SCIO. Revista de Filosofía, n.º 18, Julio de 2020, 207-234, ISSN: 1887-9853

INTELIGENCIA ARTIFICIAL LINGÜÍSTICA PERFECTA: EFECTOS SOBRE LA AUTOPERCEPCIÓN DEL SER HUMANO

PERFECT NLP-ARTIFICIAL INTELLIGENCE: EFFECTS ON HUMAN BEING'S SELF-PERCEPTION

Manuel Carabantes López[a*]

Fechas de recepción y aceptación: 25 de abril de 2019 y 29 de abril de 2020

Resumen: Recientes avances en Inteligencia Artificial (IA) como el *Project Debater* de IBM son indicadores de que la IA lingüística perfecta, es decir, capaz de superar el test de Turing y simular la conducta lingüística humana de manera indiscernible, es una realidad cercana que implicará numerosos efectos sobre nuestras vidas. Como propone la filosofía de la técnica de Langdon Winner, esos efectos deben ser evaluados antes de que la tecnología sea implantada para proporcionar así a la sociedad la posibilidad de decidir si el posible cambio es conforme a su proyecto ético y político. Para contribuir a esa decisión, en este artículo pronosticamos las consecuencias de esa tecnología sobre la autopercepción del ser humano mediante dos argumentos. Por un lado, un argumento principal utilizando como premisas la caracterización de esa IA y varias teorías de la psicología social y de la personalidad. Y, por el otro, una inferencia por analogía con la antropología cartesiana.

Palabras clave: inteligencia artificial, test de Turing, efecto Flynn, filosofía de la técnica, lenguaje natural.

[a] Facultad de Filosofía. Universidad Complutense de Madrid.

[*] Correspondencia: Universidad Complutense de Madrid. Facultad de Filosofía. Calle Profesor Aranguren 5. 28040 Madrid. España.

E-mail: manuel.carabantes@gmail.com

Abstract: Recent advances in Artificial Intelligence (AI) such as IBM's *Project Debater* are indicators that the perfect NLP-AI (Natural Language Processing-Artificial Intelligence), that is, an AI capable of passing the Turing Test and indiscernibly simulating human linguistic behavior, is a close reality that will have several effects on our lives. As proposed by Langdon Winner's philosophy of technology, these effects must be evaluated before the technology is implemented to provide society with the possibility of deciding whether the possible change is in accordance with its ethical and political project. To contribute to that decision-making, in this paper we prognosticate the consequences of that technology on human being's self-perception using two arguments. On the one hand, a main argument, using as premises the characterization of this AI and several theories of Social and Personality Psychology. And on the other, an analogical inference from the effects of the Cartesian anthropology on Descartes himself.

Keywords: artificial intelligence, Turing test, Flynn effect, philosophy of technology, natural language.

§1. Introducción

Como advierte Langdon Winner, las innovaciones tecnológicas son similares a los decretos legislativos o las fundaciones políticas en tanto que "establecen un marco de orden público que perdurará por muchas generaciones" (Winner, 2008: 68). De ahí el llamamiento de este filósofo a la reflexión anticipada sobre lo que la ciencia y sus aplicaciones técnicas están cerca de alumbrar. Porque los efectos de la técnica sobre la vida de los seres humanos son difícilmente reversibles. Es en este marco teórico donde queremos analizar una tecnología que se vislumbra ya en el horizonte y cuyos efectos pueden ser tan profundos que debemos tomar conciencia de ellos para decidir si queremos, como sociedad, seguir caminando en la dirección actual. Recordemos las palabras de Lewis Mumford: la técnica, a diferencia del universo, no es algo externo a nosotros, y por tanto lo que de ella se derive es consecuencia tanto de nuestras acciones como de nuestra pasividad (Mumford, 1963: 6). La tecnología que nos preocupa aquí es, en concreto, la inteligencia artificial (IA) lingüística perfecta, es decir, la máquina capaz de igualar la competencia

humana en el lenguaje natural. Sus efectos previsibles son de muy diversa naturaleza: económicos, políticos, culturales, etc. En este artículo nos limitaremos a examinar su impacto sobre la autopercepción del ser humano. ¿Cómo afectará a la percepción que tiene el ser humano de sí mismo la aparición de una máquina capaz de duplicar artificialmente y de manera indiscernible el lenguaje natural, facultad que hasta ahora lo ha distinguido y definido?

Comenzaremos este estudio exponiendo brevemente las causas del fenómeno (sección §2), es decir, los motivos históricos, filosóficos e instrumentales que han impulsado la búsqueda de la IA lingüística. Lo siguiente (§3) será definir con precisión esta tecnología en los términos del test de Turing, por ser este test la medida que se ha utilizado durante toda la historia de las computadoras para decidir en qué grado estas máquinas han conseguido imitar la conducta lingüística humana. A continuación (§4) describiremos los atributos objetivos observables de la máquina (§4.1) y la tendencia a la antropomorfización (§4.2) característica del ser humano que interactuará con ella en la vida ordinaria, elementos que al unirse determinan la percepción del artefacto como un agente con cierto grado de humanidad. Lo siguiente (§5) será deducir las consecuencias del contacto con ese agente percibido como similar al ser humano, pero con una inteligencia superior en los términos en los que se promociona el concepto de inteligencia en una sociedad como la nuestra dominada por la racionalidad instrumental. Sirviéndonos de la Psicología podemos afirmar que dichas consecuencias se desarrollarán en tres fases: primero (§5.1), la comparación ascendente y el descenso de la autoestima; segundo (§5.2), la devaluación social de la inteligencia agregada al actual escenario de inversión del efecto Flynn; y tercero (§5.3), la redefinición del ser humano. Cuál será la manera en que se redefinan los seres humanos tras la implantación de la IA lingüística es algo que no podemos pronosticar, pero sí esbozaremos los posibles escenarios. Como caso que inductivamente avala nuestras observaciones hasta este punto pondremos el de Descartes (§6) y su antropología mecanicista: ante la mera hipótesis –y no ya el hecho empírico como sucederá con la IA lingüística– de la recreación técnica del cuerpo y de gran parte de las facultades intelectuales del ser humano, Descartes ejemplifica con su propia vida el intento por buscar características exclusivas que lo distingan en el universo. Finalmente (§7), en la conclusión propondremos alternativas de acción para evitar los problemas de identidad generados en el ser humano por la IA lingüística.

§2. La importancia de la IA lingüística

Desde sus inicios, la IA como disciplina científica ha tenido una meta característica por encima de cualquier otra: crear una máquina capaz de imitar la conducta lingüística de un ser humano de manera perfecta. Las razones de que se fijara tal objetivo son al menos de tres tipos: históricas, filosóficas y prácticas. En cuanto a las razones históricas, debemos recordar que la IA nace en la década de 1950 de la mano de investigadores que, como Allen Newell y Herbert Simon, estaban también implicados en la fundación del cognitivismo (Franklin, 1995: 378), un nuevo paradigma de la psicología que surgía para dar explicación a las conductas humanas más complejas que el paradigma imperante en esa época, el conductismo, no era capaz de explicar; siendo la más importante el lenguaje (Gardner, 1985: 12). Explicar el lenguaje era, para esos investigadores, triunfar donde el conductismo había fracasado, y hacerlo mediante la simulación informática era la demostración empírica propia de un paradigma que, entre sus tesis nucleares, postulaba que la mente humana es un procesador de información semejante a un programa informático (García, 2001: 18).

