La pregunta correcta: cuestiones sobre la inteligencia artificial

Enrique Álvarez Villanueva. Universidad de Oviedo

Recibido 18/9/2021

Resumen

La llegada de inteligencias artificiales —que acompañen a las personas vulnerables e incluso sean nuestros amigos y amantes, por ejemplo— es cada vez menos un sueño lejano, por lo que muchas de estas cuestiones éticas han de plantearse en un horizonte temporal a medio plazo, entre los problemas actuales y la hipotética llegada de la inteligencia artificial general. Plantearse las cuestiones correctas es una labor muy complicada en general, y más si cabe en el campo de la inteligencia artificial, lleno de declaraciones y anuncios rimbombantes por parte de expertos, conspiraciones y miedos basados en las historias de ciencia ficción y un temor bien fundado a la luz del creciente número de escándalos causados por las malas prácticas de los proveedores de servicios a la hora de gestionar los datos que extraen de nuestra actividad en la red. La intención del presente trabajo es exponer una serie de problemas cuyo tratamiento normalmente queda en un segundo plano en la bibliografía a favor de otros debates, con el fin de contribuir a una visión más informada y crítica de la tecnología que parece que está por llegar para combatir el escepticismo basado en creencias de ciencia ficción y despertar un *escepticismo bien informado*.

Palabras clave: inteligencia artificial, computación afectiva, ética, filosofía de la tecnología.

Abstract

The right question. Issues about artificial intelligence

The arrival of artificial intelligence —that accompany vulnerable people and even be our friends and lovers, for example— is less and less a distant dream, so many of these ethical questions must be raised in a medium-term time horizon, between current problems and the hypothetical arrival of general artificial intelligence. Raising the right questions is a very complicated task in general, and even more so in the field of artificial intelligence, full of bombastic declarations and announcements by experts, conspiracies and fears based on science fiction stories, and a well-founded fear in light of the growing number of scandals caused by the bad practices of service providers when it comes to managing the data they extract from our activity on the network. The intention of this paper is to expose a number of issues whose treatment is usually overshadowed in the literature in favor of other debates, in order to contribute to a more informed and critical view of the technology that seems to be coming to combat skepticism based on science fiction beliefs and to awaken a *well-informed skepticism*.

Key words: Artificial Intelligence, Affective Computing, Ethics, Philosophy of Technology.

La pregunta correcta: cuestiones sobre la inteligencia artificial

Enrique Álvarez Villanueva. Universidad de Oviedo
Recibido 18/9/2021

§ 1. ¿De qué hablamos cuando hablamos de inteligencia artificial?

Existen excelentes monográficos y artículos destinados al análisis ético de la gestión de los datos que se extraen de nuestra actividad en línea, la programación de coches autónomos, la contaminación que los equipos producen y cómo se desechan, las cuestiones que los *sexbots* despiertan desde la perspectiva feminista o la multiplicidad de armas inteligentes, entre otros muchos factores (Boddington, 2017; Moniz Pereira y Barata Lopes, 2020), al igual que obras muy importantes que contemplan el futuro de la inteligencia artificial (IA) a más largo plazo (Bostrom, 2014; Tegmark, 2017). La llegada de inteligencias artificiales (en adelante, IAs) que acompañen a las personas vulnerables e incluso sean nuestros amigos y amantes es cada vez menos un sueño lejano, por lo que muchas de estas cuestiones éticas han de plantearse en un horizonte temporal a medio plazo, entre los problemas actuales y la hipotética llegada de la inteligencia artificial general.

Plantearse las cuestiones correctas es una labor muy complicada en general, y más si cabe en el campo de la inteligencia artificial, lleno de declaraciones y anuncios rimbombantes por parte de expertos, conspiraciones y miedos basados en las historias de ciencia ficción y un temor bien fundado a la luz del creciente número de escándalos causados por las malas prácticas de los proveedores de servicios a la hora de gestionar los datos que extraen de nuestra actividad en la red. La intención del presente trabajo es exponer una serie de problemas cuyo tratamiento normalmente queda en un segundo plano en la bibliografía a favor de otros debates, con el fin de contribuir a una visión más informada y crítica de la tecnología que parece que está por llegar para combatir el escepticismo basado en creencias de ciencia ficción y despertar un *escepticismo bien informado* (López Cerezo, 2018: 159).

§ 2. ¿Qué lugar ocuparán las máquinas inteligentes en unos pocos años?

A pesar de que en sus orígenes, hace ya 65 años en la Conferencia de Dartmouth, la inteligencia artificial no contemplase en absoluto la introducción de las emociones en los sistemas que se estaban planteando desarrollar, manteniendo una visión puramente cognitivista de la que pronto se quejaron algunos grandes nombres como Hubert Dreyfus (1965), cuando Rosalind Picard (1997) acuñó el término *computación afectiva* a finales de los años 90 las emociones comenzaron a recibir mucha más atención a la hora de diseñar y conceptualizar los sistemas que se estaban creando.

