RESONANCIAS. Revista de Filosofía ▪ N° 17 ▪ JULIO 2024: 85-100 ▪ ISSN: 0719-790X

La IA entre nosotros: ¿qué debemos hacer?

AI Among Us: What Should We Do

DIANA INÉS PÉREZ[1]
Universidad de Buenos Aires; IIF-SADAF-CONICET
Fecha de Recepción: 08/05/2023
Fecha de Aceptación: 19/10/2023

Resumen

En este trabajo me concentro en las maneras en las que los seres humanos interactuamos con los sistemas con IA actualmente existentes en nuestras sociedades. Muestro que son cierto tipo de formas de interacción las que están en la base de la consideración de tales sistemas como "inteligentes" dado el origen mismo del desarrollo de la IA en base al "Test de Turing". Distingo dos tipos de desarrollos y sus efectos, tomando como criterios las formas de interacción que estos sistemas demandan. Finalmente, destaco algunos de los desafíos que enfrentamos en nuestros días, el rol central que las humanidades y ciencias sociales tendrán en este ámbito, y sugiero algunas líneas posibles de acción.

CÓMO CITAR ESTE ARTÍCULO:

En APA: Pérez, D. I. (2024). La IA entre nosotros: ¿qué debemos hacer? *Resonancias*, (17), 85-100. DOI: 10.5354/0719-790X.2024.74394

En MLA: Pérez, D. I. "La IA entre nosotros: ¿qué debemos hacer?" *Resonancias*, n.° 17, julio de 2024, pp. 85-100. DOI: 10.5354/0719-790X.2024.74394

[1] Doctora en Filosofía de la Universidad de Buenos Aires. Investigadora Principal del CONICET. Directora del IIF-SADAF-CONICET. Profesora Titular regular de "Metafísica" y Profesora Asociada Regular de "Fundamentos de Filosofía" del Departamento de Filosofía, Universidad de Buenos Aires. Directora del programa de actualización de posgrado "Inteligencia artificial desde una perspectiva humanística" (UBA). Especialista en metafísica de la mente y filosofía de la psicología, con competencia en filosofía del lenguaje y filosofía del arte. https://orcid.org/0000-0002-6185-7986

Palabras clave: test de Turing, robots sociales, interacción humano-máquina, regulación, segunda persona

Keywords: Turing Test, social robots, human-machine interaction, regulation, second person

Abstract

In this work I focus on the ways in which human beings interact with AI systems currently existing in our societies. I show that certain types of forms of interaction are the basis for considering such systems as "intelligent" given the very origin of the development of AI based on the "Turing Test". I distinguish two types of developments and their effects, taking as criteria the forms of interaction that these systems demand. Finally, I highlight some of the challenges we face today, the central role that the humanities and social sciences will have in this area, and I suggest some possible lines of action.

Cuando escuchamos la expresión "inteligencia artificial", nos vemos llevados a pensar, seguramente por la influencia de las películas de ciencia ficción que consumimos, en robots que conviven con seres humanos como en *Yo Robot*, en grandes máquinas que nos esclavizan y dominan como en *Matrix*, en cyborgs como *Terminator* o en voces que nos seducen como en la película *Her*. Son todos futuros que imaginamos, pero que no están tan cercanos (y, en algunos casos, incluso son tecnológicamente imposibles en este momento). Sin embargo, esto no quiere decir que la IA esté lejos de ser una realidad en nuestras vidas cotidianas; por el contrario, está hoy entre nosotros y es mucho más ubicua -y mucho más imperceptible- de lo que suponemos. Usamos muchos desarrollos tecnológicos que tienen entre sus componentes algoritmos de IA. Para dar solo algunos ejemplos más o menos conocidos: los algoritmos de recomendación de películas o música en *streaming* (como *Netflix, Spotify,* etc.), las Apps que nos permiten geolocalizarnos y nos indican qué camino seguir (*Google Maps, Waze*, etc.), los asistentes de voz (como *Alexa, Siri* y *Cortana*) y el archiconocido *ChatGPT* así como generadores de imágenes como *DALL-E* son desarrollos que involucran IA. También puede haber IA en los sistema de diagnóstico a través de imágenes médicas, en la posibilidad de mejorar la calidad del foco en la cámara fotográfica de nuestro celular, en los modelos epidemiológicos o meteorológicos que se usan para prevenir catástrofes y desarrollar políticas públicas, en vehículos y drones autónomos y robots de compañía, por mencionar solo algunos de los muchos que nos rodean. En este trabajo me voy a centrar en este tipo de desarrollos actuales de la IA, para reflexionar acerca del lugar que ocupan en nuestras vidas y para discutir la pregunta relativa a qué podemos hacer y qué actitud debemos tomar ante estos desarrollos, nosotros, los seres humanos que habitamos el mundo actual, en las sociedades actuales. Voy a dejar la ciencia ficción y la futurología que conlleva para otra ocasión.

La perspectiva que adoptaré para dar respuesta a esta pregunta es la siguiente: voy a focalizar mis reflexiones en la forma en la que *nosotros*[2] interactuamos con la IA. Es importante destacar que nuestra forma de interacción con estos dispositivos es, justamente, aquello que le otorga su estatus de máquina inteligente. En efecto, la idea de que ciertas máquinas sean inteligencias artificiales ha quedado asociada al hecho de que puedan interactuar con nosotros de forma humana. Tal es la idea central que desarrollara provocativamente Turing en 1950 cuando, por primera vez, se buscó dar una respuesta a la pregunta "¿Pueden pensar las máquinas?". En lo que sigue, en primer lugar, voy a presentar el "test de Turing" que sigue siendo una de las canónicas piedras de toque para detectar inteligencia artificial (sección I). Mostraré una tensión que se deriva de este test en lo relativo a nuestras formas de interacción con la IA (sección II). Seguidamente, dividiré en dos grandes grupos los desarrollos de IA actuales, usando como criterio de distinción las formas en las que se espera que nosotros interactuemos con ellas (sección III). Finalmente, extraeré algunas conclusiones acerca de los desafíos que estos dos tipos de desarrollos nos plantean (sección IV).