En segundo lugar, hay razones filosóficas por las que la IA ha perseguido siempre la creación de una máquina parlante. Y es que el concepto que el ser humano tiene de sí mismo está esencialmente ligado al lenguaje. Es algo que la filosofía ha señalado en todas sus épocas. Por citar dos ejemplos, recordemos el pasaje de los *Tópicos* a propósito del predicado *proprium* en el que Aristóteles dice: "si es hombre, es capaz de leer y escribir, y, si es capaz de leer y escribir, es hombre" (*Tópicos* A, 102a20). O el aforismo del *Tractatus*: "El lenguaje ordinario es una parte del organismo humano y no menos complicado que este" (Wittgenstein, 2000: §4002). ¿Cuál es la parte de la humanidad que va ligada al lenguaje? Obviamente, la vida intelectiva, dice Aristóteles, pues aunque las bestias también tienen lenguaje, aquel es de otra clase, más limitado que el nuestro. Aristóteles lo llama voz (φωνή), frente a la palabra (λόγος), exclusiva del ser humano (*Política*, 1253a11). El lenguaje es el medio de la inteligencia humana, y no se concibe la inteligencia humana sin lenguaje.

Es una idea que encontramos también en la cultura popular y en la etimología. Por poner otros dos ejemplos, pensemos en el mito del golem y en la palabra *dumb* en inglés. En inglés *dumb* puede significar tonto o mudo. Término de raíz germánica, originariamente se utilizaba para designar a los mudos, pero como los mudos solían padecer retraso mental por falta de una educación adaptada a su discapacidad, terminó utilizándose para significar tonto en general. Aquel que no tiene lenguaje es tonto, no es inteligente. Lo mismo se observa en el mito del golem: el golem creado por el rabino Löw para defender al gueto judío de Praga de los ataques antisemitas no podía hablar (Salfellner, 2011: 45). Por eso era un golem: *golem* significa *tonto* en hebreo. Era una enorme bestia de fuerza sobrehumana creada para proporcionar a los judíos el poderío físico del que proverbialmente carecen, pero a cambio no tenía inteligencia, era mudo: era estúpido. Ciertamente, los investigadores de la IA en su mayoría no son gente con formación filosófica, pero sí que son partícipes de estas ideas a través de sus realizaciones en la cultura popular.

Y, en tercer lugar, decimos que hay razones prácticas que han motivado la búsqueda de una IA competente en el lenguaje natural. La IA es un proyecto muy costoso que ha requerido siempre financiación de gobiernos y grandes empresas, los cuales invierten en investigación con la expectativa de obtener un rédito práctico. En Estados Unidos el principal mecenas de esta disciplina desde sus inicios ha sido DARPA (Defense Advanced Research Projects Agency) (Nilsson, 2009), una agencia militar que, en plena Guerra Fría, tenía gran interés en disponer de una máquina capaz de escuchar masivamente conversaciones telefónicas en ruso y transcribirlas al inglés (Crevier, 1993: 10). En nuestros días DARPA es tan importante en el desarrollo de la IA en Estados Unidos que Peter W. Singer, antiguo empleado del Pentágono y consultor de la CIA, estima que el 80 % de la investigación norteamericana en IA en la actualidad está financiado por los militares (Singer, 2009: 78). La única razón por la que los Gobiernos desean una IA competente en el lenguaje natural, advierte el ingeniero y filósofo Joseph Weizenbaum, es para aumentar la vigilancia sobre la población (Weizenbaum, 1976: 271). La tendencia en nuestra sociedad a la reducción de las libertades civiles es evidente, desde la Executive Order 12333 de Reagan, de 1981, hasta la USA Patriot Act de George W. Bush, de 2001. Una tecnología como esta que nos ocupa vendría, sin duda, a satisfacer la demanda de control del sistema.

§3. El test de Turing

Por todas estas razones, el lenguaje natural ha sido siempre el objetivo prioritario de la IA. Y el científico que formuló de manera socialmente más exitosa los términos en los cuales una máquina debe demostrar su competencia lingüística fue el matemático Alan Turing, razón por la cual a la prueba de aptitud lingüística por él diseñada se la denomina *test de Turing*. Lo describió por primera vez con el nombre de *juego de imitación* (*imitation game*) en su artículo de 1950 *Computing machinery and intelligence*. El juego es sencillo:

> Intervienen en él tres personas, un hombre (*A*), una mujer (*B*) y un interrogador (*C*). [...] El interrogador permanece en una habitación, separado de los otros dos. El objeto del juego para el interrogador es determinar cuál de los otros dos es el hombre y cuál es la mujer. Los distingue mediante las letras *X* e *Y*, y al final del juego dice *"X es A e Y es B"* o *"X es B e Y es A"*. El interrogador puede formular a *A* y *B* preguntas de este tipo: *¿Podría decirme, X, la longitud de su pelo?* Supongamos que *X* es *A*, luego *A* ha de contestar. *A* trata de conseguir que *C* se equivoque al identificarla. [...] Para que el tono de la voz no ayude al interrogador, las respuestas deberían ser escritas, o mejor, escritas a máquina. [...] El objeto del juego para el tercer jugador (*B*) es ayudar al interrogador. [...] Preguntamos ahora, *"¿Qué sucederá cuando una máquina se encargue del papel de A en este juego?"* (Turing, 1990: 9).

La máquina no tiene que ser necesariamente una computadora, advierte Turing, aunque reconoce que es ese el artefacto que él tiene en mente (Turing, 1990: 12). ¿Qué sucederá cuando una máquina consiga imitar de manera indistinguible la conducta lingüística de un ser humano en el test de Turing? Según él, la consecuencia será que le concederemos que es inteligente, porque eso es lo que hacemos con otros seres humanos: les concedemos el atributo de la inteligencia juzgando su conducta, sin exigir prueba directa de sus estados mentales (Turing, 1990: 29) –nótese que el planteamiento de Turing es análogo al planteamiento del problema de la intersubjetividad en los críticos de Husserl: en el *Lebenswelt* suponemos la existencia de otros como yo–. Hay que puntualizar, eso sí, que la máquina sería inteligente en sentido antrópico

(Copeland, 1993: 44), pues los chimpancés, por ejemplo, no superan el test de Turing, y no por ello les negamos un cierto tipo de inteligencia.

Tras enunciar el test de Turing como criterio para decidir cuándo una máquina ha alcanzado la competencia lingüística perfecta, se nos plantean dos cuestiones intermedias cuya resolución es imperativa para responder a la pregunta final sobre cómo afectará ese artefacto a la autopercepción del ser humano. La primera es: ¿tiene razón Turing y concederíamos inteligencia a la máquina? Y segunda, en caso afirmativo, ¿qué tipo de inteligencia sería? ¿Cualitativamente distinta, como la que otorgamos al chimpancé, o semejante a la nuestra por ser lo imitado algo tan humano como el lenguaje natural?

§4. Percepción de la IA lingüística

Nuestro enfoque para responder a estas cuestiones será, como el de Turing, de tipo conductual. Por tanto, nos situamos al margen del debate de la filosofía de la mente acerca del tipo de operaciones internas de procesamiento de la información constitutivas de pensamiento y de la subsiguiente pregunta de si una computadora puede duplicarlas para en consecuencia ser considerada un ser pensante o inteligente. Lo que nos ocupa aquí es otra cosa: nos interesa saber si el ser humano que interactuase con ella le concedería inteligencia y, en caso afirmativo, qué clase de inteligencia. La respuesta, así pues, depende de los atributos objetivos observables de la máquina y de la manera en que estos serían subjetivamente percibidos por el ser humano que interactuase con ella.

4.1 *Atributos objetivos*

Que la máquina tendrá competencia lingüística indistinguible de la del ser humano es algo que se deduce *ex hypothesi* por haber superado el test de Turing. ¿Pero qué otros atributos podemos razonablemente deducir que poseerá? En una sociedad como la nuestra, dominada por la racionalidad instrumental que persigue maximizar los beneficios por encima de cualquier consideración moral (Mumford, 1963; Marcuse, 1984; Horkheimer, 2002), esos atributos

serán diseñados para aumentar la productividad. Y dentro de sus atributos no nos interesan aquí los que emplee en su uso aislado, oculto a la mirada de la mayoría, sino solo aquellos observables, que exhiba en su interacción con el ser humano, porque serán esos los que afecten a la percepción de la máquina y por tanto los que determinen las respuestas a las dos preguntas con las que finalizamos la sección anterior y las ulteriores consecuencias que deduciremos más adelante.