El tener sólo en cuenta factores cognitivos a la hora de desarrollar inteligencias recuerda a la narración de la creación de los humanos que Timeo describe a Sócrates en el diálogo del mismo nombre (*Timeo*, 33b2-9), en el que las personas fueron creadas primero sólo como cabezas, por ser elementos circulares y por tanto más perfectos. Según la narración, como estos humanos compuestos sólo de cabeza tenían problemas con los accidentes del terreno, el Demiurgo creador les dotó de extremidades. Esta analogía resulta muy iluminadora en tiempos modernos, ya que, si queremos crear máquinas inteligentes, habremos de tener en cuenta muchas cuestiones que caen fuera de lo que tradicionalmente se ha entendido como factores cognitivos, como las emociones y la encarnación (*embodiment*), que van cobrando cada vez más importancia a la hora de entender por qué la inteligencia humana ha llegado a ser lo que es.

Los avances en neurociencia y psicología han puesto de relieve la estrecha interconexión entre la corteza prefrontal, las áreas perceptivas del cerebro y los sistemas subcorticales relacionados con el procesamiento de las emociones (Rolls, 2018), de modo que el enfoque puramente cognitivista sobre la mente humana no parece que pueda ser útil para alcanzar el objetivo final de dotar de consciencia e inteligencia al estilo humano a sistemas artificiales.

David Levy (2008) escribió uno de los planteamientos pioneros sobre las futuras relaciones humano-máquina, explicando, con su optimismo característico, su convicción de que los robots emocionalmente competentes serán nuestros futuros amantes y compañeros sentimentales, una idea que ha encontrado gran aceptación desde entonces. Antes de estos planteamientos, las personas ya usaban máquinas para hacer sus experiencias sexuales más satisfactorias, e incluso mucho antes algunos

proyectos en IA se comenzaron a usar con fines psicoterapéuticos, como el programa ELIZA a mediados de los años 60 (Bostrom, 2014: 7). Pese a que ELIZA, por la tecnología disponible en su momento, no era capaz de establecer conversaciones muy convincentes, la mitad de los pacientes que la usaron en el hospital declararon que preferían hablar con ELIZA sobre sus problemas antes que con otro ser humano (Levy, 2008: 113). Este fenómeno ha dado lugar al llamado *efecto Eliza*, que describe la tendencia a antropomorfizar las conductas de las inteligencias artificiales (Zhou y Fischer, 2019: 88), también llamado *pareidolia emocional* cuando se aplica a la asunción de que los objetos tengan emociones, aunque no sea así (Vallverdú y Trovato, 2016: 7). Muchos otros estudios constatan este hecho (Zhou y Fischer, 2019: 23; Reeves y Nass, 2002). Más aún, parece que ni siquiera es necesario que el robot sea humanoide o capaz de conversar. Según un estudio, ser abrazado por un oso de peluche robótico hace a las personas más susceptibles de abrirse emocionalmente con el robot que las personas que no han tenido contacto con él (Laitinen *et al*., 2019: 380).

La constatación de estos fenómenos ha hecho más cercana la posibilidad de que estos robots acompañen a personas que necesitan asistencia o cuidados, además del ya mencionado desarrollo de robots amigos y amantes. Debido a que ya se están probando robots en acompañamiento de ancianos y en tratamientos y en psicoterapia desde hace algunos años con buenos resultados, y que incluso se ha acuñado un término para definir el estudio de la compatibilidad de interacción entre robots y personas, la robopsicología (*robopsichology*) (Libin y Libin, 2004), es de esperar que el campo continúe creciendo a medida que la tecnología mejora. La rapidez con la que este fenómeno está ocurriendo, y el hecho de que no sea necesario que las máquinas tengan funciones cognitivas superiores para que los humanos se sientan confortables con ellas da a entender que pronto veremos más frecuentemente a robots en hospitales, residencias de ancianos y quizá en nuestros propios hogares.

§ 3. Amigos, amantes y cuidadores

Por lo reseñado anteriormente parece que los robots e inteligencias artificiales podrían ser una solución al cada vez mayor envejecimiento de la población e incidencia de enfermedades mentales en Occidente, cuyos costes hacen peligrar la

estabilidad de los sistemas de bienestar (OMS, 2004: 14), además de ayudar a personas con cada vez menos tiempo libre a tener vidas emocionales y sexuales plenas (McArthur y Twist, 2017: 4).

