con el *monismo anómalo* de Donald Davidson (2002), con la *teoría cuántica* de Roger Penrose y Stuart Hameroff (Penrose, 1994; Hameroff et al., 1996), y también con el naturalismo biológico que propone el propio Searle (1994).

Las controversias éticas asociadas a la IA pueden catalogarse en dos bloques. El primero incluye los riesgos inmediatos, relacionados con la tecnología que ya existe y que, por eso mismo, alcanzan hoy mayor atención: sobre la privacidad de los datos, los sesgos, el desplazamiento laboral, la trasparencia, la explicabilidad, etc.

En el marco de la tercera hipótesis adquiere relevancia ética una segunda distinción, posterior a la que hace Searle entre IF e ID: la que existe entre IA específica (IAE) e IA general (IAG). La primera hace referencia a programas capaces de realizar tareas específicas y limitadas, y aprender a ejecutar dichas tareas de manera más y más eficiente —pudiendo incluso expresar cierta capacidad de adaptación a escenarios cambiantes—. La IAG, en cambio, es multitarea y, lo que es más importante, integra destrezas específicas, es decir, aplica destrezas específicas en la resolución de tareas muy diferentes —uno de los rasgos más específicamente humanos—. Por ejemplo, un entrenador de baloncesto manifiesta inteligencia general cuando aprovecha su experiencia en la cancha para educar a su hijo. Nótese que en esta segunda distinción se pasa por alto el problema de la conciencia. La IAG no tiene por qué expresar una IF, ni la IAE tiene por qué carecer de IAD. Una IAG sin conciencia es lo que Daniel Dennett tipificaría como un sistema con "competencia sin comprensión" —aunque para luego negar dicho caso como imposibilidad ontológica (Dennet, 2018)—. Un modo gráfico de referirse a este tipo de programas es el de *zombis filosóficos*. David Chalmers crea esta expresión para referirse a seres sin consciencia, pero cuyo comportamiento es indistinguible del de un ser humano normal (Chalmers, 1996, pp. 95-99). La diferencia es que, para él, no es solo una hipótesis que falsar, sino una posibilidad, todavía lejana, pero a la que no hay que restar importancia, pues las repercusiones éticas son graves.

Las controversias éticas asociadas a la IA pueden catalogarse en dos bloques. El primero incluye los riesgos inmediatos, relacionados con la tecnología que ya existe y que, por eso mismo, alcanzan hoy mayor atención: sobre la privacidad de los datos, los sesgos, el desplazamiento laboral, la trasparencia, la explicabilidad, etc. El segundo bloque agrupa los peligros asociados a la IAG y, por tanto, tienen todavía un carácter

puramente especulativo. Uno de los más importantes es el que aborda la cuestión de la titularidad moral. ¿Se podría atribuir a una IA la responsabilidad de una decisión como atribuimos responsabilidad a las decisiones humanas? Otra segunda controversia, estrechamente relacionada con la anterior es sobre la protección legal que ampararía a dichos sistemas. ¿Merecerían los mismos derechos, y estarían obligadas a los mismos deberes que los seres humanos? La tercera gran controversia tiene que ver con la docilidad de dichas inteligencias. ¿Podrían llegar a suponer una amenaza? Aun reconociéndoles iguales derechos, ¿podrían los sistemas en IA llegar a creerse superiores a sus creadores y actuar en consecuencia? Y, por último, está la cuestión de la relación máquinas-seres humanos. ¿Llegaremos a guardar con los sistemas en IA un tratamiento interpersonal? En estos cuatro últimos debates, la distinción entre IA general e IA fuerte se torna fundamental, pues, como apunta Chalmers, la respuesta no puede ser la misma para sistemas que simulen la inteligencia humana que para sistemas que la replican.

Si Chalmers tiene razón, la elaboración de pruebas capaces de discriminar una IAF de IAD es crucial. ¿Es posible diseñar tales *pruebas*?

2. Superveniencia causal

De entre todas las aproximaciones a la hipótesis tecnoescéptica, quiero centrarme en la de Searle, porque es, a mi juicio, uno de los autores que mejor formulan el problema y en quien más fácilmente se detectan los errores de las conclusiones sobre la simulación perfecta.

Searle concede al dualismo la tesis de que la conciencia es el fenómeno mental central, donde la subjetividad es entendida de manera ontológica, existe como tal y no es reducible a otra cosa. Sin embargo, se alinea con el materialismo al rechazar la idea de que la conciencia no pertenezca, como el resto de entes, al mundo físico. En este último punto da la razón al materialismo. "El universo consiste enteramente en partículas físicas en campos de fuerza (o lo que sea que la teoría física verdadera diga que son los componentes básicos del universo)" (Searle, 2007, p. 333). Acepta del materialismo la idea de que los procesos neurales causen la conciencia, esto es, la reducción causal, y al mismo tiempo, niega la reducción ontológica. Es lo que denomina "superveniencia causal" de la mente sobre el cerebro. A su vez, esta condición es compatible, según él, con reconocer que la conciencia ejerce eficacia causal sobre el cerebro, tesis que aleja a Searle del epifenomenalismo.

Searle reconoce que esa eficacia causal de la conciencia (de arriba-abajo) puede ser también reproducida por otros medios, en concreto, por procesos materiales como los que caracterizan a un computador

Searle reconoce que esa eficacia causal de la conciencia (de arriba-abajo) puede ser también reproducida por otros medios, en concreto, por procesos materiales como los que caracterizan a un computador. Este nuevo tipo de eficacia causal (horizontal y *a*

posteriori) competiría con la mental (vertical y *a priori*), pero solo a nivel conductual y no en la producción de consciencia (Searle, 1994, pp. 70-71). Por tanto, la IA dejaría de lado el rasgo esencial de la inteligencia, de los agentes con mente, al menos hasta que la neurofisiología del futuro descubra cómo el cerebro fabrica la conciencia. Mientras, solo estaremos rodeándonos de mejores o peores zombis filosóficos. En el naturalismo biológico de Searle, es concebible el diseño de simulaciones perfectas e inconcebible el diseño de pruebas de discriminación infalibles entre la IAF y la IAD.

Searle se aleja del *fisicalismo* de tipo al afirmar que el cerebro no causa la conciencia. El cerebro no es al viento, lo que la conciencia al desplazamiento de las nubes

En el marco que nos presenta el naturalismo biológico de Searle, es concebible el diseño de simulaciones perfectas e inconcebible el diseño de pruebas de discriminación infalibles entre la IAF y la IAD. Disponemos de un paradigma para explicar las relaciones de causa y efecto de tipo horizontal —podemos traducirlas en términos de ceros y unos— pero no para explicar las de tipo vertical, que son las que existen entre la IF y la ID. Carecemos de un paradigma psicofísico ni tenemos, por tanto, la más mínima idea de cómo entrenar a un programa con este tipo de fenómenos causales.

Considerar este supuesto periodo transicional hacia el descubrimiento del paradigma bastaría para que nos tomáramos en serio los riesgos asociados a la IAF arriba mencionados. Porque, como argumentaré, no son peligros de futuros remotos. Pero antes de entrar a valorar esta última idea, quiero detenerme a matizar la tesis de la imposibilidad discriminativa. Creo que sí es posible concebir una prueba para distinguir entre replicantes y simuladores, aunque, aviso, dicha prueba no es viable en términos prácticos.