I. El Test de Turing y nuestra interacción con "máquinas inteligentes"

Para responder a la compleja cuestión acerca de si las máquinas pueden pensar, Turing (1950) propuso reemplazar esta pregunta por otra. La nueva pregunta que formuló era si las máquinas en cuestión podrían inducir a un sujeto humano a creer que estaba interactuando (más específicamente, dialogando) con otro sujeto humano. Es decir, ofreció una "operacionalización" del pensamiento, a través de la conducta (exclusivamente) lingüística de la máquina en el contexto de una interacción dialógica con un ser humano. Si la máquina era capaz de "engañar" al ser humano, se podría considerar que estaba desarrollando una conducta inteligente y que, por lo tanto, era una máquina pensante. Quisiera detenerme en algunos de los detalles de esta propuesta que serán importantes para lo que sigue.

En primer lugar, el test de Turing está basado *exclusivamente* en la conducta lingüística de la máquina.[3] Para ello, Turing propuso diseñar un artefacto capaz de entablar un diálogo que no fuera cara a cara, sino mediado tecnológicamente

---

[2] Cada vez que use la palabra "nosotros" voy a estar refiriéndome a seres humanos de todos los géneros, etnias, religiones y edades que viven en la actualidad, segunda década del siglo XXI, en las sociedades globalizadas de hoy.

[3] Tal vez por eso ha causado tanto revuelo el ChatGPT, y cada vez que alguna versión de este test es pasada exitosamente por alguna computadora, se afirma que las computadoras ya llegaron a ser seres pensantes.

(como los que hoy tenemos cuando chateamos con otra persona o con los *chatbots* de las páginas web de empresas y organismos públicos), con lo cual, no solo se trata de una conducta exclusivamente lingüística, sino además de lenguaje escrito. Asimismo, ese diálogo, para parecer humano, debía incorporar el *timing* propio del procesamiento lingüístico y cognitivo de un humano (por ejemplo, tardar más en responder cuánto es 864 + 1896 que cuánto es 2+3), debía estar desprovisto de todo elemento expresivo (tonos de voz y entonaciones específicas) lo que queda garantizado al ser lenguaje escrito y, por supuesto, debía estar programado para "esconder" todos los rasgos no humanos de la máquina, incluyendo su aspecto físico; en síntesis, tenía que *estar diseñada para "engañar" al ser humano en la interacción*.

Como resulta evidente, de acuerdo con el test de Turing, la máquina no es inteligente ni piensa en virtud del material del que está hecha (su *hardware*), ni tampoco en virtud del diseño específico de su *software* (*i.e.* es indiferente si el engaño se logra con un sistema de reglas y representaciones, redes neuronales o cualquier otro algoritmo surgido en algún futuro). Lo relevante para pasar el test no es lo que hay adentro de la caja negra, sino la conducta exhibida por el sistema. Y tampoco es relevante cualquier conducta: lo que importa es la forma de *interacción (lingüística)* que puede llegar a sostener *con los seres humanos*.

Finalmente, la máquina es inteligente porque actúa *como lo haría un ser humano*. No es ni mejor (más inteligente) ni peor (más tonta) que un ser humano. Tenemos que vernos llevados a atribuir intencionalidad, conciencia, pensamiento a la máquina en base a nuestras interacciones con ella para que podamos afirmar (según el test de Turing) que los algoritmos piensan o son inteligentes; es decir, las máquinas deben exhibir el tipo de conducta que usualmente comprendemos adoptando hacia ellos lo que Dennett (1981, 1987) denominó la actitud intencional.

II. ¿Qué actitud adoptar ante la IA? Una discusión dennettiana

Las "maquinas inteligentes" de Turing (1950) son, entonces, artefactos *diseñados* por los seres humanos para que interactúen con nosotros como lo haría otro ser humano induciéndonos, de esta manera, a *atribuirle inteligencia y pensamientos*. La expresión "inteligencia artificial" fue acuñada un poco más tarde, en 1956, cuando se reunieron en el *Darmouth College* los fundadores de esta disciplina incluyendo, entre otros, a John McCarthy, Marvin Minsky, Claude Shannon, Allen Newell y Herbert Simon. Según Russell y Norvig (2008), dado el tenor de la reunión, hubiera sido preferible el término "racionalidad computacional"

para denominar los desarrollos propuestos. Sin embargo, no solo la racionalidad humana fue objeto de indagación desde los inicios de la IA, sino también otras facultades humanas como "la creatividad, la auto-mejora, y el uso del lenguaje" (Russell y Norvig 2008: 21), además de la capacidad clasificatoria y matemática, el aprendizaje, la detección de patrones y la resolución de problemas, entre otros. En síntesis, el proyecto de la IA, desde sus inicios busca duplicar las facultades psicológicas humanas en una estofa no biológica, sino artificial.[4] Y la única forma de reconocer capacidades psicológicas humanas en algo es a través de la conducta públicamente observable. Recordemos que todos los proyectos científicos, desde las más diversas disciplinas, que buscaron avanzar en el conocimiento de la mente humana durante el siglo XX tomaron como punto de partida la conducta pública, es decir, la perspectiva de tercera persona o punto de vista objetivo. El acceso científico a la mente está basado en protocolos experimentales en los que la conducta observable resulta ser la base empírica sobre la que la teoría psicológica es validada. De hecho, la visión acerca de la mente humana que dominó la academia anglosajona durante la segunda mitad del siglo XX fue el denominado "funcionalismo" (Putnam 1981, Lewis 1972), que culminó con el desarrollo de la así llamada teoría computacional representacional de la mente humana (Fodor 1987).