Las numerosas noticias que aparecen cada día sobre avances en el comportamiento de las IAs y el realismo de algunos robots antropomórficos alientan un optimismo que no obstante todavía ha de ser contrastado con algunas cuestiones importantes. La primera de ellas es delimitar la complejidad que estas máquinas deban tener para cumplir su función adecuadamente. David Burden y Maggi Savin-Baden (2019: 17) distinguen entre tres tipos de *humanos virtuales* según su similitud con los humanos biológicos (las descripciones han sido ligeramente adaptadas para el presente trabajo, entendiendo que en la mayoría de los casos estos humanos sintéticos serán robots con un cuerpo físico, aunque se les llame *humanos virtuales* para aprovechar la claridad conceptual de la clasificación):

- Humanoides virtuales (*Virtual Humanoids*): entidades que presentan, hasta cierto punto, algunos de los comportamientos, emociones, pensamiento, autonomía e interacción de un humano.
- Humanos virtuales (*Virtual Humans*): entidades que pueden tener comportamientos, emociones, pensamiento, autonomía o interacción modelados a imagen y semejanza de los de los humanos.
- *Sapiens* virtuales (*Virtual Sapiens*): humanos virtuales sofisticados que alcanzan similares (o superiores) niveles de presencia, comportamientos, emociones, pensamiento, autonomía, interacción, autoconciencia y narrativa interna a los humanos.

En este momento, los robots más avanzados que existen entrarían en la categoría de humanoides virtuales, aunque las características modeladas distan notablemente de las que presenta un humano medio. Si ponemos como ejemplo a Sophia, de Hanson Robotics, uno de los *robots sociales* más avanzados hasta la fecha, vemos que las características más similares a un humano que presenta son la comprensión del lenguaje natural, el aspecto físico externo y la capacidad de seleccionar frases con sentido de un acervo precargado en su memoria interna generando pequeños cambios para adaptar las locuciones de forma pertinente y ejecutarlas con una voz que suena

bastante natural. Pese a la impresión que pueda dar inicialmente Sophia, no hay nada de autoconciencia, personalidad, emocionalidad o encarnación en ella. Básicamente, es un agente conversacional con una carcasa muy sofisticada. Pese a todo, las personas que interactúan con ella parecen comúnmente considerar que sí es capaz de reflexionar sobre las cosas que escucha y dice—ilusión potenciada por la ostentosa publicidad que Hanson Robotics hace de ella—, y algunas de sus provocativas declaraciones han sido tomadas bastante en serio por el gran público (Weller, 2017). Sophia ha llegado incluso a recibir el estatus de ciudadana de Arabia Saudí, siendo el primer robot en conseguir tal distinción.

Incluso concediendo que técnicamente parece quedarnos algo de trabajo hasta ver *sapiens* artificiales, no se puede negar que la tecnología en este campo está avanzando muy rápidamente, y algunos expertos como Ray Kurzweil (2005) pronostican que ese momento se alcanzará muy pronto. En cualquier caso, los desarrollos más inmediatos deberán tener como objetivo la producción de humanos virtuales cada vez más sofisticados, emulando capacidades humanas de forma tan precisa como sea posible (o necesario). A continuación, se expondrán algunos retos que esta misión tendrá y que no son tan frecuentemente tratados en la bibliografía sobre el asunto.

3.1. ¿Qué emulamos exactamente?

La capacidad emocional de las máquinas, como se ha comentado previamente, es un elemento importante tanto para emular la inteligencia humana general como para desarrollar agentes que interactúen con las personas en contextos sociales o de cuidado. Las emociones, fingidas o sentidas, apoyan y dan consistencia a las acciones de las máquinas que las ejecutan (Pessoa, 2017: 819), además de aprovechar la *pareidolia* emocional mencionada más arriba. El modelado de las emociones depende de la teoría que manejemos, y la cuestión de qué son las emociones y cómo funcionan está muy lejos de resolverse. La evidencia proporcionada por la fisiología, la neurología y la observación directa ha dado lugar a una plétora de diferentes teorías que casan con los fenómenos observados, causando una infradeterminación difícilmente resoluble. Por tanto, la labor del ingeniero de arquitecturas emocionales consiste en muchos casos en seleccionar la que más se adapta a la naturaleza computacional —siendo la teoría OCC

de Ortony, Clore y Collins (1988) en conjunción con la de las emociones básicas de Ekman (1972) una de las más populares— y generar un modelo a partir de sus principios. Esto multiplica los problemas, ya que no sólo las teorías utilizadas suelen estar discutidas dentro del campo de la psicología, sino que las arquitecturas producidas a partir de ellas también tienen que competir con otras basadas en los mismos principios. Al seleccionar diferentes acercamientos teóricos, probablemente los robots interactivos del futuro tendrán diferentes grados de precisión en la emulación, teniendo un comportamiento adecuado en determinadas situaciones y fallando en otras, obligando al refinamiento paulatino hasta alcanzar un resultado satisfactorio.

