Artículo de reflexión

Cuestiones de Filosofía
ISSN: 0123-5095
E-ISSN: 2389-9441
Vol. 4 – Nº 22
Enero - junio, año 2018
pp. 141-170

Inteligencia artificial y ética de la responsabilidad

Artificial intelligence and the ethic of responsibility

Antonio Luis Terrones Rodríguez[1]
Pontificia Universidad Católica del Ecuador

Recepción: 23 de febrero del 2017
Evaluación: 14 de mayo del 2018
Aceptación: 3 de junio del 2018

[1] Licenciado en Filosofía por la Universidad de Murcia, Máster en "Ética y Democracia" por la Universitat de València y actualmente doctorando sobre los desafíos éticos que plantea la Inteligencia Artificial en el futuro. Facultad de Ciencias Filosófico-Teológicas de la Pontificia Universidad Católica del Ecuador, Av. 12 de octubre y Roca, Quito, Ecuador.
Correo electrónico: antonioluis.terrones@gmail.com

Terrones, A. (2018). Inteligencia artificial y ética de la responsabilidad.
Cuestiones de Filosofía, 4 (22), 141-170.
doi: https://doi.org/10.19053/01235095.v4.n22.2018.8311

Resumen

La Inteligencia Artificial (IA) ha supuesto un gran avance para la humanidad en diversos campos; sin embargo, eso no implica que su actividad esté exenta de reflexión ética. La humanidad está enfrentado, y va a enfrentar en el futuro, numerosos desafíos que van a obligar a elaborar nuevas ideas para poder vivir a la altura de los tiempos. Entre esos desafíos encontramos el laboral y económico, el de mejoramiento humano, el militar y de seguridad y el político y jurídico, entre otros. Así pues, una vez considerados los desafíos en el terreno de la IA, una referencia ética que puede servir para enfrentar dichos desafíos, puede ser el principio de responsabilidad de Hans Jonas. La lectura de este aporte ético puede facilitarnos unas primeras coordenadas para la orientación en medio de un horizonte de posibilidades nuevo para la humanidad y, adicionalmente, servir como punto de partida en el compromiso que deben asumir los diferentes saberes implicados en este novedoso campo.

Palabras clave: Inteligencia artificial, ética, responsabilidad, riegos, tecnología, desafíos.

Abstract

The artificial intelligence (IA) has set a great advance for humanity in several fields, however, that does not mean that its activity is excluded from the ethical reflection. Humanity is dealing, and it is going to deal in the future, with a number of challenges that will force us to create new ideas to be able to live at the demand of the ages. Among these challenges, we found: labor and economical; human improvement, military and security; and the political and juridical; among others. Then, once the challenges of the IA field are considered, one ethical reference that can serve to deal with those challenges can be the *The Imperative of Responsibility* of Hans Jonas. The reading of this ethical proposal can offer us the first coordinates to locate ourselves within a new horizon of possibilities for the humanity and serve as a starting point in the commitment that the different and involved knowledges should assume in this new field.

Keywords: Artificial intelligence, ethics, responsibility, risks, technology, challenges.

Intelligence artificielle et éthique de la responsabilité

Résumé

L'Intelligence Artificielle (IA) a constitué une grande avancée pour l'humanité dans divers domaines, mais cela n'implique pas que son activité soit exempte de réflexion éthique. L'humanité fait face, et va faire face dans le futur, à de nombreux défis qui vont nous obliger à développer de nouvelles idées pour pouvoir vivre à la hauteur de l'époque. Parmi ces défis, nous trouvons le professionnel et économique, le perfectionnement humain, le militaire et sécuritaire, et le politique et juridique, entre autres. Ainsi, une fois qu'on a considéré ces défis dans le terrain de l'IA, une référence éthique pouvant aider à les affronter pourrait être *Le principe de responsabilité* de Hans Jonas. La lecture de cette contribution éthique peut nous procurer les premières coordonnées pour l'orientation au milieu d'un horizon de possibilités qui est nouveau pour l'humanité, et servir ainsi de point de départ dans l'engagement que doivent assumer les différentes disciplines impliquées dans ce domaine novateur.

Mots-clés: Intelligence artificielle, éthique, responsabilité, risques, technologie, défis.