La pregunta por los atributos de la máquina en su interacción con el ser humano pasa necesariamente por revisar la bibliografía de la HCI (Human Computer Interface). En ese ámbito, Jennie Gallimore y Sasanka Prabhala tienen varios trabajos experimentales que demuestran que, en la interacción del operario humano con la computadora, para satisfacer el imperativo de la racionalidad instrumental de maximizar el rendimiento de aquel, es eficiente dotar a esta de rasgos de personalidad (Gallimorey Prabhala, 2006; Prabhalay Gallimore, 2005; Woodly, Gosnell, Gallimorey Prabhala, 2007). Por citar el contenido de uno de sus estudios (Gallimorey Prabhala, 2006), es significativo el caso de los agentes informáticos con personalidades distintas para asistir a pilotos en el control de vehículos de combate no tripulados (UCAV, Unmanned Combat Aerial Vehicles). Los científicos diseñaron dos agentes informáticos con personalidad (Computer Agent Personality): CAP-A y CAP-B, cada uno con una personalidad distinta y extrema dentro de los cinco factores del modelo FFM (Five Factor Model) (Larsen y Buss, 2008; Goldberg, 1990). CAP-A era extrovertido, amable, responsable, inteligente y emocionalmente estable. CAP-B, por su parte, era todo lo contrario. Y, adicionalmente, se empleó a modo de control un tercer agente llamado CAP-NP, carente de personalidad. El resultado fue rotundo: los agentes informáticos con personalidad mejoraban el rendimiento del operador humano, y dentro de los agentes con personalidad, CAP-A fue el más exitoso.

De estudios en HCI como los citados puede deducirse que, por razones instrumentales, la IA lingüística será dotada de personalidad. En una sociedad como la nuestra, de libre mercado, es razonable anticipar que esa personalidad dependerá de las *relaciones de propiedad*. Como sucede con toda tecnología, la máquina servirá a los intereses de su propietario. Cuando trate con el propietario asumirá siempre la personalidad del sirviente, mientras que

cuando trate con el resto de seres humanos podrá presentarse como sirviente o como capataz, en un *continuum* cuyo ajuste debe determinarse empíricamente en función de la tarea asignada mediante experimentos como los que acabamos de citar: a veces cerca del extremo de CAP-A, y otras veces más cerca de CAP-B. El único rasgo que la máquina presentará siempre, cualquiera que sea el rol que adopte, es la inteligencia.

La inteligencia es uno de los factores del FFM, pero conviene examinarla por separado, en tanto que se trata de la razón de ser de la IA –auxiliar al ser humano con una inteligencia externa y eficaz–. La máquina poseerá y mostrará una inteligencia superior en los dos grandes factores en los que se suele dividir esta facultad: fluida (Gf) y cristalizada (Gc). Ciertamente, hay muchas teorías de la inteligencia; tantas al menos como definiciones de este concepto (Davidson y Kemp, 2011: 58), pero aquí tenemos que abreviar, así que lo más conveniente es que tomemos la teoría CHC, por ser la que goza de la mayor aceptación. En la teoría CHC –acrónimo de los apellidos de sus creadores: Cattell, Horn y Carroll–, la *inteligencia fluida* es aquella relacionada con el procesamiento de información novedosa y que depende de la eficacia en el funcionamiento del sistema nervioso central, mientras que la *inteligencia cristalizada* está constituida por el conjunto de habilidades y conocimientos que se adquieren y retienen en la memoria. En los humanos es habitual que los jóvenes posean una inteligencia fluida superior a la de los adultos, mientras que los adultos, por su mayor bagaje, suelen destacar sobre los jóvenes en inteligencia cristalizada. La máquina, sin embargo, será superior a todos en todo. Su inteligencia fluida será superior a la de un ser humano gracias a su mayor velocidad de computación. En cuanto a su inteligencia cristalizada, es razonable prever que será también superior gracias a las técnicas de *big data* aplicadas a internet, la gran base de datos donde se acumula todo el conocimiento humano, la mayoría codificado en lenguaje natural. En el marco de la teoría IM de Gardner, aquellas inteligencias que no estén codificadas en lenguaje natural también serán accesibles a una IA lingüística, ya que es una característica de toda inteligencia el ser codificable en un sistema simbólico (Gardner, 1993), y el metalenguaje desde el que se define todo sistema simbólico es siempre, en última instancia, el lenguaje natural.

Somos conscientes de que la IA que acabamos de describir es de tipo general, o AGI (Artificial General Intelligence), que supera con creces el alcance estrecho que caracteriza a la IA actual. Ello se debe a que la IA capaz de entender el lenguaje natural es probablemente, como señala Nick Bostrom, un *problema IA-completo* (*AI-complete problem*) (Bostrom, 2014: 86), es decir, un problema cuya solución equivale a haber resuelto el problema de crear una IA fuerte; una máquina al menos tan inteligente como un ser humano, con toda la flexibilidad intelectual que caracteriza a este. Nuestra descripción no es, por tanto, exagerada, sino ajustada a la capacidad de una IA capaz de superar el test de Turing; sin entrar en la discusión de que no hay forma de superar el test sin algún tipo de cognición, por ser este un asunto de filosofía de la mente, mientras que nuestro interés se inscribe en el campo de la CTS (Ciencia, Tecnología y Sociedad). La IA lingüística, en resumen, se presentará como una inteligencia superior, abarcadora de más conocimientos de los que cualquier ser humano es capaz de adquirir.

4.2 *Tendencia a la antropomorfización*

Recapitulando, estamos ante una máquina que exhibe dominio del lenguaje natural, una inteligencia superior en los términos instrumentales en los que se define esta facultad en nuestra sociedad, y rasgos de personalidad de configuración variable determinados por razones de productividad. Todo esto cuenta en el lado de lo objetivo a favor de la opinión de Turing, a saber: que la IA lingüística será percibida como inteligente en sentido antrópico. Pero aún tenemos que examinar el lado subjetivo de la ecuación. Nuestra tesis es que los atributos descritos, al ser observados por un ser humano que, como vamos a justificar a continuación, tiende a antropomorfizar, producirán la percepción de un cierto grado de humanidad en la máquina y, por tanto, de similitud con esta.

Ya en 1995, cuando las computadoras eran notablemente más rudimentarias, Batya Friedman demostró en un estudio que el 83 % de los estudiantes universitarios de informática entrevistados atribuían a las computadoras aspectos clave de la subjetividad (*agency*) tales como la capacidad de toma de

decisiones y/o intenciones (Friedman, 1995). Un 21 % incluso consideraba que estas máquinas eran moralmente responsables de sus errores. Cualquiera con conocimientos básicos de ética sabe que la moralidad es exclusiva de agentes libres y conscientes, atributos ambos de los que carece cualquier máquina. No obstante, nos interesa justamente esto: la tendencia del ser humano corriente a la antropomorfización. Una década después, Friedman y su equipo investigaron las relaciones de varias personas con mascotas robóticas de aspecto y conducta similares a los perros (Friedman, Kahn yHagman, 2003; Kahn et al., 2004). Las conclusiones fueron parecidas: el 49 % atribuía a los robots cualidades propias de los seres vivos (*life-likeessences*), el 60 % les otorgaba estados mentales, y un 12 % llegaba a considerarlos sujetos dignos de un tratamiento moralmente digno. Por supuesto, los investigadores no afirmaban que esas personas realmente *pensaran* que el robot estaba vivo, sino que se *comportaban* como si el robot estuviera vivo. Es la conducta, insistimos, lo que nos interesa: tal es el enfoque de Turing y el nuestro.