La emulación de los fenómenos observados sin tener una teoría sólida sobre la naturaleza de lo emulado no es un fenómeno nuevo en el campo de la computación, siendo otros ejemplos la producción y comprensión del lenguaje natural o la visión artificial (*computer vision*). La deuda técnica (*technical debt*) es un concepto usado en ingeniería de *software* para describir el coste adicional que debe ser pagado en el futuro como resultado de tomar un atajo (*shortcut*) cuando se desarrolla un sistema de *software* para que este funcione a pesar de tener taras, que se solucionan *ad hoc* sin solventar el problema real (Cristianini, 2019). En este caso, y dadas las implicaciones de que los humanos virtuales trabajen en tan estrecha e íntima relación con los humanos, tal vez la deuda que ha de ser pagada en el futuro venga de la mano de traumas o problemas de sociabilidad con otras personas, especialmente para las personas más vulnerables en programas de psicoterapia, considerando que hoy en día aproximadamente el 10% de las personas que reciben terapia psicológica empeoran después de haber sido tratados (Jarrett, 2008). Algunos autores ya han reflexionado sobre la capacidad de los *bots* en videojuegos de rol en línea para causar serios problemas emocionales en sus usuarios, incluso sin mostrar ningún tipo de complejidad en su comportamiento (Gunkel, 2012: 37).

3.2. Los efectos de la emulación

A la hora de diseñar los sistemas artificiales que ayuden o acompañen a personas habrán de tenerse en cuenta los estilos cognitivos (*cognitive styles*) de los usuarios, que

definen sus características individuales de percibir, recordar, pensar y resolver problemas (Culley y Madhavan, 2013: 577), amén de otras cuestiones relacionadas con la cultura a la que pertenezcan los pacientes, amigos o amantes. Esto puede también ser un arma de doble filo, teniendo en cuenta que existen algunas diferencias entre diferentes grupos culturales a la hora de expresar y justificar sus reacciones emocionales (Rogers y Pilgrim, 2014: 57). Los factores sociológicos que están detrás de este comportamiento pueden agravarse dado que la interactuación más íntima y sostenida con inteligencias artificiales que sean diseñadas para conectar con grupos sociales determinados probablemente contribuyan a la fijación de los estereotipos e injusticias sociales que deberíamos estar combatiendo. El fenómeno por el cual los seres humanos adquirimos rasgos de las personas con las que interactuamos ya es estudiado en psicología bajo el nombre de «fenómeno Michelangelo» (*Michelangelo phenomenon*) (Bartneck *et al*, 2020: 58), y no es disparatado pensar que los humanos virtuales influirán notoriamente en los usuarios con los que compartan momentos de intimidad. Si estos sistemas perpetúan estereotipos o visiones culturales del mundo concretas, quizá sean más atractivos comercialmente, pero no ayudarán a hacer más abierto a un mundo que ya adolece de una cada vez mayor polarización.

Los usuarios que han interactuado con inteligencias artificiales hasta la fecha han mostrado muy buena acogida, pero no es descartar que esto cambie en el futuro, cuando interactuar con estos sistemas no sea una novedad tan grande y las expectativas crezcan —tener una conexión a internet de 200 kbps era considerado un servicio *premium* en el año 2000, y ahora 100 mbps (cincuenta mil veces más rápido) es una tarifa estándar en muchos países—. Es cierto que la continua mejora en la velocidad de computación unido a la capacidad de aprendizaje de las máquinas es un factor que va a ayudar a refinar los modelos existentes, pero esta forma de evolucionar también tiene sus contrapartidas.

Es una historia ya muy conocida la de Tay, un *chatbot* de Microsoft creado en 2016 para la empresa Twitter. La idea era que emulase el comportamiento de una persona en la red social usando únicamente lo que aprendía explorando la web. Tay tuvo que ser clausurado tras 16 horas de funcionamiento por generar tuits racistas y sexistas altamente ofensivos con la información que había obtenido de la red (Burden y Savin-Baden, 2019: 106). Este no es el único ejemplo. Por ejemplo, en Estados Unidos, la

implantación de la IA en procesos judiciales ha mostrado sesgos claramente racistas (Bartneck *et al*., 2020: 35). De todos modos, posiblemente los ingenieros logren dar con una solución para filtrar mejor el contenido que las IAs aprenden, y en ese caso este problema estaría subsanado en el caso de máquinas cuya función sea la de acompañar y prestar servicios de terapia a los usuarios. La conexión permanente a Internet entraña todavía algunos problemas serios, como la transmisión de datos de naturaleza íntima a la nube, las posibilidades de copia de la información almacenada en sus bases de datos y el riesgo de que estas sean pirateadas y los datos usados con fines ilícitos. Incluso evitando esta posibilidad, la privacidad no estaría totalmente garantizada. En Estados Unidos, por ejemplo, las compañías que prestan servicios en la nuble están obligadas, bajo la ley conocida como *Patriot Act*, a permitir al gobierno acceder a sus bases de datos (Bartneck *et al*., 2020: 28).