Inteligência artificial e ética da responsabilidade

Resumo

A Inteligência Artificial (IA) representa um grande avanço à humanidade em diversos campos. Não obstante, isso não implica que sua atividade esteja isenta de reflexão ética. A humanidade enfrenta, e vai encarar no futuro, numerosos desafios, os quais vão lhe obrigar elaborar novas ideias para poder viver nesses tempos. Entre tais desafios, encontramos, dentre outros, o laboral e econômico, o de melhoramento humano, o militar e de segurança e o político-jurídico. Assim, uma vez considerados os desafios no terreno da IA, uma referência ética que pode servir para enfrentar esses desafios pode ser o *Princípio de responsabilidade*, de Hans Jonas. A leitura deste aporte ético pode nos facilitar algumas primeiras coordenadas para a orientação em meio a um horizonte de possibilidades novas para a humanidade e servir como ponto de partida no compromisso que devem assumir os diferentes saberes implicados nesse novo campo.

Palavras-chave: Inteligência artificial, ética, responsabilidade, riscos, tecnologia, desafios.

Introducción

> *"Se trata del mandato de la cautela, en vista del carácter revolucionario que adopta la mecánica de la elección de alternativas bajo el signo de la tecnología, con su inherente «ir a por todas», tan ajeno a la evolución" (Jonas, 1995, p. 72).*

El acelerado desarrollo de la inteligencia artificial ha sido el detonante de profundas transformaciones en numerosos ámbitos de nuestra vida. El poder tecnológico interviene nuestras vidas y las altera debido a su gran fuerza de expansión. La competición entre la inteligencia humana y la inteligencia artificial está servida para las próximas décadas. Como humanidad aún no hemos logrado comprender todo lo que está en juego, las importantes implicaciones que tiene el desarrollo de la IA y el riesgo que eso supone. No se trata de hacer un canto a la tecnofobia, sino más bien de reflexionar sobre sus efectos en nuestras vidas.

Tampoco se trata de dejarnos convencer por ciertas afirmaciones sobre la lógica especial de la tecnología, como si ésta estuviera por encima del bien y del mal, exenta de toda reflexión moral. Se trata de conocer cuáles son los avances tecnológicos más relevantes en el campo de la IA, y ver qué implicaciones éticas tienen en algunos ámbitos de nuestra vida, como el profesional, el médico o el militar. Solo si sabemos a qué nos enfrentamos, podremos imaginar alternativas. Es importante asumir responsabilidad desde una heurística del temor, teniendo cautela, no teniendo excesiva confianza, cegados por un tecnocentrismo que puede llegar a esconder un dogmatismo cientificista. El sonambulismo tecnológico nos puede conducir a terrenos pantanosos de los que difícilmente podremos salir si no tomamos conciencia a tiempo. Pero para una primera toma de conciencia es fundamental asumir responsabilidad, una tarea que no es nada fácil, y más en un momento histórico en el que únicamente se nos muestran los brillos de la IA, pero, como todo en nuestra realidad, también tiene un lado, quizás, no tan brillante, sino más bien oscuro.

Concepto y significado de inteligencia artificial

Es una tarea muy difícil la de encontrar una definición de inteligencia artificial (IA) capaz de recoger todas las consideraciones que pueden existir sobre

este término. No obstante, el propósito de este apartado consiste en aclarar, en la medida de lo posible, el concepto y significado de la IA.

Como ocurre en infinidad de ámbitos, en lo que se refiere a la conceptualización, existen muchas definiciones en torno a la IA, cada una desde diferentes enfoques, aunque parece ser que todas tienen algo en común. Existe una idea fundamental en torno a la que giran las diversas propuestas, a saber, la idea de crear y dar forma a programas de ordenador o también a máquinas que sean capaces de desarrollar conductas que serían consideradas inteligentes si las realizara un ser humano. Considero que esta definición es abierta y puede generar consenso, pues la variedad de definiciones facilitadas por algunos expertos en la materia suelen ser cerradas y opuestas entre sí, por lo que, al menos en este caso, es más prudente no cerrarse. Esta definición se encuentra estrechamente relacionada con la propuesta por John McCarthy en 1955. No obstante, esa definición parte de un enfoque que aparentemente puede presentar algunos defectos, ya que hay que tener en cuenta que fue propuesta a mitad del siglo XX y aún la IA no se había desarrollado lo suficiente.

Sin embargo, lo expuesto en el párrafo anterior no niega la posibilidad de que puedan existir consensos en torno a ciertos marcadores de la inteligencia en numerosos contextos concretos. La duda surge a la hora de intentar aplicar esos marcadores a la máquina. Por ejemplo, si tenemos en cuenta la ardua tarea realizada por los escribas en el Antiguo Egipto con los manuscritos para reproducir textos con el fin de transmitir los conocimientos y la comparamos con una imprenta de libros de textos escolares de hoy en día, la máquina sería "más inteligente" porque es más rápida reproduciendo textos que un ser humano; en este caso, se ha tenido en cuenta el marcador de la velocidad. No obstante, como se puede comprobar, el marcador de la velocidad no es un indicador válido para considerar a una máquina más inteligente que a un ser humano.