Más allá de los estudios científicos, también hay argumentos filosóficos a favor de la tesis antropomórfica. John Searle, conocido por ser un enemigo acérrimo de la IA fuerte, reconoce que proyectamos estados mentales sobre animales inferiores como los perros porque es una práctica exitosa (Searle, 1981: 297). De manera similar, es exitoso proyectar características humanas y de otros seres vivos sobre algunas máquinas, y por eso ocurren fenómenos como los descritos por Friedman y compañía. Aunque, como decimos, Searle es un detractor de la IA fuerte, para ser coherente con su crítica al externalismo semántico de Putnam acerca del experimento mental de la Tierra Gemela (Putnam, 1973: 700) y su defensa final de que el agua con distinta fórmula química (H_2O frente a XYZ) también debería ser reconocida como agua a pesar de ser de otro tipo (Searle, 1983: 203), Searle debería asimismo reconocer que una IA lingüística es inteligente, aun aceptando la salvaguarda de que sería un tipo distinto de inteligencia. Otro argumento de un filósofo a favor de nuestra tesis de la percepción antropomorfa de la IA lingüística lo encontramos en John Haugeland, quien dice lo siguiente del test de Turing: "Hablar no es una mera habilidad entre otras, sino además, y esencialmente, la habilidad de *expresar* inteligentemente muchas (quizás todas) otras habilidades de la inteligencia. Y sin *poseer* esas habilidades de hecho, o al menos en cierto

grado, uno no puede hablar de forma inteligente sobre ellas. Por eso el test de Turing es tan irresistible y poderoso" (Haugeland, 1997: 4). Y, por último, merece la pena reseñar el famoso caso de la secretaria de Weizenbaum, una mujer de cultura media que pasaba horas hablando con un programa de ordenador creado por aquel, a pesar de que el propio Weizenbaum le había explicado que al otro lado no había nadie que pudiera entender lo que ella le escribía (Weizenbaum, 1976: 6).

Después de todo lo expuesto, ya podemos responder a las dos preguntas que han guiado nuestro análisis hasta este punto. ¿Se concederá inteligencia a la IA lingüística? La respuesta es afirmativa. ¿Y será de tipo antrópico esa inteligencia? La respuesta es también afirmativa. Hablamos de que *se* concederá inteligencia. Nos referimos, por supuesto, al *se* (*man*) impersonal heideggeriano, que apunta a las prácticas sociales que alejan al *Dasein* de la existencia auténtica. Individuos como Searle o los científicos detrás de la máquina tendrán argumentos para oponerse a esa práctica, pero la gran mayoría, teniendo por un lado como elementos de juicio solamente los atributos exhibidos por la máquina, y siendo, por el otro, presa de la tendencia a la antropomorfización, la secundará. Así pues, alcanzadas estas conclusiones intermedias, ya podemos preguntarnos por el efecto de la IA lingüística sobre la autopercepción del ser humano. ¿Qué ocurrirá cuando los seres humanos traten con esa máquina? ¿Qué dice la psicología social sobre los efectos del contacto con semejantes seres dotados de características intelectuales superiores?

§5. Consecuencias del contacto con la IA lingüística

Atendiendo a las teorías de la psicología social sobre la comparación, la presencia constante de la IA lingüística dará lugar a tres fases consecutivas causalmente conectadas. Primero, la comparación ascendente con la inteligencia de la máquina producirá la disminución de la autoestima de los seres humanos. Segundo, se activará un mecanismo de autoprotección consistente en devaluar la deseabilidad de las características de la máquina que han producido la disminución de la autoestima, esto es, la inteligencia. Tercero, se retornará a una situación similar a la inicial en tanto que el ser humano

habrá recuperado la autoestima, pero redefinido axiológicamente: el sujeto ha cambiado. Como se puede ver, es un movimiento dialéctico. No obstante, no vamos a utilizar la filosofía hegeliana para explicarlo, sino conocimientos científicos de la psicología social.

5.1 *Primera fase: comparación ascendente y descenso de la autoestima*

Empezando por la comparación ascendente, los seres humanos tenemos un deseo innato de ser mejores que los demás, para lo cual tendemos a evaluarnos comparando nuestros rasgos, habilidades, opiniones y emociones con los de nuestros semejantes (Festinger, 1954). Según la *hipótesis de la similitud*, realizamos esa comparación eligiendo como modelos a personas parecidas a nosotros (Worchel et al., 2000: 69). Por tanto, cuanto más se parezca la máquina a nosotros, mayor será la tendencia a compararnos con ella. Como ya hemos visto, se parecerá a nosotros en tanto que estará dotada de personalidad por motivos de productividad, y en tanto que dominará el lenguaje natural, que es nada menos que un *proprium* aristotélico del ser humano. Un segundo factor que promoverá la comparación será la *utilidad*. Los seres humanos tendemos a compararnos con otros cuando es útil hacerlo (Jones y Gerard, 1967). Útil es, por ejemplo, compararse con quien compites por el mismo puesto de trabajo, que es justo lo que hará la IA lingüística: reemplazar a millones de seres humanos en tareas intelectuales, en continuación con la actual sustitución por computadoras. En consecuencia, la comparación con la IA lingüística se producirá por razones de semejanza y utilidad.

La comparación social puede tener dos sentidos: descendente y ascendente (Swanny Bosson, 2000: 600). La *comparación descendente* (*down ward comparison*) se produce cuando nos comparamos con otros a los que consideramos inferiores respecto de los atributos comparados, y es una actitud tanto más común cuanto menor es la autoestima del sujeto (Worchel et al., 2000: 71). La *comparación ascendente* (*up ward comparison*), por el contrario, se produce cuando nos comparamos con otros a los que consideramos superiores respecto de los atributos comparados. La comparación con la IA

lingüística será ascendente, ya que el término de utilidad de la comparación, como decimos, será la inteligencia, y la máquina, tal y como la hemos caracterizado, tendrá una inteligencia superior en sus dos factores (Gf y Gc). La consecuencia de la comparación ascendente está científicamente demostrada: el sujeto inferior sufre una disminución de la autoestima. Véanse, por ejemplo, el experimento de Mr. Dirty y Mr. Clean (Morse y Gergen, 1970: 148), o el estudio clásico de Morris Rosenberg, hoy en día impensable, en el que se daba cuenta de cómo los adolescentes de minorías étnicas tienen una autoestima más baja cuando estudian en escuelas no segregadas en las que hay estudiantes blancos, ya que estos, por su pertenencia a estratos sociales superiores, suelen obtener mejores calificaciones (Rosenberg, 1965). Se podría decir que la IA lingüística convertirá a los seres humanos en la nueva minoría étnica: tendremos una autoestima más baja cuanto más tratemos con ella.

5.2 *Segunda fase: devaluación social de la inteligencia*

En la segunda fase del proceso dialéctico se extenderán, por selección darwiniana, las conductas que permitan superar la crisis de autoestima producida por la comparación con la IA lingüística. Esas conductas serán de dos tipos no excluyentes: la comparación descendente compensatoria y la devaluación social de las características que hacen a la máquina superior. La *comparación descendente compensatoria* es, como hemos dicho, una conducta tanto más frecuente cuanto más baja sea la autoestima del sujeto. Un ejemplo histórico es el antisemitismo que permitió al pueblo alemán sobreponerse a las humillantes condiciones del Tratado de Versalles. Es de prever, por tanto, que la IA lingüística causará odio y desprecio. ¿Se focalizará esa reacción en la propia máquina, contribuyendo a la oleada de *neoludismo* que algunos anticipan (Singer, 2009: 292), o en un tercero, como hicieron los fascistas alemanes con los judíos? Eso es imposible anticiparlo.