Una de las ventajas que tienen los sistemas artificiales es la continua disponibilidad y mejora del servicio que prestan; por ejemplo, es claramente una ventaja que un terapeuta pueda ser transferido a la nube y encarnado en otro dispositivo conservando todo el conocimiento que tenía sin la necesidad para el usuario de encontrar a otro profesional y ponerse al día con él. No está claro, sin embargo, lo bien que funcionaría esto en el caso de amantes robóticos, ya que uno de los elementos que tienen en común la mayoría las relaciones sentimentales es un cierto sentimiento de exclusividad: la relación es entre dos individuos, y aún en relaciones de poliamor las relaciones son entre un número limitado de personas. Que un novio o esposo pueda ser transferido a la nube y copiado ilimitadamente quizá generara una nueva dimensión para los celos e hiciera perder atractivo a este tipo de relaciones humano-robot.

Otro factor es que en las relaciones sentimentales humanas normalmente existe cierto tipo de incertidumbre; no sabemos si la persona que nos atrae nos corresponderá. Los humanos virtuales tendrán la ventaja de que se plegarán a nuestros deseos, lo que puede eliminar algunos problemas como las confrontaciones en pareja o los celos, aunque no está del todo claro si también volverán las relaciones más previsibles y menos interesantes. Levy (2008 :138) argumenta que en la capacidad de cambio de los robots estará precisamente buena parte de su interés, ya que siempre podrán dar lo que los usuarios desean, pero quizá precisamente el problema que presenten sea que la accesibilidad permanente se vuelva poco deseable. Dominique

Cardon (2018) y Byung-Chul Han (2017) exponen también la constante recolección y procesamiento de los datos de los que proveemos a los algoritmos con cada acción que hacemos en línea, y sobre todo del peligro que entraña que una entidad conectada permanentemente a Internet conozca absolutamente todas nuestras aficiones, fobias y filias. Si una pareja con rasgos psicopáticos puede arruinar la vida de una persona destrozando su autoestima y convenciéndola para hacer muchas cosas contra su propio interés, ¿qué no podrá hacer un humano virtual que comparta su vida con nosotros?

En lo que se refiere a la apariencia, existe un fenómeno ampliamente documentado desde hace cincuenta años, el conocido como «valle inquietante» (*uncanny valley*). El fenómeno consiste en una sensación de extrañeza y rechazo ante robots humanoides muy cercanos en apariencia a los humanos. Melinda Mende y cols. (2019: 48) exponen que algunos estudiantes que han asistido a clase con profesores humanoides y niños en presencia de estas máquinas tenían cierta sensación de incomodidad, al igual que en el caso de ancianos cuidados por este tipo de robots, declarando preferir estar solos o ser asistidos por humanos. Este fenómeno, de todos modos, todavía no está resuelto y hay cierta controversia con su apoyo empírico (Szczuka y Krämer, 2017). Jordi Vallverdú y cols. (2018) han expuesto también cómo un mayor realismo en los robots no hace necesariamente que las respuestas que los humanos tienen hacia ellos sean más cercanas a cómo actuarían con otras personas. En el terreno sexual, además, se observa una suspensión de la incredulidad (*willing suspensión of disbelief*) por la que los humanos inmersos en relaciones íntimas con los robots actúan con ellos como harían con otras personas a pesar de ser conscientes de que no lo son, un fenómeno conocido como *ethopoieia* (Nass y Moon, 2000: 94). En cualquier caso, los riesgos de que los robots sean pirateados o influyan en las opiniones de sus usuarios permanece, con más gravedad aún si cabe, en estos casos.

Como coda a todas estas reflexiones, cabe reflexionar, siguiendo a Jacques Ellul, sobre el hecho de que los problemas causados por la tecnología son, comúnmente, corregidos con más tecnología, limitando de manera dramática la posibilidad de retornar a un mundo donde estas inteligencias artificiales no ocupen espacios íntimos. Mientras los humanos virtuales no se conviertan en *sapiens* virtuales, además, siempre será necesario un grado mayor o menor de adaptación por parte de los usuarios que

interactúen con los robots, entrando dentro del alcance operativo (*reach envelope*) del humano virtual, aceptando sus limitaciones para sacar más partido a sus ventajas. No sólo a nivel individual, sino es de esperar que la irrupción de los humanos virtuales constituya una revolución en sí misma, obligando a las personas a adaptarse y aceptar las nuevas condiciones que esta convivencia con los humanos artificiales imponga.