Esta visión de la mente tenía una serie de presupuestos que posibilitaron tanto el desarrollo de la IA como el de las ciencias cognitivas en su conjunto.[5] El funcionalismo sostiene la autonomía del nivel psicológico (*i.e.* la autonomía del estudio de las facultades psicológicas humanas) respecto del nivel biológico. Esto posibilitó el estudio de las facultades cognitivas (por ejemplo, el desarrollo de los estudios chomskianos sobre la facultad del lenguaje) con independencia de los estudios de los mecanismos neurobiológicos involucrados en la *performance* lingüística. Lo mismo ocurrió en otros ámbitos de los estudios cognitivos de la mente, por ejemplo, en la visión, la capacidad conceptual, la teoría de la mente, etc. Esta misma autonomía de lo psicológico respecto de lo biológico está en la base del desarrollo de la IA: se busca diseñar algoritmos que reproduzcan la conducta humana en la resolución de las más variadas tareas "inteligentes", sin necesidad de preocuparnos por el material en el que estos diseños se realizan.[6] Así, el

---

[4] Russell y Norvig hablan de *duplicar*, otros autores prefieren hablar de *simular* o *modelar* las mentes humanas. Esta discusión terminológica está asociada a la cuestión metafísica de si eso que hacen las máquinas es o no es "realmente" pensamiento. Como estoy enfocándome en la forma en la que nosotros interactuamos con las máquinas, dejo de lado en este trabajo esta cuestión metafísica, y voy a usar indistintamente todos estos términos para referirme a los estados "mentales" de las máquinas; es decir, a los estados mentales atribuidos por nosotros a las máquinas, más allá de la cuestión metafísica de si esta atribución se corresponde con algún proceso que genuinamente podamos considerar un proceso mental, con independencia de nuestras prácticas atributivas.

[5] Recordemos que las ciencias cognitivas son un proyecto interdisciplinario que incluye la IA entre otras disciplinas, como la psicología, la filosofía, la lingüística, etc.

[6] En realidad, la base material sí que importa porque determina, por ejemplo, la velocidad de procesamiento, la capacidad de memoria, etc. El desarrollo de la IA no es *solamente* el desarrollo de algoritmos matemáticos, sino que es

desarrollo de duplicados, simulaciones o modelos computacionales capaces de realizar eficientemente tareas cognitivas humanas corrió en paralelo al estudio empírico de las capacidades cognitivas humanas.

Es importante remarcar que la teoría computacional representacional de la mente, tomando como caso paradigmático el lenguaje humano y la hipótesis del lenguaje del pensamiento, fue un paradigma experimental que acumuló mucha evidencia en su favor en los primeros años del desarrollo de las ciencias cognitivas en variadas áreas de investigación, incluyendo facultades de nivel inferior como la visión (Marr 1982). Sin embargo, los desarrollos de la IA que toman como punto de partida las capacidades superiores y los pensamientos proposicionales, rápidamente se encontraron con limitaciones de desarrollo en su aplicación a las capacidades más básicas.[7] Rápidamente, en el ámbito computacional se comenzó a desarrollar un paradigma alternativo, basado esta vez en las computaciones y tráfico de información que el cerebro humano maneja (en lugar de basarse en la mente humana). Nacían así las redes neuronales y los modelos conexionistas de la cognición. No me voy a detener en esta disputa; simplemente quiero remarcar que, en ambos casos (es decir, las teorías computacionales representacionales proposicionales y los modelos conexionistas), la idea que motoriza el desarrollo es la misma: que se puede diseñar un sistema artificial que duplique o simule las capacidades psicológicas humanas con independencia de lo biológico. La mente/cerebro es así independiente de la materia en el cual se realice (e independiente también de las características del cuerpo en el que se realice). Y, en ambos casos, es la conducta observable lo que resulta la base evidencial de la atribución de intencionalidad o mentalidad a un sistema (biológico o no).

Como señalé al finalizar el apartado anterior, si lo pensamos desde el punto de vista de nuestra interacción con estas máquinas inteligentes, resulta importante señalar que se trata de máquinas diseñadas por nosotros, para que nosotros adoptemos hacia ellas lo que Dennett 1987 denominó la "actitud intencional". En efecto, estas máquinas están diseñadas de tal manera que nosotros nos veamos obligados a atribuirles deseos, creencias y otros estados psicológicos si queremos tener éxito en nuestra interacción. En palabras de Dennett:

---

a la par el desarrollo de materiales y tecnologías que ejecutan los algoritmos, almacenan y transmiten información. Pero este desarrollo de la materialidad de la IA no está atada al conocimiento de la materialidad en la que los pensamientos humanos se ejecutan (es decir, en el cerebro), ni busca asemejase a ella (a diferencia de la mente artificial, el programa, que busca asemejarse a la mente humana). Es en este sentido que la materialidad "no importa", quiere decir que podría ser cualquiera, la que sea necesaria para que los algoritmos "corran".

7 El problema del marco y el problema de la explosión computacional fueron los primeros en popularizarse en contra de esta versión de la IA de "reglas y representaciones".

> Las movidas futuras de las mejores máquinas computadoras que juegan al ajedrez en la actualidad no pueden predecirse usando la actitud de diseño ni la actitud física; se han vuelto tan complejas que ni sus diseñadores pueden evaluarlas desde la actitud de diseño. La mejor manera de vencerlas es predecir sus respuestas preguntándose cuál sería la movida más racional que la máquina podría hacer, dadas las reglas y los objetivos del ajedrez. Es decir, debemos asumir … que la computadora "elegirá" la movida más racional. (Dennett 1981)[8]

Pero adoptar la actitud intencional supone rechazar la actitud de diseño, es decir, dejar de considerar a la máquina inteligente como una *máquina diseñada* por alguien para realizar una determinada función o propósito, o sea, para resolver un problema que la persona que la diseñó percibió como tal, y que buscó solucionar con ese diseño.[9] En este sentido, la IA es un objeto de diseño humano que nos obliga a no considerarlo un artefacto, una máquina, sino a considerarlo uno de nosotros. Y, es importante destacar, no podemos hacer las dos cosas simultáneamente: si bien es cierto que podemos alternar (más o menos voluntariamente) la actitud que adoptemos, no podemos sostener las dos simultáneamente (así como no podemos ver al mismo tiempo un pato y un conejo en una figura ambigua). Más aún, una de esas dos actitudes es la actitud *por default*, es decir, la que adoptamos irreflexivamente en virtud de las conductas que observemos en el sistema. Así, en virtud del diseño, la IA nos lleva a adoptar hacia ella, irreflexivamente, la actitud intencional.[10]