En cuanto a la *devaluación social de la inteligencia* por ser esta la característica que hace a la máquina superior, será consecuencia de la inclinación humana a autodefinirse en base a rasgos distintivos (Worchel et al., 2000: 69). El ser humano se apartará progresivamente de la inteligencia como rasgo

distintivo de su humanidad. Además, no será difícil abandonar el cultivo de la inteligencia a propósito, sino todo lo contrario: será lo más natural para una civilización que, como la nuestra, lleva ya dos décadas descuidando el cultivo de la inteligencia, tal y como se desprende de la inversión del efecto Flynn. Este fenómeno es, obviamente, lo opuesto al *efecto Flynn*, que en psicometría describe el sorprendente aumento que ha acontecido a lo largo del siglo xx del *cociente intelectual* (IQ) (Flynn, 2009: 4), concepto que mide cuantitativamente el factor de inteligencia general *g*, que nosotros antes hemos desglosado en los factores Gf (inteligencia fluida) y Gc (inteligencia cristalizada). Pues bien, desde hace dos décadas, como decimos, se viene observando que el IQ no solo no aumenta, sino que se está produciendo una *inversión del efecto Flynn*: la población en general tiene un IQ cada vez más bajo. Las causas de esta pérdida no están claras pero, dado el corto período de tiempo en que se ha producido, muchos descartan que sean genéticas (Bratsberg y Rogeberg, 2018): tienen que ser ambientales. La identidad de esas causas ambientales, sin embargo, es objeto de debate (Rindermann, Becker y Coyle, 2017: 242). Sin poder precisarlas, James Flynn y Michael Shayer apuntan de manera general a que desde 1995 se ha producido un descenso en la demanda social de habilidades intelectuales (Flynn y Shayer, 2018: 120). Al disminuir la demanda de inteligencia, los individuos, lamarckianamente, han dejado de cultivarla.

¿Y por qué ha disminuido la demanda de habilidades intelectuales precisamente desde mediados de los años noventa? Haciendo propios los argumentos de Nicholas Carr, afirmamos que la causa principal ha sido la penetración de la informática en la vida cotidiana (Carr, 2010). Es cierto que las computadoras pequeñas y relativamente asequibles existen desde la década de los setenta (Ceruzzi, 2002: 226), pero eran máquinas que solo despertaban el interés de unos pocos entusiastas de la electrónica. Es en los noventa cuando las computadoras penetran de manera generalizada en las empresas y en los hogares de los países desarrollados. En la empresa se produce una profunda reestructuración de las plantillas a causa de la *paradoja de Moravec*: lo que es difícil para el ser humano, es fácil para las computadoras, y viceversa. Esto da lugar a que las tareas intelectuales más difíciles se asignen a la máquina. O, dicho de otra forma, donde antes había trabajo para tres ingenieros, ahora solo se demanda uno con un ordenador. El empleo cualificado, por tanto,

desciende (Flynn y Shayer, 2018: 120). Al mismo tiempo, en los hogares, el ser humano adquiere la condición de *homo protesicus* dependiente de la computadora en su vida personal, en un grado que aumenta cuantos más ámbitos de su existencia se informatizan: entretenimiento, consumo, relaciones sociales, economía, educación, política. En definitiva, llevamos dos décadas de incremento de la dependencia intelectual en las computadoras, tanto en lo laboral como en lo personal, y por eso el IQ de la población ha caído.

La IA lingüística, por tanto, surgirá en un escenario de inversión del efecto Flynn agravado, y en esa circunstancia vendrá a hacer lo que hicieron los ordenadores tras la Segunda Guerra Mundial: una *revolución sin revolución*. Así describe Javier Bustamante, citando a Salvador Giner, la revolución informática (Bustamante, 1993: 39). De manera análoga a como la computadora vino a satisfacer unas necesidades de control de un sistema político y económico que de no haber sido atendidas quizás habrían derivado en una transformación social, la IA lingüística vendrá a satisfacer la necesidad de un bien cada vez más escaso para mantener la última versión del mismo sistema: la inteligencia, definida en su sentido instrumental moderno. Esto, por supuesto, acentuará la inversión del efecto Flynn, pues un efecto que se origina por la dependencia de la computadora no puede sino aumentar a consecuencia de la introducción de un nuevo artefacto en el que se delegan todavía más operaciones intelectuales.

5.3 *Tercera fase: redefinición del ser humano*

Por último, la tercera fase del proceso dialéctico será la redefinición del ser humano en términos de características únicas. ¿En qué cifrarán los seres humanos su valía tras haber sido superados en lo intelectual por un ente inalcanzable? Atendiendo la división platónica del alma y de la sociedad podemos establecer tres tipos de éticas que se dan simultáneamente en cualquier sociedad y que sirven como guía de vida para los individuos: la ética socrática, que ensalza el cultivo del intelecto como es propio de los gobernantes; la ética hesiódica, que promueve los valores que convienen a la clase de los productores; y la ética homérica o *heroica*, que deben practicar los guerreros. Siendo

imposible la excelencia en el cultivo del intelecto en comparación con la IA lingüística, es seguro que las otras dos crecerán en extensión. Es un escenario tan lejano que solo podemos esbozar las posibilidades abiertas.

¿Adónde irán los seres humanos expulsados del cultivo del intelecto propio de la ética socrática que ha constituido el núcleo identitario de nuestra civilización? ¿A la ética hesiódica o a la heroica? La hesiódica es poco probable por dos razones principales. La primera es su incompatibilidad esencial con la alienación material del sistema capitalista. La segunda es que su posibilidad relativa de ser en una economía industrializada dirigida por los parámetros de racionalidad instrumental imperantes disminuye cuanto mayor sea la automatización del trabajo, por lo que el paso del tiempo va en su contra. La economía mundial ya está, en buena parte, delegada en máquinas, los Automated Trading Systems (ATS), que realizan miles de transacciones bursátiles por segundo, a velocidades y con capacidades de cálculo que dan una ventaja decisiva a sus propietarios frente a los inversores que operan de la manera tradicional y ya casi obsoleta (Pasquale, 2015). La IA lingüística, en una sociedad como la nuestra, tendrá el mismo efecto que los ATS y que todas las tecnologías productivas: acrecentará el poder económico de las élites. En ese escenario, la ética hesiódica solo podrá subsistir en entornos paralelos, a la manera de la economía de falansterios, que no pasarían de ser fútiles reediciones de experiencias utópicas que siempre han fracasado, como la *Nueva Era* de los setenta (Winner, 1989: 74).

Descartando posibilidades, los seres humanos del futuro solo tendrán dos formas de percibirse a sí mismos: a través de la ética heroica o, si directamente renuncian al cultivo de la virtud de ninguna de las partes de su alma, como lotófagos (Horkheimer y Adorno, 2009). El ya mencionado momento de odio causado por la comparación descendente que seguirá a la disminución de la autoestima es más compatible con lo primero: ya hemos dicho que fue lo que sucedió en la Alemania posterior al Tratado de Versalles. La humillación y el sentimiento de inferioridad producidos por la IA lingüística serán dos fenómenos fácilmente capitalizables por grupos antisistema. La neutralización de esa reacción dependerá de la capacidad de los Gobiernos para anestesiar a la población con un nivel de bienestar material lo suficientemente alto como para que tema perderlo, tal y como sucedió durante la Guerra

Fría. Previsiblemente, esa capacidad será alta, ya que la IA lingüística será un poderoso instrumento de control: habrá un espía escuchando y leyendo cada interacción informatizada de cada sujeto para evaluar su peligrosidad. ¿Por qué si no iba a estar un gobierno totalitario como el chino invirtiendo hoy tantos recursos en IA? Suscribimos la pregunta casi retórica de Weizenbaum: ¿para qué querrían los gobiernos y las grandes empresas una máquina así? (Weizenbaum, 1976: 271).