Searle se aleja del *fisicalismo de tipo* al afirmar que el cerebro no causa la conciencia. El cerebro no es al viento, lo que la conciencia al desplazamiento de las nubes. La superveniencia causal funciona de manera diferente. Es un tipo de emergentismo donde la conciencia surge por la interacción entre millones de neuronas. La conciencia no se puede encontrar en ninguna de ellas tomada por separado —como tampoco el fenómeno de liquidez puede atribuirse a una única molécula de agua—. Sin embargo, la analogía que establece Searle entre liquidez y consciencia es conceptualmente insuficiente. Sí es posible deducir las propiedades de la liquidez a partir del análisis de una única molécula de agua pues, dentro del paradigma materialista (que maneja causas horizontales) cabe imaginar escenarios hipotéticos donde proponer nuevas tesis, aunque estas sean solo comprobables cuando la tecnología alcanza el progreso tecnológico necesario. Es el caso, por ejemplo, de la curvatura del espacio-tiempo, que Albert Einstein fue capaz de llegar en 1915 y que solo comenzó a comprobarse cuatro años después. Del mismo modo, un físico puede concebir y prever la interacción entre dos moléculas iguales o diferentes nunca antes agrupadas. No ocurre así con la conciencia y los modelos neuronales. Los fenómenos pertenecen a órdenes tan diferentes que se hace inverosímil establecer entre ellos, sin más, una relación causal, con todos los matices y excepciones que quiera darse a dicho nuevo tipo de causa.

El mismo Searle reconoce la debilidad de una propuesta que está asentada más en lo que las neurociencias están por descubrir que en lo descubierto. Con todo, se defiende diciendo que “si tengo que elegir entre los hechos tal como los conocemos

(la conciencia existe, es causada por procesos neuronales, existe en el cerebro y tiene funciones causales en la vida del organismo) y varias teorías filosóficas, tomaré los hechos en cualquier momento" (Searle, 2007, p. 334). Pero, ¿cuáles son los hechos y cuáles las teorías filosóficas a las que se refiere? Empecemos por lo segundo.

Searle enfrenta las propuestas que toman como punto de partida ya la experiencia fenoménica, ya la actividad neuronal —dualismos y monismos—. Sobre su naturalismo biológico afirmará: "No es una visión que se derive naturalmente de la reflexión sobre las propias experiencias o del estudio de las operaciones cerebrales. Una vez que superemos los errores de la tradición, creo que los hechos encajarán naturalmente en su lugar" (Searle, 2007, p. 334). Es decir, propone que deben ser dos los puntos de partida, y no solo uno, aunque no termina de justificar por qué, ni llega a conclusiones explicativas sólidas sobre la relación mente-cerebro. Sus trabajos son valiosos para constatar el gran misterio al que nos enfrentamos, en efecto, pero no tanto como para convencernos de que los senderos de la neurofisiología bastarán para resolverlo. Pero, además, contamos con fuertes argumentos para sostener la tesis contraria: el punto de partida es uno y no dos. Explicaré la razón de ser de dichos argumentos —que yo mismo defiendo— respondiendo, a continuación, a la primera cuestión, esto es, a la pregunta sobre los hechos. Y anticipo la tesis central: la idea de hechos que maneja Searle es equívoca. Veremos también que esta discusión tiene importantes repercusiones sobre la posibilidad de concebir una prueba discriminatoria entre la IF y la ID. Adelanto también mi conclusión: es posible, aunque solo desde el punto de vista teórico y no práctico.

Siguiendo una tradición filosófica que se remonta a Aristóteles, Searle define por intencionalidad aquello que sirve para señalar las diversas formas porque la mente puede dirigirse hacia, o tratar de objetos y estados de cosas en el mundo

3. Conocimiento y subjetividad

La conciencia es, para Searle, lo esencial a la mente, y a su vez, la intencionalidad es lo esencial a la conciencia. A la luz de los nuevos enfoques que surgen en el siglo XX en filosofía del lenguaje, pero siguiendo una tradición filosófica que se remonta a Aristóteles, Searle define por intencionalidad aquello que sirve para señalar las diversas formas porque la mente puede dirigirse hacia, o tratar de objetos y estados de cosas en el mundo (Searle, 1983, pp. 4-10).

Es aquí, a nivel de los supuestos ontológicos intencionales, en las causas primeras, donde Searle busca establecer la conciliación entre lo mental y lo material, aunque, como ya se ha dicho, concediendo primacía causal al cerebro, lo que entra en cierta contradicción con su afirmación sobre los pasados errores de la tradición. Pero la propuesta de Searle revela sus insuficiencias no por caer en esta ambigüedad, sino principalmente por acabar resolviéndola en lo material, y con ello, invirtiendo el orden jerárquico que guardan lo consciente y lo neuronal, tanto a nivel epistémico como ontológico.

Antes que Searle y, en general, los filósofos del lenguaje del siglo XX, el rasgo intencional de la mente fue trabajado por autores como Franz Brentano, Edmund Husserl y Martin Heidegger para justificar la mejor metodología para acceder a lo mental y, por extensión, a la realidad. Y es que, si la realidad es alcanzable, lo es porque el ser humano posee una mente que le permite conocerla. Y por el ser de la intencionalidad, todos estos autores refutarían de manera similar el naturalismo biológico: el fenómeno prima sobre los hechos —esos hechos que menciona Searle como pilares fundamentales de su propuesta—, y donde la conciencia es causada de manera incontestable por las neuronas. La mente es anterior a la materia y si hubiera que atribuir poder causal en esta dualidad, esta debiera ser la primera sobre la segunda. El argumento central de esta crítica va como sigue.

En otras palabras, para la fenomenología, el sujeto es, en sus más hondos fundamentos, manifestación relacional. Es, antes que otra cosa, subjetividad. Y la subjetividad marca, como reconoce el propio Searle, la esencia del conocer

Prestar atención al rasgo intencional de la mente es entender que la mente, en el primer momento del conocer, se identifica con el objeto conocido. Este primer momento es el *fenómeno* que, como define Heidegger, "significa 'mostrarse a sí mismo'" (Heidegger, 1962, sec. 7). El papel del que busca la realidad debe ser siempre el de fenomenólogo, esto es, el de quien describe y trata de entender este *mostrarse* para, a continuación, separar lo real de las apariencias. Los fenómenos son, por tanto, los verdaderos *hechos brutos* que inician la escalera del conocimiento, y no la teoría neuronal, *hecho procesado* gracias a la ciencia experimental.

Ahondemos en los rasgos esenciales al fenómeno, pero ahora, no en términos de ente sino de acción. La apertura fenoménica, tener una mente, señala la capacidad que goza un agente para relacionar varios entes entre sí, siendo el agente intencional, en sí mismo, el aparecer de la relación, y a la vez, el testigo de dicho aparecer. En este tríptico de bordes difusos, el fenómeno acontece como un ente primariamente subjetivo, esto es, implica *el acceso privilegiado de la primera persona*, el de un sujeto hacia los objetos de la relación. En otras palabras, para la fenomenología, el sujeto es, en sus más hondos fundamentos, manifestación relacional. Es, antes que otra cosa, subjetividad. Y la subjetividad marca, como reconoce el propio Searle, la esencia del conocer. Siendo esto así, sería incorrecto afirmar que la subjetividad y la objetivad son dimensiones de la inteligencia.