En mi opinión, la actitud intencional solo recoge una parte de lo involucrado en nuestras formas de comprensión de las mentes humanas en la interacción cara a cara. La perspectiva de segunda persona (Pérez y Gomila 2022) presenta una visión más acertada de los múltiples aspectos multimodales involucrados en la interacción exitosa. No solo lo expresado lingüísticamente cuenta a la hora de atribuir estados mentales,[11] sino también importa el tono de voz, la prosodia, las pausas, la dirección de la mirada, las expresiones faciales y corporales, por

---

8 La traducción es mía.

9 Adoptar la actitud de diseño hacia un sistema S, supone aceptar que S tiene un diseñador, un agente intencional que tiene propósitos y buscar crear un objeto con una función específica, para resolver un problema que tiene el diseñador. Por ejemplo, un termostato es un objeto diseñado por seres humanos para regular la temperatura de los ambientes. Adoptando la actitud de diseño, el comportamiento del sistema bajo consideración no es visto como inteligente, sino adecuado o no para cumplir el propósito del agente inteligente, que en este caso es solo el diseñador.

10 Si alguien tiene alguna duda, pensemos en la sensación que tenemos de que hay otra persona (otra conciencia) del otro lado del ChatGPT cuando fluye la interacción lingüística, sensación frenada *por diseño* cuando ante ciertas preguntas el chat nos responde "No puedo responder esa pregunta, soy un sistema de IA diseñado por la empresa OpenAI…". Alternativamente pensemos en cómo nos irrita el mal diseño de un *chatbot* cuando buscamos respuesta para un trámite en línea en nuestro banco, compañía de comunicaciones o municipio, y no nos da la opción que le pedimos, nos responde con opciones prediseñadas que no nos sirven, etc. No podemos evitar sentir otra conciencia presente cuando la interacción fluye de manera humana y no podemos evitar irritarnos por el mal diseño cuando no fluye la interacción y deseamos un ser humano del otro lado.

11 Contrariamente a lo que el test de Turing presupone.

mencionar solo algunos elementos obvios. Así, desde esta perspectiva podemos atribuir una gama más amplia de estados psicológicos (no solo los que dependen de las normas de racionalidad central a la actitud intencional dennettiana),[12] dado que parte de todas las expresiones corporales (incluyendo las emocionales) y está fundada en nuestra sensibilidad afectiva (por lo que se trata de una atribución valorativa, no neutra como la propia de la actitud intencional). Desde esta perspectiva, la presencia de elementos lingüísticos es opcional (y no se da necesariamente en el caso canónico del que parte la segunda persona: las interacciones adulto-bebé en el primer año de vida). Tal como sostiene Dennett respecto de la actitud intencional, la adopción de la perspectiva de segunda persona resulta inevitable en las interacciones humanas: se trata de la forma de comprensión humana más básica y perdurable a lo largo de toda nuestra vida, y la aplicamos no solo a los seres humanos con los que interactuamos sino también a mascotas, juguetes, personajes de ficción y, por supuesto, a artefactos con el nivel de sofisticación adecuado, como nuestras computadoras, robots, etc.

III. La IA entre nosotros: ¿personas o herramientas?

Ahora bien, todos somos conscientes de que no es el propósito de los desarrollos más interesantes en IA crear máquinas que sean *exactamente* como los seres humanos. En realidad, el objetivo más publicitado en la industria de la IA es crear artefactos con una inteligencia *superior* a la humana, máquinas que realicen las mismas tareas que los seres humanos, pero que lo hagan de una manera más eficiente, con menos errores, más rápidamente, etc. Es la existencia de este tipo de desarrollos de la IA la que nos lleva a considerar la posibilidad de ser reemplazados por ella. Así, los desarrollos actuales de IA se inclinan en dos direcciones opuestas: (1) la creación de máquinas que simulen las conductas humanas apropiadas para la interacción con seres humanos *in the wild,* es decir, en nuestro entorno social cotidiano; (2) la creación de sistemas *mejores* que los seres humanos, que realicen las tareas asignadas de manera más eficiente y rápida que los seres humanos en entornos específicos para los que fueron especialmente diseñados, aun cuando no interactúen con nosotros como lo hacen

[12] En el fondo, como señalan Russell y Norvig, parece ser más bien la racionalidad humana y no la mente humana lo que la IA busca duplicar. Es cierto que nos autocomprendemos como animales racionales, pero la mente humana no está constituida exclusivamente por la racionalidad lógico-matemática, lingüística, abstracta, sino también por múltiples estrategias inteligentes de supervivencia en la interacción con el entorno físico y social. Todos estos aspectos son los que resultan ajenos al test de Turing y a la IA concebida como modelo de la mente (racional) humana. La mente de un animal racional es mucho más que lenguaje escrito en un *display* y nuestra comprensión de estas mentes depende de mucho más que de la interacción por medio del lenguaje.

los otros seres humanos. Estos dos tipos de desarrollos, como es claro, suponen diferentes formas de interacción con nosotros: en un caso, se busca una interacción social humana típica; en el segundo caso se busca un reemplazo de un humano por una máquina para la realización de una tarea específica (queda abierta la cuestión de cómo habremos de tratar a este tipo de máquinas, dado que resuelven tareas inteligentes humanas, pero no está diseñadas para actuar como un ser humano en el contexto de la interacción, con lo que parecen quedar relegadas al ámbito del diseño artefactual).

Entre aquellos diseños de IA que se orientan a la interacción a la manera humana, se encuentran los robots sociales, que buscan reemplazar a seres humanos que realizan tareas de cuidado (de adultos mayores que viven solos, de infantes). También serían "mejores" que los seres humanos en ciertos aspectos: no se cansan ni necesitan dormir, por lo que realizarían su tarea los 365 días del año, las 24 horas del día y nunca se "olvidan" de acercarle la medicación o alimentar, siempre en el momento apropiado, a sus dueños. Algunos de ellos tienen aspecto humanoide,[13] otros están diseñados más bien a la manera de animales de compañía. En algunos casos son simples experimentos de laboratorio (todavía no hay robots que puedan interactuar en nuestra vida cotidiana como Robotina, de los Supersónicos -los *Jetsons*- o como los robots que sustituyen a una pareja muerta, como ocurre en algún capítulo de *Black Mirror*). En otros casos han sido probados en contextos terapéuticos,[14] y algunos pueden ser adquiridos hoy para compañía en portales de venta en línea. Estos artefactos incorporan en sus diseños formas de interacción a la manera humana que van más allá de lo lingüístico; por ejemplo, respondiendo al tacto con un sonido amistoso, adaptando sus respuestas a las preferencias de su usuario, interactuando corporalmente a través de movimientos y sonidos similares a los que se dan en interacciones con seres humanos (o mascotas, según corresponda).