§6. Un caso histórico: la antropología mecanicista cartesiana

Expuesto ya nuestro argumento principal a falta solo de las conclusiones finales, nos queda dar cuenta del anunciado argumento por analogía: el caso de Descartes y su reacción a la fisiología mecanicista que él mismo propuso. Como todo argumento por analogía, este es inductivo, y por tanto puede considerarse como el camino inverso al argumento principal ya expuesto. Sin embargo, dada la problemática validez de los argumentos por analogía (Douglas, 2008), no lo proponemos como prueba, sino como simple ilustración de un caso histórico en el que se cumplió la regla general que hemos defendido como nuclear de este trabajo: ante la visión del doble que reproduce lo más propio, el ser humano retrocede y se redefine a partir de aquellas cualidades que considera que todavía le son exclusivas. Es un hecho avalado por la psicología social: "Cuando se pide a los sujetos que respondan a la pregunta "¿quién soy?", mencionan aspectos que los distinguen de los demás y en los que son únicos" (Worchel et al., 2000: 69).

El *Tratado del hombre* de Descartes es una de las obras más influyentes de fisiología humana del siglo xvii (López-Muñoz y Álamo, 2000: 247). En esa obra, que según el editor del filósofo francés, Claude Clerselier, fue proyectada como el capítulo XVIII de un libro más amplio titulado *El mundo*, Descartes rompe con la visión aristotélica de la naturaleza humana paradigmática hasta ese momento para proponer en su lugar un mecanicismo inspirado en la teoría neumática de Galeno. A modo de hipótesis –quizás para evitar ser acusado de herejía–, postula la existencia de unos seres parecidos a nosotros (Descartes, 1990: 23; AT XI: 120). Dado que es una entidad material, pura

res extensa, el cuerpo funciona en base a los dos conceptos centrales de la física cartesiana, esto es, la extensión y el movimiento (Benítez, 2013: 235). Descartes utiliza dos analogías principales para explicar el funcionamiento de esa máquina que es nuestro cuerpo: los relojes (Descartes, 1985: §16; AT XI: 341-342) y las fuentes (Descartes, 1990: 35; AT XI: 130-131). Así se describe el funcionamiento del cuerpo humano al comienzo del *Tratado del hombre*:

> Supongo que el cuerpo no es otra cosa que una estatua o máquina de tierra a la que Dios forma con el propósito de hacerla tan semejante a nosotros como sea posible, de modo que no solo confiere al exterior de la misma el color y la forma de todos nuestros miembros, sino que también dispone en su interior todas las piezas requeridas para lograr que se mueva, coma, respire y, en resumen, imite todas las funciones que nos son propias, así como cuantas podemos imaginar que tienen su origen en la materia y solo dependen de la disposición de los órganos (Descartes, 1990: 22; AT XI: 120).

La pregunta clave es: ¿cuáles son esas funciones que tienen su origen en la materia? La respuesta está en las últimas páginas:

> La digestión de los alimentos, el latido del corazón y de las arterias, la alimentación y crecimiento de los miembros, la respiración, la vigilia y el sueño; la recepción de la luz, de los sonidos, de los olores, de los sabores, del calor y tantas otras cualidades, mediante los órganos de los sentidos exteriores; la impresión de sus ideas en el órgano del sentido común y de la imaginación, la retención o la huella que las mismas dejan en la memoria; los movimientos interiores de los apetitos y de las pasiones y, finalmente, los movimientos exteriores de todos los miembros (Descartes, 1990: 109; AT XI: 201-202).

Como señala el filósofo especialista en cibernética André Robinet, la lista es significativamente larga (Robinet, 1973). Descartes apunta en el *Discurso del método* que, dado que el cuerpo es una máquina, no hay objeción de principio contra la posibilidad de replicarlo de manera artificial por completo (Descartes, 1983: 92; AT VI: 56). El ser humano, en la antropología cartesiana, queda así *casi* reducido a un trozo de materia, un ente similar a los animales, que son solo máquinas.

No obstante, hay dos facultades del ser humano que no proceden de ese complejo de tuberías que es su cerebro, sino del alma, y que son las que le confieren un lugar único en la creación: el lenguaje natural y la flexibilidad del intelecto (Descartes, 1983: 92; AT VI: 56). Y no solo no son *de facto* facultades resultantes de un mecanismo, sino que Descartes declara por principio la imposibilidad de que puedan imitarse de manera perfecta mediante un mecanismo. Lo único que puede recrearse mediante ingeniería inversa es lo que hacen los animales, debido a que en la cosmovisión cartesiana son máquinas. En cuanto al *lenguaje*, el filósofo francés, en una línea de pensamiento similar a la ya antes mencionada de Aristóteles, reconoce que los animales pueden comunicar con su voz ciertas pasiones, pero son expresiones simples, dice, mientras que el lenguaje del ser humano es complejo, capaz de una combinatoria mucho mayor que permite "responder al sentido de todo lo que se diga en su presencia" (Descartes, 1983: 92; AT VI: 57). Y respecto a la *flexibilidad del intelecto*, Descartes admite que los animales son capaces de hacer muchas cosas, como construir nidos o excavar grutas, pero entre ellos no hay ninguna especie capaz de acometerlas todas, es decir: están limitados a pocas tareas. La causa de tal limitación es la *combinatoria*: dado que sus movimientos se producen de manera semejante a como funcionan los relojes, es *moralmente* imposible, dice, que pueda haber una máquina que tenga sus disposiciones internas "en número suficiente para permitirle obrar en todas las ocurrencias de la vida de la misma manera que nuestra razón nos lo permite" (Descartes, 1983: 92; AT VI: 57). La imposibilidad de esa máquina es moral porque Descartes no dispone de demostración (Robinet, 1973).

¿A qué dedicó Descartes toda su vida? A cultivar las dos facultades que, según su teoría antropológica, lo hacían humano (ἄνθρωπος): el lenguaje natural y la flexibilidad del intelecto. Fue filósofo, físico, matemático, fisiólogo: cultivó su mente en tantas disciplinas como pudo en su corta vida. Y lo hizo, por supuesto, a través del lenguaje natural, en el que plasmó sus obras, junto con el lenguaje artificial de la matemática. Descartes, por tanto, no solo elaboró una teoría antropológica, sino que buscó su identidad como ser humano en coherencia con ella. Se dedicó a aquello que lo hacía único. Eso es justo lo que dice la psicología social que hacemos los seres humanos: tendemos a definirnos en base a aquellos rasgos que nos diferencian, que nos hacen especiales, tanto a nivel individual como de especie. Podemos afirmar, así

pues, que el espiritualismo cartesiano es la consecuencia de un proceso de descubrimiento científico que obligaba al ser humano Descartes a cultivar ciertas facultades para seguir siendo humano. El sujeto no es solo el punto de partida de la filosofía cartesiana, sino también el punto de destino sobre el que se aplican las consecuencias del pensamiento en forma de una vida dedicada a lo propiamente humano. Descartes es, en definitiva, tanto en el plano teórico de su obra como en el factual de su vida, una instancia que avala inductivamente por analogía nuestra tesis: ante la reproducción artificial de lo más propio, el ser humano da un paso atrás para refugiarse en lo que le queda como distintivo.