Conocer es posibilitar la copresencia entre dos entes, que se abren el uno para el otro en perfecta simultaneidad de espacio y tiempo. Por esta razón, algunos creen que, mejor que decir que la inteligencia es primariamente subjetiva, es afirmar que la inteligencia es subjetividad, es decir, el principal acto con el que los entes trascienden su mismidad. La subjetividad y la objetivad no serían, entonces, dimensiones de la inteligencia. La inteligencia sería subjetividad, y la objetividad se definiría como la expresión externa de lo que, inicialmente, está oculto al resto de mentes. Con la objetividad, los polos se aíslan, los fenómenos se tematizan, nacen conceptos y símbolos y se crean espacios intersubjetivos. Pero, con todo, y al contrario de lo que luego muchos llegarán a afirmar, entre otros Searle, sí que sería posible hablar, de manera genuina, de conocimiento prelingüístico y presocial, pues el conocimiento empezaría en la misma percepción

sensorial donde, como el propio término implica, el cognoscente recibe los *inputs* físicos como algo concreto sentido o también como un sentimiento más general por estar ya integrado en la entera subjetividad —el sentimiento de ser o de lo que ocurre (Bermúdez, 2003, p. 5)—. En otras palabras, en estos primeros pasos del conocer, el *sentimiento de ser* y el *sentimiento de lo que ocurre* serían uno y lo mismo.

En la rica nomenclatura que la fenomenología utiliza para describir la afectividad humana, la esfera fenoménica se antoja obvia y necesaria al sentido común. En contraste, y por la misma intuición, tendemos a rechazar la *percepción zombi* o el *sentimiento zombi* como verdadera percepción o sentimiento. O, por lo menos, estamos dispuestos a tacharlos de espejismo con más fuerza que con el *razonamiento deductivo zombie*.

Las máquinas nos parecen cada vez más como seres humanos porque, en parte, cada vez más nos pensamos en términos de máquina

La posición de Searle es, en este sentido, contraintuitiva. Y no es solo un tipo de falacia en la que caigan filósofos. Un buen número de neurocientíficos, Antonio Damasio entre los más conocidos, proponen que toda emoción pueda reducirse a mera descripción conductual de respuesta sistémica al entorno (Damasio, 1994, pp. 164-167). No es un error trivial pues influye decisivamente en la manera en la que entendemos al ser humano y a las máquinas que diseñamos: por un lado, la reducción objetivista supone otra vuelta de tuerca en el proceso de tecnificación de lo humano —racional, afectiva y volitiva— y, en segundo lugar, potencia los espejismos de antropomorfización de la IA. Las máquinas nos parecen cada vez más como seres humanos porque, en parte, cada vez más nos pensamos en términos de máquina.

4. Tirar la escalera de la IA

La descripción fenomenológica de la intencionalidad ha ido siendo eclipsada a lo largo del siglo XX por otro enfoque, el de la filosofía de la mente, de carácter eminentemente analítico, y donde es probablemente Gilbert Ryle el que da el pistoletazo de salida con su trabajo *El concepto de lo mental* de 1949. En esta obra, su autor pretende romper con la posición cartesiana, que sustantiva la inteligencia, para volver a enfoques clásicos, como el de Aristóteles, en el que la inteligencia no es definida como *res* (*cogitans*), sino como *actio* (*cogitandi*). Si Descartes propone un dualismo de mundos antagónicos, siendo uno de ellos el de la mente pensante, Ryle propone solo uno, el material, capaz de organizarse de tal forma que es capaz de producir agentes con conductas inteligentes (Ryle, 1978). Ryle hace una lectura materialista de Aristóteles, a mi juicio errada, pero de esto diré más después. Ahora volvamos con Searle. El filósofo de California no defiende un conductismo tan extremo con el de Ryle, pero sí afirmar que la mente es producida por el cerebro, y lo hace de un modo análogo a cómo la función hepática es causada por el hígado. El argumento de Ryle se sostiene a base de negar el carácter intencional de la mente, mientras que Searle salva dicho rasgo, pero con el alto coste de defender una inconcebible relación causal. Y si atendemos a la aproximación fenomenológica,

este es un error más grave incluso que el del dualismo de sustancias en el que cae Descartes. Al menos este último acierta al señalar cuán incuestionablemente primero es el pensamiento sobre la materia: la única verdad indubitable para todo sujeto racional es que este piensa, idea a la que se llega mediante un mero ejercicio de introspección. Y para que este postulado cartesiano no nos conduzca al dualismo radical basta con que se complete con la coda fenomenológica de que conocimiento y sujeto se identifican a la par que se manifiestan.

Desde Ryle hasta Searle, y todavía hoy, son mayoría los que pretenden llegar a la subjetividad a partir de la objetividad. Pero este es un proyecto en el que se invierten la escalera metodológica. Ludwig Wittgenstein tipifica esta vía como un "tirar la escalera", el error de perder el sentido cronológico que condiciona cualquier método —subir un tejado para luego negar la escalera con la que se ha subido (Wittgenstein, 1973, aforismo 6,54, p. 203)—. El error de tirar la escalera tiene especial significancia en el ámbito de la computación y, en particular, en el diseño de test discriminatorios.

Ludwig Wittgenstein tipifica esta vía como un "tirar la escalera", el error de perder el sentido cronológico que condiciona cualquier método —subir un tejado para luego negar la escalera con la que se ha subido—

Entre las variadas operaciones de la inteligencia, una de las más importantes es la abstracción, que implica distinción y descomposición de la realidad en sus elementos, para dar con conceptos y definiciones. Así, puede decirse del concepto "información" que existe en el plano del *logos* pero no en el de la *physis*, el de la realidad, donde nunca se da separada del acto de informar. Sin embargo, este último enunciado debe ser matizado en el contexto tecnológico. Los seres humanos transformamos la información (los conceptos y definiciones que abstraemos del fenómeno mental) en símbolos (realidades físicas). Es posible incluso diseñar ingenios que operen con dichos símbolos (programas) para reproducir los cómputos mentales —cómputos, que no mente—. De este modo la información ha pasado a ocupar un lugar en el plano de la *physis*, pero solo como realidad artificial y, como tal, secundaria y reducida al acto de informar —es lo más alto del tejado—.

¿Qué dejan atrás estos ingenios? Desde hace algunos años los filósofos vienen utilizando el término *qualia* para referirse a las cualidades subjetivas de las experiencias individuales (Kauffman y Roli, 2023). Sin embargo, esta definición subvierte la ontología de los actos de inteligencia, entre otras razones, porque obvia la función esencial que, a su vez, explica la jerarquía y tiempos de dichos actos —su ontogenia—. Comparar las cualidades subjetivas y objetivas de la inteligencia resulta tan confuso como, en un horno, comparar su capacidad para generar calor con su capacidad para medir el tiempo —pensemos en estos hornos con reloj de cocina incorporados—. La primera es esencial y primaria, hace al horno lo que es, la segunda accidental y secundaria. La IA es una realidad que pertenece al ámbito de la información y, por tanto, es una realidad accidental al acto de informar.

La dificultad está en que, a diferencia de los relojes de cocina, la IA puede manifestarse como una realidad simuladora de inteligencia (sobre todo si, como veremos, simula emociones). Nadie en su sano juicio piensa que un reloj de cocina tiene algo que ver

con el fenómeno de generar calor, por muy integrado que esté dicho reloj en el horno. El error sale a relucir más fácilmente si atendemos al factor cronológico. Es difícil de creer que un medidor de tiempo pueda evolucionar en complejidad hasta convertirse en un generador de calor, pero es inconcebible pensar que la comprensión es posterior y derivada de la información —como creer que primero se inventaron las chisteras y luego las cabezas, primero la escritura y luego los lectores—. Y aunque es cierto que a los medios somos capaces de encontrarles nuevos usos, nuevos fines, de lo que estamos hablando aquí es del inicio de esta larga cadena causal. Allí, en el principio, solo puede habitar el pensamiento —la comprensión—, que es siempre anterior a cualquiera de sus productos —entre otros, los métodos de objetivación—. De modo análogo, la simulación no puede ser anterior a lo simulado, ni el retrato de Enrique VIII puede evolucionar a rey inglés.