§ 4. El futuro de las máquinas pensantes y sus implicaciones

Al plantear la posibilidad de que los *sapiens* virtuales sean una realidad en algún momento, hemos también de analizar las implicaciones que este hecho implica más allá del punto de vista puramente utilitarista para los humanos. El concepto de singularidad ha recibido mucho protagonismo y diferentes acepciones desde que Vernor Vinge lo acuñara; estas acepciones suelen atender a criterios cognitivos, como la dada por Bostrom (2014: 4), que define singularidad como una explosión de inteligencia (*intelligence explosión*), refiriéndose a capacidad intelectiva abstracta. Otras acepciones populares de singularidad son la de que las máquinas puedan mejorarse a sí mismas o alcancen la inteligencia artificial general, que implicaría que un sistema artificial podría tener las mismas capacidades intelectuales que un humano o superarlas. Las personas que contemplan que tal escenario está cerca, como el caso citado de Ray Kurzweil (2005) o Ben Goertzel (Gunkel, 2012: 32), siguen por lo regular esta visión, en consonancia con la que Allen Newell y Herbert Simon propusieron con su hipótesis del sistema de símbolos físicos (SSF), según la cual «todo sistema de símbolos físicos posee los medios necesarios y suficientes para llevar a cabo acciones inteligentes» (López de Mantarás y Meseguer González, 2017: 9).

No obstante, y en la línea de que lo que se argumentado hasta ahora, para alcanzar niveles de inteligencia humana en sistemas artificiales debemos tener en cuenta otros factores como las emociones, la motivación o la encarnación. Esto aleja el horizonte de la creación de *sapiens* virtuales, pero esta sigue siendo una posibilidad que todavía tenemos que contemplar. En este apartado se harán algunas reflexiones sobre esta posibilidad y algunas implicaciones que este proceso conlleva y que no aparecen reflejadas comúnmente en la bibliografía sobre el tema.

Como se ha comentado en el apartado 3.1., a falta de una teoría inequívoca sobre las emociones, por el momento los esfuerzos de implantar la capacidad emocional en los sistemas artificiales pasan por estudiar los fenómenos observables de la conducta humana y asociarlos con los datos de los que la neurociencia nos provee. Bostrom (2014: 35) o Talanov y cols. (2019), entre otros, han propuesto simulaciones del cerebro o de algunas de sus partes relacionadas con la conducta emocional con la intención de alcanzar una genuina experiencia emocional para las máquinas.

Una de las cuestiones más importantes es el hecho de la agencia moral de las máquinas, produciendo una amplia y exhaustiva bibliografía que goza de gran interés para la cuestión misma de qué es un sujeto con capacidades de agencia y paciencia (*patience*) moral, lo que fuerza, como ya había pasado con la cuestión animal (Singer, 2002), a mover el foco desde el antropocentrismo hacia una concepción más amplia. Por motivos de espacio, no será posible entrar en más detalles sobre esta interesante discusión, pero sí quizá comentar algunas cosas que se siguen de la simulación de emociones para este debate.

Los seres humanos actuamos con máquinas más o menos sofisticadas desde hace ya bastante tiempo, teniendo algunas de ellas, como Siri y Alexa, la capacidad de hablar e incluso de hacer comentarios más o menos imaginativos. A pesar de todo esto, no parece ser un elemento de debate serio si estamos esclavizando inapropiadamente a nuestro coche cada vez que requerimos de sus servicios o a nuestro congelador, que trabaja sin descanso 24 horas al día sin gozar nunca de vacaciones (excepto quizá si necesitamos descongelarlo). La pregunta por nuestro comportamiento ético hacia las máquinas parece estar ligada de alguna manera con la forma en que estas son capaces de percibir el abuso que se les hace y *sufrirlo*. Es, por tanto, un enfoque básicamente utilitarista, al menos de fondo, ya que una vez que las máquinas sean consideradas agentes morales de pleno derecho, será el momento de debatir muchos otros asuntos.