§7. Conclusión

¿Cómo afectará la IA lingüística a la autopercepción del ser humano? Recordemos las consecuencias deducidas: disminución de la autoestima, comparación descendente compensatoria y devaluación social de la inteligencia. Ante la aparición de esa máquina intelectualmente antropoide, el ser humano se comparará con ella motivada por razones de semejanza y utilidad. La consecuencia inmediata será, en una primera fase, un sentimiento de inferioridad por incapacidad de igualar la competencia intelectual de la IA. En la segunda fase se activarán dos mecanismos de supervivencia: la comparación descendente y la devaluación de los atributos en los que se ha sido superado. La comparación descendente es propia de individuos y colectivos con la autoestima mermada, y puede producirse respecto del causante de esa merma, es decir, buscando características que lo sitúen de manera real o imaginada en una posición superior respecto de la IA (neoludismo), o bien respecto de terceros. En cuanto a la devaluación de los atributos causantes del sentimiento de inferioridad, decíamos que se producirá un desprecio de lo intelectual en favor de otros rasgos, y que será una conducta natural en una sociedad que, desde la popularización de la informática, cada vez delega más tareas intelectuales en la máquina. Por último, la tercera fase consistirá en la autodefinición en base a otros rasgos distintivos. El sujeto que sufra ese proceso dialéctico de transformación será, en virtud de la previsible agudización de la inversión del efecto Flynn por el aumento de la informatización de la vida, todavía

menos inteligente que el actual, es decir, será menos capaz de pensamiento profundo: "adquisición consciente de conocimiento, análisis inductivo, pensamiento crítico, imaginación y reflexión" (Greenfield, 2009: 69). Esa menor inteligencia amortiguará el descenso de autoestima y marcará la nueva forma en que se autoperciba el ser humano.

Algunos objetarán que ese futuro está demasiado lejos como para preocuparse por él, o rebatirán con argumentos filosóficos que es imposible que una IA pase el test de Turing. Nuestra respuesta en ambos casos es invitarlos a observar la curva de progreso de esta tecnología. Aun siendo conscientes de que la recalcitrancia puede ser variable, y de que la falacia del primer paso exitoso es siempre una posibilidad (Dreyfus, 1992: 147), la evidencia empírica indica que estamos cada vez más cerca de la IA lingüística. Un ejemplo reciente es el *Project Debater* de IBM: una computadora capaz de debatir con seres humanos tomando información de internet. Aunque hemos planteado este trabajo como una reflexión de máximos partiendo de la hipótesis de un test de Turing superado, no es necesario que la IA lingüística sea perfecta para que el ser humano la perciba con la suficiente semejanza como para cuestionarse su propia identidad. Pensemos en el citado experimento de los perros robóticos: a pesar de su evidente desemejanza con los perros reales, sus dueños se comportaban en buena medida como si fueran de carne y hueso. Una semejanza aproximada es suficiente para caer en*el valle inquietante* (*the uncanny valley*): el momento en el que la máquina se parece demasiado al ser humano, pero no es igual, y eso resulta perturbador: un casi-humano que no es humano. ¿Y qué hay más humano que el lenguaje natural? La IA lingüística avanza, y aunque no supere el test de Turing, con que alcance el valle inquietante será suficiente para desencadenar una crisis de identidad.

Este escenario es tan preocupante que deberíamos seguir el consejo de Winner y cobrar conciencia de los efectos de esa máquina antes de que se implante para poder decidir si la transformación social que ella impondrá coincide con el futuro que queremos. La manera de evitar las consecuencias descritas es mantener a la máquina alejada del ser humano. Hay dos maneras de hacerlo: una radical y la otra moderada. La radical es suprimir los programas de IA lingüística. La moderada, sacrificar la productividad prohibiendo su diseño con rasgos de personalidad, para así mantener a la máquina lejos

del ser humano y evitar que sea percibida como similar a nosotros. La radical es, como señala Bostrom, poco probable, ya que en la carrera por una tecnología poderosa todos los competidores desean ser los primeros, porque nadie quiere ser el segundo y perder la ventaja estratégica decisiva (Bostrom, 2014: 96). La moderada depende de la voluntad política, y su determinación para rebelarse contra los intereses económicos de los que depende.

Siendo este el estado de cosas, nos atrevemos a pronosticar que no se darán ni el escenario moderado ni el radical contra el desarrollo y la implantación de la IA lingüística. El capitalismo es, como la evolución darwiniana, un sistema ciego de competencia iterada sin ninguna finalidad. En ese marco, el sujeto no tiene otra opción más que la de seguir evolucionando hacia el aumento de poder sobre el resto para evitar ser aniquilado por una competencia en continua evolución. Un ejemplo histórico importante lo encontramos en la bomba atómica. Se trata de una tecnología que, desde cierta posición moral, podría argumentarse que habría sido mejor no descubrirla. Sin embargo, al enterarse Estados Unidos en 1939 gracias a la carta Einstein-Szilárd de que la Alemania nazi estaba trabajando en semejante concepto bélico, no les quedó más remedio que impulsar un programa nuclear homólogo, que culminó con el éxito que ya conocemos. Fue ese acto de respuesta competitiva –junto con el error alemán de apostar por el deuterio en lugar de por el grafito (Galison, 2008: 46)– lo que salvó al mundo de un mal mayor. De manera análoga, ninguna gran potencia quiere ser la segunda en llegar a la IA lingüística. La IA puede ser a la guerra informática lo que la bomba atómica a la guerra convencional. El valor de esta tecnología para la defensa de cualquier nación resulta evidente cuando reparamos en hechos, como el ya mencionado, de que su desarrollo en Estados Unidos corre a cargo principalmente de fondos militares (Singer, 2009: 78).

Contemplando todos los elementos que hemos expuesto y deducido a futuro, resulta difícil no rendirse a una visión sustantiva de la técnica tan pesimista como la de Heidegger (Feenberg, 2014). Sin embargo, creemos que buscar una salvación es un imperativo moral. El propio Heidegger no tenía claro dónde podía estar la salvación del ser humano al destino al que lo *arroja* la técnica, por decirlo en su terminología. En un lugar sugiere que el crecimiento de lo salvador está en el arte, por compartir esta raíz con la técnica

(Heidegger, 1997: 146). En otro, se anuncia más fatalista, y declara que "solo un dios puede salvarnos ahora" (Heidegger, 1977: 17). Nosotros proponemos, como Winner, utilizar la reflexión filosófica para abrir la posibilidad de un futuro distinto al previsible, libre de los peligros que se advierten en el horizonte. Como dice Mumford, si la vida humana consistiera solamente en el ajuste al medio físico y social dominante, entonces habríamos dejado el mundo intacto, como hacen los demás animales (Mumford, 1963: 317). La técnica no es una fuerza transformadora del medio trascendente a este, sino parte del mismo, y como tal es un elemento más al que podemos dar forma. Por cerrado que parezca el futuro, el examen del pasado revela que todo lo sucedido pudo haber sido de otra manera. Los americanos pudieron haber desestimado el proyecto de la bomba atómica, así como los europeos podrían haber importado las tecnologías abortifacientes que encontraron en el Caribe (Schiebinger, 2008), o Ford podría haber salvado a sus clientes llamando a revisión a todas las unidades del Ford Pinto en lugar de dejar que murieran quemados porque era más barato pagar las indemnizaciones por fallecimiento que cambiar la ubicación del depósito de gasolina (Lee, 1998). La historia de la técnica es una historia de decisiones humanas guiadas por intereses. La IA lingüística es una técnica más, y como tal la posibilidad de controlar su desarrollo tiene como condición necesaria una reflexión anticipatoria de sus efectos como la que hemos realizado en este artículo para que, partiendo de ahí, podamos decidir si es compatible con nuestro proyecto ético y político.

Referencias bibliográficas

Aristóteles. *Política*. Madrid: Gredos.

Aristóteles. *Tópicos*. En *Tratados de Lógica (Órganon) I*. Madrid: Gredos.