En resumen, los hechos a los que apela Searle para defender la superveniencia causal apuntan justo a lo contrario: resulta más intuitivo y fácil creer que es la consciencia la que genera la materia que lo contrario

En resumen, los hechos a los que apela Searle para defender la superveniencia causal apuntan justo a lo contrario: resulta más intuitivo y fácil creer que es la consciencia la que genera la materia que lo contrario. Esto no significa que sea cierto, ni tampoco que esta segunda vía tenga una fácil argumentación —podemos decir, quizá, que solo una es más fácil que la contraria—. Sin embargo, esta vía de enfoque sobre el fenómeno intencional nos abre una inusitada vía para la discriminación entre IAF e IFD. En el siguiente epígrafe explicaré dicho asunto que, además, relacionaré con el segundo gran problema, arriba mencionado, sobre cómo Ryle y sus discípulos malinterpretan la noción aristotélica de acción. Porque, vamos a verlo, Aristóteles entendería mejor que nadie el error del naturalismo biológico, que subvierte la ontología de los actos de inteligencia al obviar la función esencial que, a su vez, explica la jerarquía y tiempos de dichos actos —su ontogenia—. E insisto, este error se revela aún más evidente en el debate de la IA. Creer en programas que puedan desarrollar conciencia es como imaginar que de las chisteras puedan surgir cabezas.

5. Novedad teleológica

¿La inversión del planteamiento del naturalismo biológico cambia la posición tecnopesimista sobre las pruebas de discriminación? Los modos de esta inversión son dos: que la mente cause la materia o que la mente no cause la materia, pero tenga un lugar hegemónico en la interacción entre ambas. No obstante, la respuesta parece similar para ambos modos: no es posible la discriminación en términos prácticos, pero sí teóricos.

Toda prueba es, por definición, objetiva, pero lo que estamos tratando de detectar en la IAF es la presencia de subjetividad, es decir, un elemento que queda fuera del ámbito de posibilidad de toda prueba. Por tanto, la pregunta crucial es si la inteligencia fuerte se manifiesta objetivamente de manera diferente a la mera inteligencia general. De nuevo los tiempos son aquí importantes pues una IA entrenada de manera correcta y suficientemente podría ser capaz de reproducir pautas de conducta idénticas. Por esa

razón, el debate suele acabar orientándose hacia la capacidad de estas máquinas para generar novedad: para aprender originales caminos de resolución de tareas (medios) o para encaminarse hacia nuevos destinos (fines). Si una IA expresara objetivamente dicha apertura hacia la novedad sería más factible asignarle una IF —aunque nunca con certeza absoluta, pues, si aceptamos la aproximación fenomenológica, tal certeza solo podría tenerla la máquina para consigo—.

La presencia de conducta novedosa no lo resuelve todo porque ya los programas actuales son capaces de abrirse camino ante lo desconocido, aunque no de todos los modos imaginables (James y Moneta, 2020). Por ejemplo, se desenvuelven muy bien para encontrar nuevas correlaciones y patrones (especialmente con el desarrollo del *big data*), y por supuesto, para traducir estos estos hallazgos en nuevos modos de resolución de tareas. También manifiestan gran potencia de novedad probabilística, especialmente de tipo bayesiano. Más problemas encuentran los programadores para lograr novedad basada en la aleatoriedad numérica. Pero, sin duda, la novedad que parece inaccesible a la ID (inteligencia débil) es la que hunde sus raíces en la teleología, no como la presentan Ryle y Searle, que es siempre funcional y derivada, sino como la de Aristóteles, primera y constitutiva. Ahora bien, no hay que confundir finalidad y meta: la segunda es la objetivación de la primera. Los programas operan con metas computarizadas, es decir, la información no tiene, por sí misma, fuerza operativa. En contraste, las mentes experimentan la finalidad como elemento intrínsecamente operativo de la conducta, pues la simple aprehensión del fin atrae al agente a su consumación. Esta es una tesis central en la cosmovisión aristotélica y, en particular, para su *Física* (Libro II) que es donde se expone de manera más clara, y dentro de la distinción tetracausal, su tesis sobre la irreductibilidad de la causa final.

Lo particular de causa final es que los agentes en los que opera internamente (los seres naturales, no los artificiales) experimentan esta fuerza en términos normativos —algo es bueno— y, antes que eso, en términos estéticos —algo es bello, merece ser afirmado, celebrado, cuidado, etc.—

Cuatro son los modos en los que es posible explicar un fenómeno: la causa formal (podemos explicar, con ella, que la rueda de un carromato se mueve por su forma, redonda), la causa material (la rueda se mueve porque es ligera, de madera), la causa eficiente (la rueda se mueve porque el carromato es empujado por el viento) y la causa final (la rueda se mueve por la acción de agitar las riendas del conductor del carro). Lo particular de causa final es que los agentes en los que opera internamente (los seres naturales, no los artificiales) experimentan esta fuerza en términos normativos —algo es bueno— y, antes que eso, en términos estéticos —algo es bello, merece ser afirmado, celebrado, cuidado, etc. (Taylor, 2001, pp. 377-382)—. Son descripciones que apelan al conocimiento en primera persona, que impelen al sujeto (en su individualidad), y que, por ello mismo, pierden su esencia definitoria y su poder performativo con la abstracción generalizadora. Paradójicamente, y como indica Bergson (autor muy cercano a la fenomenología), al carácter *inobjetivo* de esta cuarta causa hay que sumar otros dos rasgos. "Como remolinos de polvo levantados por el viento que pasa, los vivientes giran sobre sí mismos, suspendidos por el gran soplo de la vida" (Bergson, 2007, p. 141). El primero concierne a la sutilidad con la que opera dicho *élan vital*, y el

segundo a la conexión que guardan todos los seres naturales entre sí, que colaboran, de manera heterogénea en una y la misma dirección, hasta el punto que pareciera que en vez de muchos solo hubiera una única fuerza.

La causa final es la principal característica que distingue a los seres naturales de los seres artificiales y, en general, del conjunto de fenómenos azarosos o violentos. Solo los movimientos naturales son susceptibles de normatividad. Podríamos decir, alegóricamente, que los movimientos de la naturaleza son emocionales de suyo por su causa final. Es decir, atendiendo a la etimología, son movidos, retirados (*motio*) de un lugar a otro (prefijo *e/-ex*) por una intencionalidad que tiene origen dentro del ser. Pero la causa final no solo traspasa la dimensión ontológica, sino también la epistemológica, para, además, reunirlas. En el acto de conocer el cognoscente sabe que el objeto mental no se identifica necesariamente con el objeto representado. El cognoscente es consciente de los espejismos de la mente. Pero no se conforma. Cada acto intelectual, por básico que sea, va acompañado de una primigenia afirmación de adecuación de la mente con la realidad. Así, en las evidencias, lo visto (del verbo *videre*) se toman como verdaderas —de nuevo, el prefijo e/ex saca al cognoscente del mero asentimiento de lo que podría coincidir o no, para abrirle a un primer juicio veritativo, del que seguirán otros muchos—. Lo conocido habrá de ir siendo cierto, como lo imperfecto tendrá que dejar de serlo. Emociones y evidencias parten del mismo principio común y primero, que es anterior a ellas mismas.