Sin duda, la incorporación de estos robots en nuestras vidas abre la pregunta acerca de cómo se verán afectadas, cómo se alterarán las formas actuales de interacción humana y de organización social. Pero me gustaría destacar que el éxito de estos robots sociales parece depender de que efectivamente nos engañen y nos hagan creer que no son robots sino seres humanos (o mascotas). Así surge un dilema para el diseño y la introducción de estos artefactos en las sociedades humanas: o bien se diseñan dejando claro su carácter robótico (no solo en su apariencia, sino también en sus formas de interactuar con seres humanos), en cuyo caso nunca serán realmente otros como nosotros (pero no seremos engañados), o

---

[13] Por ejemplo, Myon, desarollado por el *Neurorobotics Research Laboratory* del *Berliner Hochschule für Technik* ("Neurorobotics Research Laboratory (NRL) - Myon").

[14] Por ejemplo, Paro, una foca terapéutica desarrollada por la industria japonesa AIST. "PARO Therapeutic Robot". *PARO Therapeutic Robot*, www.parorobots.com. Accedido el 16 de junio de 2024

bien se diseñan para que parezcan y actúen como seres humanos (o mascotas) y nos estarán engañando.

El segundo caso que quiero analizar es el de los sistemas con IA orientados a la resolución de una tarea puntal de una manera más eficiente, rápida y desprovista de errores que la humana, tales como muchos de los desarrollos que mencioné al iniciar este trabajo: sistemas que nos orientan en el espacio para viajar por el camino más rápido, con menos congestión de tránsito, que traducen un texto de una lengua que no conocemos a la nuestra, que diagnostican enfermedades o que deciden a quién otorgarle un crédito o un puesto de trabajo, etc. Sin duda, la mayor parte de los sistemas con IA que nos rodean hoy podríamos incluirlos aquí. Respecto de este tipo de desarrollos me gustaría señalar varias cosas.

En primer lugar, cada uno de estos desarrollos sería, en principio, excelente (mejor que los seres humanos) *para una tarea específica*. Pero estamos muy lejos de lograr que un sistema que es muy bueno en algo pueda también ser muy bueno en otra cosa. Es decir, si entrenamos un algoritmo para reconocer tumores en imágenes de ultrasonido, ese mismo algoritmo no estará *ipso facto* calificado para reconocer tumores malignos en fotografías tomadas de la piel de un paciente ni, mucho menos, para reconocer caras de personas en cámara de vigilancia. Para cada propósito específico hay que entrenar un algoritmo diferente. En segundo lugar, cada tarea específica estará orientada por un problema humano que se busca resolver con esta herramienta, por lo que tanto el planteo del problema como las soluciones que se consideran correctas del problema son problemas y soluciones humanos. Así, un algoritmo está entrenado para reconocer el tipo de tumores que conocemos y queremos diagnosticar para tratar, pero no cualquier otra cosa en las imágenes; como sabemos, las formas alternativas de cortar al mundo en partes son muchas, pero solo algunas de ellas son significativas para nosotros (Dupré 1993). En tercer lugar, todo mecanismo clasificatorio tiene falsos positivos y falsos negativos, por lo que hay que tomar decisiones de diseño respecto de qué tasa de casos como estos es aceptable.[15] Esto representa un importante desafío no desprovisto de consecuencias en las vidas humanas (pensamos en los sub- o sobre-diagnósticos y lo que produciría en el sistema de salud y en las vidas de los pacientes). En cuarto lugar, todo sistema clasificatorio tiene sesgos, es decir, prejuicios y presupuestos explícitos o implícitos que forman parte constitutiva del sistema dadas las decisiones de su diseño. Estos sesgos son esenciales para que el algoritmo funcione apropiadamente (por ejemplo, que detecte los tumores que conocemos y sabemos cómo tratar), pero puede resultar un problema si buscamos clasificar personas para entrevistarlas para un trabajo o para otorgarles libertad provisional, porque esas clasificaciones van a depender

[15] Son decisiones que toma el diseñador del sistema; es decir, la empresa que lo diseña y comercializa.

del entrenamiento que haya tenido el sistema en base a los casos anteriores, fundados en los sesgos humanos propios de las prácticas humanas relevantes, y esos ejemplos de entrenamiento del algoritmo pueden estar teñidos de prejuicios y sesgos raciales, de género, etc. que no quisiéramos repetir. En otras palabras, nos basamos en nuestro conocimiento anterior de tipos de tumores y tipos de personas para entrenar los algoritmos, pero mientras que el primer sistema clasificatorio está basado en conocimiento provisto por la comunidad científica, el segundo está basado en prácticas humanas que podemos cuestionar. Más aún, en estos casos las consecuencias sociales son mucho más complejas, profundas e injustas. Un mal diagnóstico (un falso positivo o falso negativo) que nos pueda dar la IA se puede complementar con otros medios diagnósticos, clínicos, análisis de sangre, biopsias, etc., de tal manera que la decisión final de diagnóstico quedará en manos del médico (o médicos, siempre existe la posibilidad de la interconsulta), así como la decisión de qué tratamiento realizar que queda en manos del médico en acuerdo explícito con el paciente. En cambio, la exclusión de alguien de un potencial trabajo, el rechazo al pedido de libertad condicional de alguien o el otorgamiento indebido de libertad condicional a quien volverá a delinquir, o el rechazo de un pedido de crédito bancario para iniciar un emprendimiento, en tanto sean decisiones delegadas a sistemas de IA, causan un impacto irreversible en las personas afectadas y, por lo tanto, en la sociedad en su conjunto, sin que ellas puedan participar de decisión alguna.