Siguiendo con el argumento de que la capacidad emocional y de autoconciencia que se está ensayando en las máquinas tiene su origen y modelo en la humana —y dejando de lado por el momento el debate de si el cerebro humano tiene alguna capacidad que lo distinga esencialmente de cualquier intento de duplicación y que lo dote de su capacidad para producir sujetos morales (Wallach y Allen, 2009: 59)—, debemos entender que, de tener éxito en este esfuerzo, los *sapiens* virtuales del futuro tendrían

una capacidad moral análoga a la humana, toda vez que sus sistemas de sensación emocional y encarnación han sido creados a imagen y semejanza de los humanos, y que la cultura de la que han aprendido todo lo que saben ha surgido de la historia humana —aunque es cierto que quizá estos *sapiens* virtuales forjen su propia historia cultural en un futuro—. Todavía nos enfrentaremos aquí al problema de la inconmensurabilidad de la experiencia de los otros, pero si las máquinas actúan *como nosotros* y parecen tener autoconciencia y racionalidad *como las nuestras*, lo más lógico es pensar en ellas como sujetos morales análogos a nosotros. Incluso si la emulación es imperfecta y no se logra que los *sapiens* virtuales tengan el mismo grado de autoconciencia y capacidad emocional, todavía cabría pensar que son capaces de placer y disgusto, aunque su responsabilidad moral sería sujeto de debate.

Hasta que los *sapiens* virtuales estén entre nosotros, si ponemos la responsabilidad en las máquinas estaremos culpando a la herramienta, ya que de ninguna manera pueden asociarse estos humanoides artificiales con estados mentales que puedan causar un efecto moral en el mundo. Es cierto que en el caso del aprendizaje profundo (*deep learning*) los ingenieros en muchas ocasiones no saben cómo una máquina ha llegado a una conclusión concreta (que puede ser racista, homófoba, etc.), pero el problema ahí reside en las acciones de millones de personas que han vertido los contenidos en la red que la máquina ha aprovechado. Lo que sí podría considerarse es ensayar una programación basada en una ética de la virtud para los humanoides y humanos virtuales, que pusiese ciertos valores como básicos para todo sistema artificial con el objetivo de limitar en lo posible el daño moral que estas máquinas producen. Esto, sin embargo, es muy difícil de establecer como principio básico, no sólo por la dificultad de establecer la base de esta ética de la virtud, sino para forzar esta ética como principio en todos los proyectos de programación de humanoides y humanos virtuales.

§ 5. Apunte final

Sin ánimo de exhaustividad, se han propuesto en el presente trabajo una batería de cuestiones abiertas a las que nos veremos expuestos en los próximos años si el ritmo de perfeccionamiento en el campo de la computación afectiva no se reduce. Si las

máquinas han de ser nuestros amigos y confidentes en el futuro, debemos estar listos para crear un ambiente en el que estas puedan sernos de utilidad sin crear excesivos problemas, considerando que la tecnología, y más una tan holística como esta, obliga siempre a la sociedad a adaptarse a ella (Winner, 1977).

Bibliografía

Bartneck, Christoph; Cristoph Lütge; Alan Wagner y Sean Welsh (2020), *An Introduction to Ethics in Robotics and IA*. Cham, Springer.

Boddington, Paula (2017), *Towards a Code of Ethics for Artificial Intelligence*. Nueva York, Springer.

Bostrom, Nick (2014), *Superintelligence. Paths, Dangers, Strategies*. Oxford, Oxford.

Burden, David y Maggi Savin-Baden (2019), *Virtual Humans Today and Tomorrow*. Nueva York, CRC Press.

Cardon, Dominique (2018), *Con qué sueñan los algoritmos. Nuestras vidas en el tiempo de los big data*. Madrid, Dado.

Culley, Kimberly y Poornima Madhavan (2013), «A Note of Caution Regarding Anthropomorphism in HCI Agents», en *Computers in Human Behavior*, 29, 3. 577-579.

Cristianini, Nello (2019), «Shortcuts to Artificial Intelligence» en Marcello Pelillo y Teresa Scantamburlo (eds.), *Machines We Trust*. Cambridge, MIT Press.

Dreyfus, Hubert (1965), *Alchemy and Artificial Intelligence*. RAND Corporation, 1965. <https://www.rand.org/pubs/papers/P3244.html> [20/11/2021]

Ekman, Paul; Friesen V. Wallace y Phoebe Ellsworth (1972), *Emotion in the Human Face: Guidelines for Research and an Integration of Findings*. Nueva York, Pergamon Press.

Gunkel, David J. (2012), *The Machine Question*. Cambridge, MIT Press.

Han, Byung-Chul (2017), *Psychopolitics: Neoliberalism and New Technologies of Power*. London, Verso.

Jarrett, Christian (2008), «When therapy causes harm», en *The Psychologist*, vol. 21, n. º1. The British Psycological Society. <https://thepsychologist.bps.org.uk/volume-21/edition-1/when-therapy-causes-harm> [20/11/2021]

Kurzweil, Ray (2005), *The Singularity is Near*. New York, Viking.

Laitinen, Arto; Marketta Niemelä y Jari Pirhonen (2019), «Demands of Dignity in Robotic Care: Recognizing Vulnerability, Agency, and Subjectivity in Robot-based, Robot-assisted, and Teleoperated Elderly Care», en *Techné: Research in Philosophy and Technology*, 23: 3. 366-401.