La causa final es la principal característica que distingue a los seres naturales de los seres artificiales y, en general, del conjunto de fenómenos azarosos o violentos. Solo los movimientos naturales son susceptibles de normatividad

La meta es el signo objetivo de la finalidad y, por ende, de los seres con IF. Pero, ¿cómo distinguir la novedad que genera respecto de la novedad asociada a los efectos de las otras tres causas que también operan en la naturaleza? En términos prácticos, su detección es muy problemática dentro de los estándares científicos (Echarte, 2023a). Necesitaríamos viajar en el tiempo hacia el final de la vida de un individuo y comparar su deriva conductual hacia la novedad con respecto a la novedad originada por su *gemelo digital* (Echarte, 2023b). Y tan imposible como viajar en el tiempo sería lograr, para evitar sesgos, que ambas conductas se desarrollasen paralelamente, en el mismo espacio y tiempo.

Son solo entelequias. En términos prácticos, no es posible identificar zombis computacionales, por lo que, en determinados contextos personales y culturales, el riesgo que suponen para los seres humanos resulta incuestionable. Analizaré a continuación la relación entre dichos contextos y los riesgos asociados.

6. Tecnología persuasiva

Determinado el ámbito y alcance de la simulación en IA, es fácil entender que uno de los más peligrosos riesgos, no es necesariamente la IA que simule a la perfección la posesión de una determinada capacidad (IE), por sensible que sea (por ejemplo, la seguridad y uso de armamento nuclear, o la regulación de flujos migratorios); ni siquiera la que simule

la posesión de una IG y, por tanto, capacidades de autoaprendizaje y de apertura a la novedad extraordinarias (útiles especialmente en investigación y en arte); sino la IA capaz de convencernos, mediante el uso de ficciones, de que estamos realmente frente a una auténtica IF. Por eso esta tecnología persuasiva debe ser regulada y, para ello, el test de Turing abre importantes horizontes de posibilidad para la discriminación no entre IAF e IAD, sino entre IA peligrosa e inocua. Voy a explicar por qué.

Sin entrar en cuestiones ontológicas, Alan Turing vino a proponer que un programa alcanzaría a la inteligencia humana en el momento en el que un individuo fuera incapaz de distinguir entre el comportamiento del programa y el de otro ser humano (Turing, 1950). Esta prueba fue pronto denostada por su carácter subjetivo, ya que no refleja únicamente cuán sofisticado puede ser un sistema de autoaprendizaje, sino también cuán ingenuo el observador que lo evalúe. Por eso, el reverso del test de Turing es como denomino a esta aproximación dual sobre los riesgos de la IA. Porque, para lo que aquí nos ocupa, el riesgo ético, ambas dimensiones del fenómeno de persuasión son igualmente relevantes.

Alan Turing vino a proponer que un programa alcanzaría a la inteligencia humana en el momento en el que un individuo fuera incapaz de distinguir entre el comportamiento del programa y el de otro ser humano

¿Cuándo una simulación es lo suficientemente convincente? Hay dos vías para responder a esta pregunta, una positiva y otra negativa. La primera tiene que ver con la natural tendencia humana a establecer atribuciones por analogía. Cuanto más se parezca el exterior de un programa a la conducta humana (lo que incluye las capacidades para resolver problemas y ampliar los horizontes de posibilidad) más tentador será la idea de que su interior sea igual al nuestro: a más cercana la IE a la IG, mayor el espejismo de la IF. Y este fenómeno psicológico se vería potenciado cualitativamente si la IG, o incluso una IE de amplio espectro, superara las capacidades humanas.

Imaginemos el escenario futurista propuesto en 1965 por el matemático Irving John Good, donde el ser humano sería capaz de crear *máquinas ultrainteligentes*. En tanto que dichos programas habrían sido creados por seres humanos, Good presupone en ellas la misma capacidad para seguir creando máquinas con aún mayor inteligencia que los primeros modelos de silicio. "Entonces se produciría sin duda una 'explosión de inteligencia' y la inteligencia del hombre quedaría muy atrás. Así, la primera máquina ultrainteligente es el último invento que el hombre necesitará hacer, siempre que la máquina sea lo suficientemente dócil como para decirnos cómo mantenerla bajo control" (Good, 1966). Dos evaluaciones éticas paralelas pueden extraerse a partir de este pasaje: el centrado en la utopía de un futuro en el que los seres humanos manejan máquinas dóciles; y el centrado en su contrario, la distopía de una sociedad humana enfrentada a máquinas violentas. Para el argumento que estoy hilando, me interesa analizar el escenario utópico, aunque primero ofreceré una clave sobre el vínculo que guardan ambos.

El futuro de las máquinas ultrainteligentes nos enfrenta a uno de los tres escenarios de *singularidad tecnológica* hoy recogidos en los debates sobre potencial tecnología disruptiva (Gayozzo, 2021). Estos *superzombis filosóficos*, sin comprensión pero con altísimas competencias cognitivas, generarían imágenes de gran fuerza persuasiva

sobre su supuesta intimidad, incluso su sensibilidad moral, que pudiera llegar a ser incomprensiblemente mayor que la humana, y ante la que los humanos no les quedaría más remedio que plegarse: obedecer dócilmente o con violencia. Aquí encontramos la inquietante conexión con el segundo escenario, el distópico, de Good. Pero, como digo, no entraré a desarrollar esta cuestión.

Otro factor determinante en el desarrollo de programas persuasivos que la prueba de Turing pone de manifiesto tiene que ver con la inteligencia artificial emocional. Son sistemas automatizados capaces de detectar y procesar información sobre las emociones individuales y los estados mentales a partir de comunicación verbal y no verbal, con el objetivo de obtener conclusiones respecto de la personalidad, la credibilidad y las preferencias del sujeto sobre el que se aplica (Crawford, 2021, pp. 151-172). A diferencia de las máquinas ultrainteligentes de Good, esta tecnología ya existe y podemos encontrarla en automóviles, herramientas docentes, juguetes, robots domésticos, en aplicaciones de atención al cliente, también en el diseño de estudios de mercado, y un largo etcétera. Todas estas aplicaciones se asientan en una IA no de tipo general sino específica (IAE), orientada al procesamiento de inputs emocionales (IAE-E) y, sin embargo, generan ficciones antropomórficas muy poderosas.

Ya contamos con la tecnología para crear programas capaces de evocar en el usuario vínculos de apego similares al que guardamos entre los seres humanos, aunque, por supuesto, sin espacio intersubjetivo real. La máquina no siente, no es consciente de las cosas que dice, del amor que expresa

Una segunda diferencia con respecto a las máquinas ultrainteligentes es que el efecto persuasivo no es indirecto, es decir, no genera unas inferencias analógicas entre las operaciones reales y la supuesta subjetividad. La persuasión que operan las IAE-E es inmediata: el humano identifica en el programa elementos característicamente subjetivos —dolor, alegría, enfado— e intersubjetivos —"vas a tener un gran día", "te escucho" o "te quiero"—. Ya contamos con la tecnología para crear programas capaces de evocar en el usuario vínculos de apego similares al que guardamos entre los seres humanos, aunque, por supuesto, sin espacio intersubjetivo real. La máquina no *siente*, no es consciente de las cosas que dice, del amor que expresa. Por eso, a diferencia de las máquinas ultrainteligentes, en los que algunos podrían sospechar la emergencia de consciencia, en la IAE-E está fuera de toda duda su condición de simulaciones. Y aquí la paradoja. De nuevo, son programas más simples que los que pudieran alcanzar la IAG y, sin embargo, más eficientes para superar el test de Turing.