Una vez más, la incorporación en nuestras vidas y sociedades de estos sistemas no es inocua. Trae muchos beneficios y puede ser muy útil, pero hay que ser consciente de sus limitaciones y posibles errores (sea por sesgos humanos filtrados en el diseño, sea porque ofrece falsos positivos o falsos negativos). Estos desarrollos no son infalibles y, aunque nos empeñemos en denominarlos "inteligencia" artificial, no son perfectos, cometen errores, como los seres humanos.[16] Abandonar la actitud intencional y volver a la actitud de diseño para interactuar con este tipo de desarrollos parece ser una buena estrategia para hacer visibles estas cuestiones y nos permite realizar otro tipo de preguntas con el fin de decidir si incorporar o no estos diseños en nuestras vidas, en políticas públicas, etc. Tener presente que estamos ante artefactos diseñados por seres humanos para realizar tareas humanas en sociedades humanas es fundamental para nuestra interacción éticamente responsable con estos sistemas.

---

16 Obviamente, la estrategia de seguir denominándolos "inteligencia artificial" responde a una cuestión de *marketing*, porque lo que tenga esta etiqueta se vende mejor: está instalada en nuestra sociedad la idea de que estos desarrollos son mejores que nosotros y que no fallan. Nada de eso es verdad: se equivocan y son tan buenos o malos como seamos los seres humanos que los diseñamos.

IV. Y entonces… ¿qué hacemos con la IA?

En este último apartado quisiera hacer dos propuestas concretas ante la pregunta del título: ¿Qué hacemos con la IA? La primera propuesta es promover la regulación de los desarrollos que incluyen IA. En el caso de los robots sociales, las razones son bastante claras: muchos de esos desarrollos (los que salieron del laboratorio y están siendo comercializados) están orientados a población vulnerable (infantes, adultos mayores y personas con capacidades cognitivas limitadas) y en muchos casos se ofrecen como opciones terapéuticas, y todas las opciones terapéuticas están (o deberían estar) fuertemente reguladas tanto en su ejercicio y uso, como en el proceso de producción, y deben ser cuidadosamente testeadas antes de ofrecerse a la venta al público en general. Un muñeco de peluche debe cumplir con ciertas normativas relativas, por ejemplo, a su producción con materiales no tóxicos; pero si se trata de un robot mascota terapéutico debería, además, estar regulada su introducción en las prácticas terapéuticas no supervisadas por personas autorizadas. Asimismo, dado que estos robots buscan reemplazar la compañía humana de cuidadores o amigos (en el caso de aquellos robots que se proponen como compañía para personas que están solas) o parejas (en el caso de los robots sexuales), también se requiere, en mi opinión, algún tipo de estudio acerca del impacto de la introducción de este tipo de artefactos en la calidad de las relaciones interpersonales. Ya está estudiado el efecto de las pantallas en el desarrollo cognitivo y afectivo de los infantes, de tal manera que muchas sociedades de pediatría sugieren limitar el tiempo de exposición en niños y evitar a exposición a pantallas en menores de 2 años.[17] Las redes sociales han contribuido con sus algoritmos de recomendación de *feeds* a ampliar la brecha entre personas con opiniones diferentes, al punto que uno termina leyendo posteos de quienes confirman lo que uno piensa, escuchando la música que uno ya conoce, encontrando en Google lo que uno ya buscó, etc. Como modelo de negocios puede ser muy exitoso, pero como modelo del tipo de relaciones interpersonales que consideramos valiosas, la cuestión queda por ser examinada.[18]

---

17 Véase: Levanta la Cabeza. "La Asociación Española de Pediatría pide restringir el uso de pantallas entre los más pequeños». *Compromiso Atresmedia*, 21 de septiembre de 2023, https://compromiso.atresmedia.com/levanta-la-cabeza/actualidad/asociacion-espanola-pediatria-pide-restringir-uso-pantallas-mas-pequenos_20230921650bffba98383a00012b2f39.html. Accedido el 16 de junio de 2024; también: "Niños y Pantallas | Comunidad SAP". *Comunidad SAP | El sitio para la comunidad de la Sociedad Argentina de Pediatría*, http://comunidad.sap.org.ar/index.php/2017/07/31/ninos-y-pantallas/. Accedido el 16 de junio de 2024; y por último: "Cuánto tiempo puede o debe un niño estar frente a una pantalla". *HealthyChildren.org*, www.healthychildren.org/Spanish/family-life/Media/Paginas/where-we-stand-tv-viewing-time.aspx. Accedido el 16 de junio de 2024.

18 Cada vez más voces se levantan en dirección a limitar la exposición de las personas, sobre todo infantes y adolescentes, a las redes sociales. Véase INFOBAE, "Nueva York declaró a las redes sociales como una amenaza para la salud mental de los menores". *infobae*, 27 de enero de 2024, www.infobae.com/estados-unidos/2024/01/27/nueva-york-declaro-a-las-redes-sociales-como-una-amenaza-para-la-salud-mental-de-los-menores. Accedido el 16 de junio de 2024.

¿Es deseable tener un amigo robot que se acomode a nuestros deseos, nos sugiera cosas que siempre nos gustan y nos acaricie como queremos? ¿O es preferible tener como amigo a alguien que a veces nos contradice, que nos expone otro punto de vista, que nos ofrece una caricia inesperada? ¿Cuál de estos dos tipos de compañía hará nuestra vida más desafiante e interesante, aunque, tal vez, no más cómoda? Yo creo que es preferible el segundo tipo de relación interpersonal, pero más allá de las preferencias personales, hay que evaluar el impacto psicológico y social que estas tecnologías tendrían.