Levy, David (2008), *Love and Sex with Robots*. Wiltshire, Cromwell Press Ltd.

Libin, Alexander y Elena Libin (2004), «Person-Robot Interactions From the Robopsychologist's Point of View: The Robotic Psychology and Robotherapy Approach», en *Proceedings of the IEEE*, 92, 11. 1789-1803.

López Cerezo, José Antonio (2018), *La confianza en la sociedad del riesgo*. Madrid, Sello.

López de Mántaras Badia, Ramón y Pedro Meseguer González (2017), *¿Qué sabemos de? Inteligencia artificial*. Madrid, CSIC.

McArthur, Neil y Markie L. C. Twist (2017),«The rise of digisexuality: therapeutic challenges and possibilities», en *Sexual and Relationship Therapy*, vol. 32. 334-344. <https://doi.org/10.1080/14681994.2017.1397950> [21/10/2021]

Moniz Pereira, Luís y António Barata Lopes (2020), *Machine Ethics. From Machine Morals to the Machinery of Morality*. Cham, Springer.

Nass, Clifford y Youngme Moon (2000), «Machines and Mindlessness: Social Responses to Computers», en *Journal of Social Issues*, 56, 1. 81-103.

Organización Mundial de la Salud (OMS) (2004), *Invertir en salud mental*. Ginebra, OMS. <https://apps.who.int/iris/rest/bitstreams/50904/retrieve> [09/01/2022].

Ortony, Andrew; Gerald L. Clore y Allan Collins (1988), *The cognitive structure of emotions*. New York, Cambridge University Press.

Picard, Rosalind (1997), *Affective Computing*. Cambridge, MIT Press.

Pessoa, Luiz (2017), «Do Intelligent Robots Need Emotion?», en *Trends in Cognitive Sciences*, 21, 11. 817-819.

Platón (1992), *Diálogos VI*. Madrid, Gredos.

Reeves, Byron y Clifford Nass (2002), *The Media Equation. How People Treat Computers, Television, and New Media Like Real People and Places*. Stanford, CSLI.

Rogers, Anne y David Pilgrim (2014), *A Sociology of Mental Health and Illness*. New York, McGraw-Hill.

Rolls, Edmund T. (2018), *The Brain, Emotion and Depression*. New York, Oxford University Press.

Singer, Peter (2002), *Animal Liberation*. New York, Harper Collins.

Szczuka, Jessica y Nicole Krämer C. (2017), «Not Only the Lonely —How Men Explicitly and Implicitly Evaluate the Attractiveness of Sex Robots in Comparison to the Attractiveness of Women, and Personal Characteristics Influencing This Evaluation», en *Multimodal Technologies and Interaction*, 1, 3. <https://doi.org/10.3390/mti1010003> [15/11/2021].

Talanov, Max; Alexey Leukhin, Hugo Lövheim, Jordi Vallverdú, Alexander Toschev, Fail Gafarov (2019), «Modeling Psycho-Emotional States via Neurosimulation of Monoamine Neurotransmitters», en J. Vallverdú y Vincent C. Müller, *Blended Cognition. The Robotic Challenge*. New York, Springer, 127-156.

Tegmark, Max (2017), *Life 3.0*. Westminster, Penguin.

Vallverdú, Jordi y Vincent C. Müller (ed.) (2019), *Blended Cognition. The Robotic Challenge*. Cham, Springer.

Vallverdú, Jordi; Toyoaki Nishida; Yoshimata Ohmoto; Stuart Moran y Sarah Lazare (2018), «Fake empathy and Human-Robot Interaction (HRI): A Preliminary Study», en *International Journal of Technology and Human Interaction*, 12 (1). 44-59.

Vallverdú,. Jordi y Gabriele Trovato (2016), «Emotional Affordances for Human-Robot Interaction», en *Adaptive Behavior*, 24, 5. 1-15.

Wallach, Wendell y Collin Allen (2009), *Moral Machines. Teaching Robots Right from Wrong*. New York, Oxford University Press.

Weller, Chris (2017), «The First 'Robot Citizen' in the World Once Said She Wants to 'Destroy Humans'», en *Inc.com*, recuperado de <https://www.inc.com/business-insider/sophia-humanoid-first-robot-citizen-of-the-world-saudi-arabia-2017.html> [21/11/2021]

Winner, Langdon (1977), *Autonomous Technology: Technics-out-of-control as a Theme in Political Thought*. Cambridge, MIT Press.

Zhou, Yuefang y Martin H. Fisher (eds.) (2019), *AI Love You. Developments in Human-Robot Intimate Relationships*. New York, Springer.