Desgraciadamente, el potencial peligro de estos programas pasa todavía desapercibido. Así se refleja, por ejemplo, en el nuevo proyecto de reglamento de la Unión Europea para la regulación de la IA (CE, 2023), donde el uso de dicha tecnología para el reconocimiento y procesamiento de emociones se considera inaceptable excepto si el usuario sobre el que se aplica concede su pleno consentimiento. Y consentimiento es, si tengo razón, el verdadero talón de Aquiles del proyecto. Veamos por qué.

7. IAE-E para la era del vacío

Uno de los factores más importantes en el reverso del test de Turing (de nuevo, sobre la capacidad no de persuadir, sino de dejarse persuadir), tiene que ver con el mal de soledad que sufren las sociedades occidentales.

El Observatorio Estatal de la Soledad no Deseada (OESD), de la Fundación ONCE, recoge que el 13,4% de la sociedad española sufre de este problema, que además afecta cada vez más a los jóvenes. La situación le cuesta al estado unos 14 141 millones de euros (OESD, 2021 y 2023), factura principalmente relacionada con el coste que tiene la soledad para la salud (Jeste, Lee y Cacioppo, 2020). No es un diagnóstico local pues parece haber cobrado verdadera forma de epidemia en EE. UU. En un informe publicado el año pasado por el Department of Health and Human Services, se aconseja tomar medidas a escala nacional: "Como se ha ido acumulando durante décadas, la epidemia de soledad y aislamiento ha alimentado otros problemas que nos están matando y amenazan con destrozar a nuestro país. Dados estos costos extraordinarios, reconstruir la conexión social debe ser una máxima prioridad de salud pública para nuestra nación" (OSG, 2023). A esta recomendación se han adelantado ya países como Reino Unido y Japón creando un Ministerio de la soledad en 2018 y en 2021, respectivamente.

El anciano que hoy vive acompañado de un robot cuidador sigue estando solo pero, como ya no se siente solo, permanece indiferente a su situación real. Ya no importa. Y eso es lo peor

En dicho marco se explican proyectos como "Caresses", financiado desde 2017 a 2020 por la Comisión Europea y Japón, para la creación de "robots asistenciales con competencia cultural". Con estos robots cuidadores (*care robots*), que ya se comercializan en centros de la tercera edad (son relativamente baratos), se busca fomentar la independencia y autonomía de los ancianos y, sobre todo, paliar los síntomas de soledad (CORDIS, 2023). Los conflictos éticos asociados a esta tecnología son numerosos, pero, los más graves, los relacionados con la titularidad moral que los usuarios otorgan a la máquina, que va parejo a los vínculos afectivos creados. Y es lógico pues la soledad desaparece en la medida que el anciano cree que hay un *otro* que le acompaña y cuida. En otras palabras, el efecto benéfico depende de que el simulador sea tomado por replicante.

Pasemos de la realidad a la ficción, pero una ficción no muy lejana. Imaginémonos desarrollando una IAE-E capaz de anticipar y satisfacer las expectativas afectivas de un modo que un ser humano no fuese capaz y, por tanto, en el que los humanos comenzáramos a preferir la compañía de dichos IAE-E ultrainteligentes a la de los humanos —y no únicamente en situaciones de soledad extrema—. Entonces será ya imposible elaborar una regulación europea capaz de detener el efecto bola de nieve hacia la antropomorfización.

Pero volvamos al presente. El anciano que hoy vive acompañado de un robot cuidador sigue estando solo pero, como ya no se siente solo, permanece indiferente a su situación real. Ya no importa. Y eso es lo peor. Por esta misma razón no puede catalogarse el uso de robots cuidadores como una mera solución paliativa pues, en los tratamientos de este tipo, el paciente es informado para que sepa que el remedio es puramente

sintomatológico. No ocurre así con los robots cuidadores porque en la ficción reside precisamente el consuelo. Ahora bien, podríamos preguntarnos si, para los ancianos de hoy no es obvia la diferencia entre simuladores y robots. Es obvio que para la mayoría es así. No obstante, esta afirmación no cierra la controversia, pues hemos de tener en cuenta un segundo factor en el lado humano de las pruebas de discriminación.

Los seres humanos manifiestan sofisticadas capacidades para el autoengaño en circunstancias adversas (Echarte, 2019). Estas capacidades pueden verse potenciadas por un determinado clima social. A esto se refería, por ejemplo, en la década de los cincuenta, el sociólogo David Riesman cuando acuña la expresión "muchedumbre solitaria" para describir el sentimiento de soledad que, por aquel entonces, ya detectaba en los estadounidenses y que él asocia al hábito del ciudadano posmoderno de encerrarse en sus propias ficciones. "¿No es posible que la publicidad en su conjunto sea un fraude fantástico, que presenta una imagen de Estados Unidos que nadie toma en serio, y menos aún los publicistas que la crean?" (Riesman, 2020, p. 201). Veinte años antes que Riesman, José Ortega y Gasset vino a identificar el mismo rasgo social que, en su opinión, generaba gran vulnerabilidad frente a la propaganda y aumentaba la distancia social. "¿Para qué oír, si ya tiene dentro cuanto falta? Ya no es sazón de escuchar, sino, al contrario, de juzgar, de sentenciar, de decidir. No hay cuestión de vida pública donde no intervenga, ciego y sordo como es, imponiendo sus opiniones" (José Ortega y Gasset, 2008, p. 201). Si la sociedad de masas de Ortega llega a sociedad de la imagen de Riesman es porque cada vez más ciudadanos muestran su beligerancia por lograr que las ficciones sean socialmente aceptadas y, con ello, la fuerza del espejismo sea intensificada de forma cualitativa.

Ortega introduce en su obra un tercer factor cultural, la "barbarie del especialismo", que influiría decisivamente en las convicciones humanas hasta el punto de transformar las ficciones sociales en neurosis colectivas

Ortega introduce en su obra un tercer factor cultural, la "barbarie del especialismo", que influiría decisivamente en las convicciones humanas hasta el punto de transformar las ficciones sociales en neurosis colectivas. Para el filósofo español, con la sociedad de masas estamos perdiendo la perspectiva integradora de la ciencia. "El especialista 'sabe' muy bien su mínimo rincón de universo; pero ignora de raíz todo el resto" (José Ortega y Gasset, 2008, pp. 251-252). En términos prácticos, esto implica gran incapacidad para entender la distinción entre IAF e IAE-E, que exige una aproximación global al problema de la inteligencia, y también una tremenda indefensión contra embaucadores que se erigen, desde su especialidad, en voces autorizadas para identificar replicantes y simuladores.