En el caso de aquellos sistemas con IA que no están desarrollados para convivir como otras personas entre nosotros, sino que están orientados a tareas específicas, creo preferible adoptar la estrategia de diseño para su evaluación y, por lo tanto, para pensar sus formas de regulación. Seguramente, esta adoptará formas muy diversas, teniendo en cuenta la tarea que el sistema esté diseñado para realizar. Hay muchas actividades humanas en las que la introducción de tecnología supone una serie de regulaciones propias del ámbito en cuestión, tal como es el ámbito de la medicina. No me voy a detener en estos casos. Me interesa destacar la necesidad de regulación para las tecnologías con IA que se introducen en ámbitos de la vida cotidiana y de políticas públicas. Hoy es relativamente fácil y barato crear una *App* con IA para realizar una amplia variedad de tareas, y estos desarrollos son introducidos al mercado (*i.e.* a las sociedades humanas) sin regulación alguna. Y, como bien sabemos, los "términos y condiciones" son ilegibles e incomprensibles para el usuario típico; nadie sabe exactamente a qué se somete cuando los acepta.[19] Pero muchos de estos desarrollos exacerban sesgos y desigualdades, como señalé en el apartado anterior, y generan situaciones conflictivas en varios órdenes de la vida humana. Por ejemplo, recientemente se dio a conocer en la Argentina el caso de un joven que había desarrollado una *App* para "desnudar" gente; es decir, para generar una imagen de una persona desnuda a partir de una foto de esa persona vestida.[20] Obviamente las personas "desnudadas" por la *App*, cuyas fotos circularon en redes, no tenían ninguna legislación que las protegiera, dado que eran fotos generadas por IA, o sea, ficticias (la regulación vigente impide la circulación de fotos no autorizadas de las personas, no impide la difusión de imágenes de esas mismas personas que parecen fotos, pero que fueron producidas artificialmente). Los límites pueden ser difíciles de establecer, pero está claro que este desarrollo no regulado generó perjuicios a personas (obviamente, fueron mujeres las perjudicadas; siempre el

---

19 Una de las cuestiones sobre las que más se ha trabajado en el ámbito de la ética y regulación de la IA es la de la privacidad y uso de los datos. Dado que esta cuestión no está directamente relacionada con el tema que abordo aquí, es decir, con nuestras formas de interacción con estos desarrollos una vez producidos y ofrecidos al público, no me voy a ocupar del tema de los datos, aunque resulta un tema central para el diseño de regulaciones apropiadas.

20 Segulin, Mariana. "Creó imágenes de sus compañeras sin ropa con inteligencia artificial: ¿hay delito?" Últimas Noticias de Argentina y del Mundo | Todo Noticias, 27 de septiembre de 2023, https://tn.com.ar/tecno/2023/09/27/creo-imagenes-de-sus-companeras-sin-ropa-con-inteligencia-artificial-hay-delito/. Accedido el 16 de junio de 2024.

más vulnerable es el más perjudicado). Si, por ejemplo, fuera obligatorio para todo sistema que genera imágenes por IA incluir de una manera inviolable (aunque el concepto mismo de inviolable en la era digital es en sí mismo complejo) una especie de marca de agua, algún tipo de identificación visible de la imagen como ficticia, al menos, las personas perjudicadas podrían fácilmente argüir que no son ellas las que están en esa fotografía, sino que es una fotografía trucada. El daño que una imagen puede hacer en la honorabilidad de una persona, sabemos, es muy grande. No parece razonable que sea posible apropiarse de la imagen de alguien y alterarla sin su consentimiento. Pero no hay (en Argentina, y creo que en ningún lado) regulaciones al respecto.

Otro caso conflictivo con la IA generativa es, obviamente, el de los derechos de autor. Podemos pedir a un sistema con IA que genere un cuadro "a lo Van Gogh". Aun cuando no lo usemos para engañar a nadie, ¿quién puede comercializarlo? ¿El usuario que le indicó a la IA que haga un cuadro de un unicornio a lo Van Gogh? Al fin y al cabo, fue quien tuvo la idea… ¿Y si el artista está vivo y sigue pintando?, ¿quién puede comercializar la imagen? Al fin y al cabo, tal vez a ese artista, en el futuro se le ocurra pintar un cuadro similar…. Una vez más, si no se exige algún tipo de marca de agua que permita distinguir imágenes (o textos) generados por IA de los generados por seres humanos, estamos ante graves problemas que alteran nuestras formas de convivencia actual.[21]

Sé que es difícil pensar el tema regulatorio y es difícil, entre otras razones, porque la tecnología se está desarrollando a una velocidad mucho más grande de lo que pueden desarrollarse y consensuarse nuestros sistemas legales. Sin embargo, cuando es posible avizorar efectos adversos de manera tan clara como en estos casos, tal vez la mejor estrategia sea la precautoria (es decir, "todo está prohibido hasta que se pruebe que no tiene efectos adversos") e invertir la carga de la prueba: es quien ofrece el desarrollo quien debería probar que no tiene impacto (efectos secundarios) negativos, antes de ofrecer al mercado su desarrollo. Tal como ocurre con la industria farmacéutica: los estudios preclínicos y clínicos para lanzar una nueva droga al mercado corren por cuenta de la empresa que ofrece el medicamento o vacuna, y hay una autoridad regulatoria (como la FDA) que dados estos estudios ya realizados analiza la viabilidad de la introducción de ese producto en el mundo humano.

Por otra parte, si un desarrollo tecnológico resulta tan peligroso *prima facie*, que se estima que sus efectos pueden generar un efecto adverso masivo, siempre es posible limitar el desarrollo en cuestión, tal como ocurrió con la energía

---

21 Véase el manifiesto de los artistas reunidos en el colectivo ArteesEtica: "Arte es Ética - Artivismo y Autorías de Habla Hispana". Arte es Ética - Artivismo y Autorías de Habla Hispana, https://arteesetica.org/. Accedido el 16 de junio de 2024.

atómica, la edición genética y la clonación humana. Si bien no parece ser este el caso con los desarrollos de IA, no es algo que haya que descartar tan rápidamente. Por un lado, varios desarrolladores han realizado llamados públicos para frenar el desarrollo de la IA dado que se está visualizando el peligro del desarrollo exponencial no controlado de estas tecnologías. Pero más allá del escenario postapocalíptico de computadoras tomando el poder, los efectos de las AI actualmente existentes que generan más intolerancia, más segregación y más desigualdades en las sociedades humanas y que alteran nuestras formas humanas de interacción y valoración, deberían ser considerados lo suficientemente importantes como para prender una luz roja. Y esto me lleva a mi segundo punto.