Benítez, A. (2013). *Lógica*. Madrid: Escolar y Mayo.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Bratsberg, B., Rogeberg, O. (2018). Flynn effect and its reversal are both environmentally caused. *Proceedings of the National Academy of Sciences*, *115*(26), 6674-6678.

Bustamante, J. (1993). *Sociedad informatizada, ¿sociedad deshumanizada?* Madrid: Gaia.

Carr, N. (2010). *The shallows: How the Internet is changing the way we read, think and remember*. London: Atlantic Books.

Ceruzzi, P. (2002). *A history of modern computing*. Cambridge: The MIT Press.

Copeland, J. (1993). *Artificial intelligence: A philosophical introduction*. Oxford: Blackwell.

Crevier, D. (1993). *The tumultuous history of the research for AI*. New York: Basic Books.

Davidson, J., Kemp, I. (2011). Contemporary models of intelligence. En R. Sternberg y S. Kaufman (eds.), *The Cambridge handbook of intelligence* (pp. 58-82). Cambridge: Cambridge University Press.

Descartes, R. (1983). *Discurso del método & Reglas para la dirección de la mente*. Barcelona: Orbis.

Descartes, R. (1985). *Meditaciones metafísicas & Las pasiones del alma*. Madrid: Orbis.

Descartes, R. (1990). *El tratado del hombre*. Madrid: Alianza.

Descartes, R. (1996). *Oeuvres complètes*. Paris: Charles Adam et Paul Tannery.

Douglas, W. (2008). *Informal logic: A pragmatic approach*. Cambridge: Cambridge University Press.

Dreyfus, H. (1992). *What computers still can't do*. Cambridge: The MIT Press.

Feenberg, A. (2014). What is philosophy of technology? En *Defining Technological Literacy: Towards an Epistemological Framework*, 2nd ed., editado por John Dakers (pp. 11-21). London: Palgrave McMillan.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*, 117-140.

Flynn, J. (2009). *What is intelligence? Beyond the Flynn effect*. Cambridge: Cambridge University Press.

Flynn, J., Shayer, M. (2018). IQ decline and Piaget: Does the rot start at the top? *Intelligence* (66), 112-121.

Franklin, S. (1995). *Artificial minds*. Cambridge: The MIT Press.

Friedman, B. (1995). It's the computer's fault: Reasoning about computers as moral agents. En *Con. Companion of CHI 1995* (pp. 226-227). ACM Press.

Galison, P. (2008). Removing knowledge. En *Agnotology: The Making and Unmaking of ignorance*, editado por Robert Proctor y Londa Schiebinger (pp. 37-54). Stanford: Stanford University Press.

Friedman, B., Kahn, P., Hagman, J. (2003). Hardware companions? What online AIBO discussion forums reveal about the human-robotic relationship. *Digital Sociability*, 5(1), 273-280.

Gallimore, J., Prabhala, S. (2006). Creating collaborative agents with personality for supervisory control of multiple UCAVs. *Proceedings of the NATO HFM Symposium of Uninhabited Military Vehicles as Force Multipliers*.

García, E. (2001). *Mente y cerebro*. Madrid: Síntesis.

Gardner, H. (1985). *The mind's new science: A history of the cognitive revolution*. New York: Basic Books.

Gardner, H. (1993). *Multiple intelligences: The theory in practice*. New York: Basic Books.

Goldberg, L. (1990). An alternative 'description of personality': The Big Five-Factor structure. *Journal of Personality and Social Psychology*, 59(6), 1216-1229.

Greenfield, P. (2009). Technology and informal education: What is taught, what is learned. *Science*, 323(5910), 69-71.

Haugeland, J. (1997). What is mind design? En J. Haugeland (ed.), *Mind design II* (pp. 1-28). Cambridge: The MIT Press.

Heidegger, Martin. 1977. Only a god can save us now. *Graduate Faculty Philosophy Journal*, 6(1): 5-27.

Heidegger, M. (1997). *Filosofía, ciencia y técnica*. Santiago de Chile: Editorial Universitaria.

Horkheimer, M. (2002). *Crítica de la razón instrumental*. Madrid: Trotta.

Horkheimer, M., Adorno, Th. (2009). *Dialéctica de la Ilustración*. Madrid: Trotta.

Jones, E., Gerard, H. (1967). *Foundations of social psychology*. New York: Wiley.

Kahn, P., Freier, N., Friedman, B., Severson, R., Feldman, E. (2004). Social and moral relationships with robotic others? *Proceedings of the 2004 IEEE International Workshop on Robot and Human Interactive Communication*, 545-550.

Larsen, R., Buss, D. (2008). *Personality psychology*. New York: McGraw Hill.

Lee, Matthew. The Ford Pinto case and the development of auto safety regulations, 1893-1978. *Business and Economic History*, 27(2): 390-401.

López-Muñoz, F., Álamo, C. (2000). El tratado del hombre: Interpretación cartesiana de la neurofisiología del dolor. *Asclepio*, *52*(1), 239-267.

Marcuse, H. (1984). *El hombre unidimensional*. Barcelona: Orbis.

Morse, S., Gergen, K. (1970). Social comparison, self-consistency, and the concept of self. *Journal of Personality and Social* Psychology, *1*(1), 148-156.

Mumford, L. (1963). *Technics and civilization*. New York: Harbinger Books.

Nilsson, N. (2009). *The quest for Artificial Intelligence: A history of ideas and achievements*. New York: Cambridge University Press.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Cambridge: Harvard University Press.

Prabhala, S., Gallimore, J. (2005). Developing computer agents with personalities. *Proceedings of the 11th International Conference in Human-Computer Interaction*.

Putnam, H. (1973). Meaning and reference. *The Journal of Philosophy*, *70*(19), 699-711.

Rindermann, H., Becker, D., Coyle, T. (2017). Survey of expert opinion on intelligence: The Flynn effect and the future of intelligence. *Personality and Individual Differences* (106), 242-247.

Robinet, A. (1973). *Le déficybernétique*. Paris: Editions Gallimard.

Rosenberg, M. (1965). *Society and the adolescent self-image*. New Jersey: Princeton University Press.

Salfellner, H. (2011). *El golem de Praga: Leyendas judías del gueto*. E.U.: Vitalis.

Schiebinger, L. (2008). West indian abortifacients and the making of ignorance. En *Agnotology: The Making and Unmaking of ignorance*, editado por Robert Proctor y Londa Schiebinger (pp. 149-162). Stanford: Stanford University Press.

Searle, J. (1981). Minds, brains and programs. En J. Haugeland (ed.), *Mind design* (pp. 282-306). Cambridge: The MIT Press.

Searle, J. (1983). *Intentionality: An essasy in the philosophy of mind*. Cambridge: Cambridge University Press.

Singer, P. W. (2009). *Wired for war: The robotics revolution and conflict in the 21st century*. New York: Penguin Books.

Swann, W., Bosson, J. (2010). Self and identity. En S. Fiske, D. Gilbert, G. Lindzey (eds.), *Handbook of social psychology, vol. 1* (pp. 589-628). New Jersey: Wiley.

Turing, A. (1990). *¿Puede pensar una máquina?* Buenos Aires: Almagesto.

Weizenbaum, J. (1976). *Computer power and human reason: From judgement to calculation*. San Francisco: W. H. Freeman & Company.

Winner, L. (2008). *La ballena y el reactor: Una búsqueda de los límites en la era de la alta tecnología*. Barcelona: Gedisa.

Wittgenstein, L. (2000). *Tractatus logico-philosophicus*. Madrid: Alianza.

Woodly, R., Gosnell, M., Gallimore, J., Prabhala, S. (2007). Agents with personality: Human operator Assistants. *Proceedings of the 2007 Summer Computer Simulation Conference*, 1139-1146.

Worchel, S., Cooper, J., Goethals, G., Olson., J. (2000). *Psicología social*. Madrid: Thomson.