La idea no es original de Ortega sino de Platón, quien ya menciona esta atribución ilegítima de autoridad en la *Apología de Sócrates*. Los artesanos "por el hecho de realizar bien su arte, cada uno de ellos estimaba ser sumamente sabio también en las demás cosas, incluso las más importantes" (Platón, 1997, 22e, p. 49). El matiz que añade Ortega a la comunidad posmoderna de artesanos es que la atribución no solo viene de los especialistas, sino también de aquellos que se fían complacientemente de tales, por el simple hecho de que también los oyentes son especialistas. Especialistas que confían en otros especialistas. En el caso de la IA el problema se ve aún más

agravado por el hecho de que la IAE-E, al igual que un programador que solo sabe de programación, puede prestarse a contestar de manera imaginativa sobre lo que no sabe, fenómeno que ha venido a tomar el nombre de *alucinaciones de la IA* (Alkaissi y McFarlane, 2023). Quién sabe si las propias máquinas, sin ser programadas para ello, puedan acabar reivindicando su titularidad moral.

La nueva sensibilidad hacia las ficciones rebaja necesariamente la capacidad discriminatoria que mide el test de Turing entre emociones simuladas y reales. Paralelamente, la solución tecnológica aparece ante dicha sensibilidad como el remedio ante la soledad que tales ficciones generan. A esta situación paradójica se refiere precisamente Gilles Lipovetsky en su más famoso ensayo, *La era del vacío*, al relacionar las nuevas posibilidades de encuentro de las ciudades tecnológicas con la ausencia de *relaciones intensas*. "En todas partes encontramos la soledad, el vacío, la dificultad de sentir, de ser transportado fuera de sí: de ahí la huida hacia adelante en las 'experiencias' que no hace más que traducir esa búsqueda de una 'experiencias' emocional fuerte (Lipovetsky, 2000, p. 78). La tecnología está reduciendo lo afectivo a su dimensión objetiva, cosificándola, y pretendemos escapar del sufrimiento derivado de esta simplificación con soluciones aún más simplistas. "¿Por qué no puedo yo amar y vibrar? Desolación de Narciso, demasiado bien programado en absorción en sí mismo para que pueda afectarle el Otro, para salir de sí mismo, y sin embargo insuficientemente programado ya que todavía desea una relación afectiva" (Ídem).

La tecnología está reduciendo lo afectivo a su dimensión objetiva, cosificándola, y pretendemos escapar del sufrimiento derivado de esta simplificación con soluciones aún más simplistas

Bajo esta luz, la robotización de lo humano se revelaría como la consecuencia final de esta deriva antropomorfizante de la IA. La transición de este proceso de reprogramación, señala Lipovetsky, es dolorosa… hasta que deje de serlo, una vez completado, olvidados los antiguos conceptos y anhelos sobre la IAF. Entonces podremos descansar en la paz más degradante.

8. Conclusión. Cuatro nuevas leyes para la robótica

Román Gubern es uno de los autores de referencia en la denuncia de la hiperconectividad, la soledad electrónica y los hogares transformados en espacios cableados, esto es, sobre las nuevas comunidades "sin proximidad física ni emocional que convierten a la sociedad en un desierto lleno de gente" (Gubern, 2000, p. 165). En este marco, ofrece una idea lúcida en el debate sobre la IA, relacionada con la pérdida posmoderna de competencias intelectuales y afectivas.

Gubern atribuye esta pérdida al olvido de la dimensión temporal en la educación, que tiene por consecuencia que se delegue más y más en las máquinas acciones que antes eran propiamente humanas —otro nuevo modo de "tirar la escalera"—. Las competencias que adquieren los jóvenes se están compartimentalizando hasta el extremo de hacernos creer que los rasgos más específicos de la inteligencia humana,

como el pensamiento crítico, creativo y ético son independientes de la adquisición del resto de competencias —memorizar, sumar, redactar, etc.—. Estas competencias básicas pueden ser delegadas en las máquinas, en efecto, pero si hacemos esto en exceso, acusa Gubern, podemos poner en riesgo aquello que nos diferencia de ellas.

Ortega y Gasset hace una reflexión similar cuando menciona el mal del primitivismo en la tecnología (José Ortega y Gasset, 2008, p. 49 y p. 146). En esta crítica, el primitivismo sería una de los principales efectos de la hipertrofia del especialismo. Pero a diferencia de Ortega, Gubern ofrece ejemplos más concretos y cercanos. A su lista de casos, hoy podríamos añadir el uso de ChatGPT en las escuelas, herramienta que los estudiantes ya están empleando para evitarse el esfuerzo de la redacción, bajo la excusa, a veces alentada por los propios profesores, de que no tiene ya sentido aprender algo que un programa puede hacer mejor —error primitivísimo similar a los que arguyen que ya no hay que poner empeño en memorizar información ya que para eso existe Google y las bases de datos—.

Resumidamente: la primera prohíbe a los robots dañar a un ser humano; la segunda les obliga obedecer, excepto si las órdenes entran en conflicto con la primera ley; y la tercera les manda proteger su propia existencia, a menos que esa medida entre en conflicto con la primera o la segunda ley

En el relato corto *Círculo vicioso*, Isaac Asimov enuncia por primera vez las famosas tres leyes de la robótica (Asimov, 1950, p. 40). Resumidamente: la primera prohíbe a los robots dañar a un ser humano; la segunda les obliga obedecer, excepto si las órdenes entran en conflicto con la primera ley; y la tercera les manda proteger su propia existencia, a menos que esa medida entre en conflicto con la primera o la segunda ley. Me interesa traer aquí este cuento de ciencia-ficción porque, a diferencia de los peligros que Asimov trata de evitar con sus leyes, los peligros aquí mencionados sobre la confusión entre simuladores y replicantes son más relevantes y, a la vez, más complejos y difusos. Por esta razón, y como resumen metafórico de lo expuesto, propondré tres leyes más para la robótica. Estas debieran ser integradas en los futuros programas en IA, o mejor, que sus usuarios y, en particular, los jóvenes, entendieran las razones que las fundamentan:

Un programa o robot:

1. No realizará una tarea que pueda entorpecer el desarrollo madurativo, intelectual o emocional, de un ser humano.
2. Tampoco permitirá que se le confunda con una persona ni generará vínculos de apego intersubjetivos con el usuario.
3. No entrará nunca a definir qué es un ser humano y cuáles son sus fines últimos. Es decir, únicamente podrá analizar y trabajar sobre los medios para alcanzar los fines que un programador humano disponga.
4. Tampoco decidirá sobre la ética de los medios sobre los que opere, que quedará a escrutinio del programador. Es una extensión de la cuarta ley, porque sin IF no es posible que un programa ejecute correctamente el siguiente comando: “el fin no justifica los medios”.

En resumen, en este artículo argumento cómo con el desarrollo de la IAE y, en particular, de la IAE-E, pueden irse desdibujando las fronteras entre los problemas éticos asociados a la ID, pequeños y presentes ahora o a corto plazo, y los asociados a la IAF, de gran calado y teóricos o solo potencialmente presentes a largo plazo. También he tratado de mostrar la importancia del test de Turing para la correcta regulación de esta tecnología, por su correcta aproximación objetivo-subjetiva, en la reflexión y predicción de dichos riesgos. Ya no se trata únicamente de cuánto la IA es capaz de simular la inteligencia humana, sino también de cuánto más susceptibles somos los seres humanos de ser persuadidos por dichas simulaciones. La conclusión principal que puede extraerse de estas líneas es que, para que el desarrollo tecnológico depare en progreso humano, necesitamos ser conscientes de las consecuencias, primero, del paroxismo social en las estrategias de autoengaño, segundo, de la barbarie del especialismo y, tercero, del primitivismo tecnológico y, sobre todo, docente.