Las personas que nos dedicamos a las humanidades y a las ciencias sociales somos quienes más hemos reflexionado y estudiado las prácticas humanas, nuestras formas de interacción y de organización, nuestras formas de autocomprensión y de organización social y política, somos quienes hemos reflexionado sobre el significado de la vida y sobre los valores y normas que rigen las sociedades humanas.[22] Es por eso que creo que la perspectiva humanística debe ser parte de las discusiones y reflexiones alrededor del presente y futuro de la IA. No porque tengamos respuesta a todos los interrogantes que esta tecnología nos plantea, sino porque ya nos hemos formulado las preguntas que estos desarrollos nos plantean. No es la primera vez que una tecnología produce una revolución en el mundo humano; pensemos en la máquina a vapor, la escritura o la imprenta. Estudiar estos casos históricos y los desafíos que plantearon, puede iluminar nuestro presente. No es la primera vez que identificamos conductas discriminatorias y juicios sesgados. Las soluciones que en otras ocasiones se adoptaron pueden servirnos para repensar los desarrollos actuales. Mirar las formas de organización real (que nos provee la antropología) o posible (que nos provee la literatura, el cine y la filosofía política) de las sociedades humanas nos permite replantear los problemas que hoy enfrentamos, que tratamos de resolver con tecnología, pero que también podemos resolver de otras formas. Una vez escuché argumentar que era necesario desarrollar *Apps* de apoyo psicológico porque cada vez hay más gente con trastornos de ansiedad, fobias o depresión. A lo que respondí: "Mejor invitar a esa gente a una clase de Tai Chi o de zumba a la plaza con otra gente!" Si el problema es la soledad, un robot no lo va a solucionar. No es cierto que los problemas que genera la tecnología se resuelven con más tecnología. A veces es tomar distancia, ver dónde estamos parados y a dónde

---

22 Asimismo, los diversos tipos de estudios sugeridos en el párrafo anterior corresponden a disciplinas sociales y humanísticas, por ejemplo, una evaluación de diversos tipos de terapias psicológicas, incluyendo explícitamente las que quedan en manos de sistemas con IA, el impacto psicológico (afectivo y cognitivo) de las horas ante pantallas de personas de diferentes edades, el impacto educativo en la incorporación de IA en las escuelas, la evaluación sociológica y antropológica del impacto de las redes sociales, el impacto en la personalidad e identidad personal producida por la adopción de avatares en entornos virtuales y el grado comparativo de responsabilidad asociado a las acciones que realizamos en entornos reales versus entornos digitales, por mencionar solo algunos casos.

estamos yendo, y preguntarnos a dónde queremos ir. Y para responder esta pregunta nada mejor que imaginar escenarios alternativos posibles. Para eso están las humanidades y las artes: las alteridades disruptivas, no las que se acomodan a nuestros deseos para mantenernos atados.

■ ■ ■ ■ ■

Referencias

"Arte es Ética - Artivismo y Autorías de Habla Hispana". *Arte es Ética - Artivismo y Autorías de Habla Hispana*, https://arteesetica.org/. Accedido el 16 de junio de 2024.

INFOBAE. "Nueva York declaró a las redes sociales como una amenaza para la salud mental de los menores". infobae, 27 de enero de 2024, www.infobae.com/estados-unidos/2024/01/27/nueva-york-declaro-a-las-redes-sociales-como-una-amenaza-para-la-salud-mental-de-los-menores. Accedido el 16 de junio de 2024.

Dennett, Daniel. *Brainstorms. Philosophical Essays on Mind and Psychology.* Cambridge: MIT Press, 1981.

Dennett, Daniel. *The Intentional Stance.* Cambridge: MIT Press, 1987.

Dupré, John. *The Disorder of Things. Metaphysical Foundations of the Disunity of Science.* Cambridge: Harvard University Press, 1993.

Fodor, Jerry. *Psicosemántica*, Madrid: Técnos, 1994.

Lewis, David. "Psychophysical and Theoretical Identifications", *Australasian Journal of Philosophy*, 50 (1972): 249-58

Levanta la Cabeza. "La Asociación Española de Pediatría pide restringir el uso de pantallas entre los más pequeños". *Compromiso Atresmedia*, 21 de septiembre de 2023, https://compromiso.atresmedia.com/levanta-la-cabeza/actualidad/asociacion-espanola-pediatria-pide-restringir-uso-pantallas-mas-pequenos_20230921650bff-ba98383a00012b2f39.html. Accedido el 16 de junio de 2024.

Marr, David. *Vision. A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco: W. H. Freeman, 1982.

"Neurorobotics Research Laboratory (NRL) - Myon." Neurorobotics Research Laboratory, www.neurorobotik.de/robots/myon_en.php. Accedido el 16 de junio de 2024.

"Niños y Pantallas. Comunidad SAP". *Comunidad SAP. El sitio para la comunidad de la Sociedad Argentina de Pediatría,* http://comunidad.sap.org.ar/index.php/2017/07/31/ninos-y-pantallas/. Accedido el 16 de junio de 2024.

"PARO Therapeutic Robot". *PARO Therapeutic Robot*, www.parorobots.com. Accedido el 16 de junio de 2024

Pérez, Diana & Gomila, Antoni. *Social cognition and the second person in human interaction.* London: Routledge, 2022.

Putnam, Hilary. "La naturaleza de los estados mentales". *Cuadernos de Crítica*, vol. 15 México: UNAM, 1981.

Russell, Stuart & Norvig, Peter. *Inteligencia artificial. Un enfoque moderno.* México: Pearson Prentice Hall, 2008.

Segulin, Mariana. "Creó imágenes de sus compañeras sin ropa con inteligencia artificial: ¿hay delito?" Últimas Noticias de Argentina y del Mundo | Todo Noticias, 27 de septiembre de 2023, https://tn.com.ar/tecno/2023/09/27/creo-imagenes-de-sus-companeras-sin-ropa-con-inteligencia-artificial-hay-delito/. Accedido el 16 de junio de 2024.

Turing, Alan. "Computing Machinery and Intelligence", *Mind*, 1950, 49 (1950): 433-460.

■ ■ ■ ■ ■