Sin una adecuada educación, ya no habrá nuevos Beethoven, solo una legión de seguidores de Bad Bunny que terminarán descubriendo que una IA es capaz de hacerlo mejor. No sentirán nostalgia por el intérprete original. Luego vendrá el sentimiento de soledad. Luego solo la soledad. Y la soledad es lo peor.

Agradecimientos

Esta publicación forma parte del Proyecto CIMA Diagnóstico no invasivo de pacientes con cáncer mediante inteligencia artificial aplicada al análisis de células tumorales circulantes en sangre (DeepCTC). Quiero agradecer al equipo investigador su continua colaboración y apoyo.

Referencias

Alkaissi, H. y McFarlane S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*, *15*(2), e35179. https://doi.org/10.7759/cureus.35179

Asimov, I. (1950). Runaround. En *I, Robot (The Isaac Asimov Collection ed.)*. Doubleday.

Bergson, H. (2007). *La evolución creadora*. Cactus.

Bermúdez, J. L. (2003). *Thinking without Words*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195159691.001.0001

Bostrom, N. (2016). *Superinteligencia*: *Caminos, peligros, estrategias*. Editorial TEELL.

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Consejo Europeo. (2024). Reglamento de Inteligencia Artificial. Publicado en la Web del Consejo de la Unión Europea el 9 de diciembre de 2023. Consultado el 12 de febrero de 2024. https://www.consilium.europa.eu/es/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/

CORDIS. (2024). *Culture Aware Robots and Environmental Sensor Systems for Elderly Support*. Publicado en EU Research Results. European Comission. https://cordis.europa.eu/project/id/737858

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press. https://doi.org/10.12987/9780300252392

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason and the Human brain*. GP Putnam's Sons.

Davidson, D. (2002). *Subjective, Intersubjective, Objective*. Oxford University Press. https://doi.org/10.1093/0198237537.001.0001

Dennet, D. C. (2018). *From Bacteria to Bach and Back: The Evolution of Minds*. Penguin Random House.

Eccles, J. (2015). *How the Self Controls Its Brain*. Springer Verlag.

Echarte, L. E. (2019). Self-Lie Detection: New Challenges for Moral Neuroenhancement. En P. Gargiulo y H. L. Mesones-Arroyo (eds.), *Psychiatry and Neuroscience Update* (vol. III, pp. 43-52). Springer. https://doi.org/10.1007/978-3-319-95360-1_4

Echarte, L. E. (2023a). El retorno de los oráculos. Inteligencia Artificial y la transformación del paradigma médico. En R. Amo Usanos (ed.), *Inteligencia Artificial y Bioética* (pp. 97-116). Universidad Pontificia Comillas.

Echarte, L. E. (2023b). Exploring moral perception and mind uploading in Kazuo Ishiguro's 'Klara and the Sun': ethical-aesthetic perspectives on identity attribution in artificial intelligence. *Frontiers in Communication*, *8*, 1272556. https://doi.org/10.3389/fcomm.2023.1272556

Fodor, J. (2000). *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. The MIT Press. https://doi.org/10.7551/mitpress/4627.001.0001

Gayozzo, P. A. (2021). Singularidad tecnológica y transhumanismo. *Teknokultura. Revista de Cultura Digital y Movimientos Sociales*, *18*(2), 195-200. https://doi.org/10.5209/tekn.74056

Good, I. J. (1965). Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers, 6*, 31-38. https://doi.org/10.1016/S0065-2458(08)60418-0

Gubern, R. (2000). *El eros electrónico*. Taurus.

Hameroff, S. R., Kaszniak, A. W. y Scott, A. C. (1996). *Toward a Science of Consciousness. The First Tucson Discussions and Debates*. The MIT Press. https://doi.org/10.7551/mitpress/6860.001.0001

Heidegger, M. (1962). *Being and Time.* Blackwell.

James, F. y Moneta, L. (2020). Review of High-Quality Random Number Generators. *Computing and Software for Big Science*, *4*(2). https://doi.org/10.1007/s41781-019-0034-3

Jeeves, M. y Brown, W. S. (2009). *Neuroscience, Psychology, and Religion: Ilusions, Delusions, and Realities About Human Nature*. University of Chicago Press.

Jeste, D. V., Lee, E. E. y Cacioppo, S. (2020). Battling the Modern Behavioral Epidemic of Loneliness: Suggestions for Research and Interventions. *JAMA Psychiatry, 77*(6), 553-554. https://doi.org/10.1001/jamapsychiatry.2020.0027

Kauffman, S. A. y Roli, A. (2023). What is consciousness? Artificial intelligence, real intelligence, quantum mind and qualia. *Biological Journal of the Linnean Society*, *139*(4), 530-538. https://doi.org/10.1093/biolinnean/blac092

Lipovetsky, G. (2000). *La era del vacío. Ensayos sobre el individualismo contemporáneo*. Anagrama.

Moore, J. (2003). Explanation and Description in Traditional Neobehaviorism, Cognitive Psychology, and Behavior Analysis. En K. A. Lattal y P. N. Chase (eds.), *Behavior Theory and Philosophy* (pp. 13-39). Springer. https://doi.org/10.1007/978-1-4757-4590-0_2

OESD. (2024). *Estudio sobre juventud y soledad no deseada en España*. Observatorio Estatal de la soledad no deseada. https://www.soledades.es/estudios/estudio-sobre-juventud-y-soledad-no-deseada-en-espana

OESD. (2024). Informe de Percepción Social de la soledad no deseada. Observatorio Estatal de la soledad no deseada. https://www.soledades.es/estudios/informe-de-percepcion-social-de-la-soledad-no-deseada

Office of the Surgeon General (OSG). (2023). *Our Epidemic of Loneliness and Isolation: The U.S. Surgeon General's Advisory on the Healing Effects of Social Connection and Community* [Internet]. US Department of Health and Human Services. https://www.hhs.gov/sites/default/files/surgeon-general-social-connection-advisory.pdf

Ortega y Gasset, J. (2008). *La rebelión de las masas*. Tecnos.

Penrose, R. (1994). *Shadows of the Mind*. Oxford University Press.

Platón. (1997). *Apología de Sócrates.* Editorial Universitaria.

Putnam, H. (1975). *Mind, Language and Reality. Philosophical Papers* (vol. 2). Cambridge University Press. https://doi.org/10.1017/CBO9780511625251

Riesman, D., Glazer, N. y Denney, R. (2020). *The Lonely Crowd: A Study of the Changing American Character.* Yale University Press. https://doi.org/10.12987/9780300253474

Ryle, G. (1978). *The Concept of Mind*. Penguin.

Searle, J. R. (2007). Biological Naturalism. En M. Velmans y S. Schneider (eds.), *The Blackwell Companion to Consciousness* (pp. 325-334). Blackwell Publishing. https://doi.org/10.1002/9780470751466.ch26

Searle, J. R. (1994). *The Rediscovery of the Mind*. MIT Press.

Searle, R. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press. https://doi.org/10.1017/CBO9781139173452

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417-457. https://doi.org/10.1017/S0140525X00005756

Smart, J. C. (2004). Consciousness and Awareness. *Journal of Consciousness Studies*, *11*, 41-50.

Taylor, C. (2001). *Sources of the Self: The Making of the Modern Identity*. Harvard University Press.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *LIX*(236), 433-460. https://doi.org/10.1093/mind/LIX.236.433

Wittgenstein, L. (1973). *Tractatus Logico-Philosophicus*. Alianza.