Partir de considerar las capacidades humanas como un criterio válido para valorar la IA puede suponer un problema. Una máquina puede llegar a realizar una tarea en milisegundos, y una persona no sería capaz de realizar una tarea parecida en tan poco tiempo; por eso, podríamos pensar que la máquina parece dar muestras de inteligencia. Episodios de ese tipo van a ocurrir en las próximas décadas en cientos de ámbitos, incluso en muchos ya se han producido, por lo que utilizar el método comparativo entre la inteligencia humana y la IA es algo que puede conducir al absurdo, ya que la inteligencia humana

siempre saldría perdiendo. Esto nos muestra, una vez más, que la tarea de aportar una definición concreta de la IA no es nada fácil.

Considerado lo anterior, sería importante esclarecer algunas cuestiones que han podido ir surgiendo. En medio del abanico de posibilidades de dar una definición de la IA podría suponerse que un sistema de IA debe poseer algunas características que consideraríamos como básicas en todo sistema de IA. La capacidad para aprender es una de esas características del diseño básico de un sistema para alcanzar la inteligencia artificial, lo que en el campo computacional se denomina *machine learning*. Otra característica básica es manejar la incertidumbre y la información probable, así como formar conceptos a partir de representaciones combinatorias que se usan en el razonamiento lógico e intuitivo y, adicionalmente, poseer esos caracteres básicos. En palabras de Nilsson:

> La inteligencia artificial (IA), en una definición amplia y un tanto circular, tiene por objeto el estudio del comportamiento inteligente en las máquinas. A su vez, el comportamiento inteligente supone percibir, razonar, aprender, comunicarse y actuar en entornos completos. Una de las metas a largo plazo de la IA es el desarrollo de máquinas que puedan hacer todas estas cosas igual, o quizá incluso mejor, que los humanos. Otra meta de la IA es llegar a comprender este tipo de comportamiento, sea en las máquinas, en los humanos o en otros animales (Nilsson, 2001, p. 1).

En definitiva, la IA pretende desarrollar comportamientos en las máquinas que sean inteligentes en entornos complejos. Como se puede constatar, esta definición de Nilsson encaja perfectamente con el espíritu original de McCarthy, que fue mencionado anteriormente, pues el objetivo fundamental es averiguar si la inteligencia que puede llegar a alcanzar una máquina llegaría a ser similar a la del humano.

Pasado de la inteligencia artificial

Para conocer de mejor manera qué es la IA y en qué estado se encuentra en la actualidad, es importante tener en cuenta su pasado histórico y para ello es necesario remontarse a sus orígenes, pues esos orígenes nos van a dar algunas pistas para comprender mejor este asombroso campo.

Antes de poder hablar de la IA como ámbito de investigación material, resulta pertinente destacar el espíritu inicial del campo de la computación, que estaría encarnado en la figura del lógico y matemático británico Alan Turing,

que pasaría a la historia por su amplia contribución a dicho campo. Hay ocasiones en las que algunos autores se refieren a Turing como el padre fundador de la IA, cuando en realidad eso no es cierto, ya que Turing no trabajó en ningún programa concreto de IA como tal. Sin embargo, lo que sí hizo, como ya se ha dicho, fue impulsar la reflexión filosófica sobre los aspectos pensantes de las máquinas, cosa muy diferente a desarrollar un verdadero programa de IA.

En el ámbito de la IA una de sus principales contribuciones fue el famoso Test de Turing, publicado en 1950 en la revista *Mind*. Este test nace como un método para comprobar y determinar si una máquina puede pensar. Se introduce en el terreno de probar la habilidad de una máquina, para comparar el conocimiento de ésta con el de un humano, y saber si es similar o se distingue. El británico realiza una verdadera reflexión filosófica sobre los horizontes pensantes de una máquina. El Test de Turing abrió un abanico de posibilidades para el desarrollo del campo de la IA, facilitando que la programación sea cada día más compleja y sofisticada.

Otro momento importante para ser tenido en cuenta como punto de partida tiene que ver con lo acontecido en el Dartmouth Summer Research Project on Artificial Intelligence en el Dartmouth College. Este encuentro se desarrolló en el verano de 1956 y simboliza el germen de la IA como campo de investigación. Este evento contó con la presencia de importantes investigadores, que luego pasarían a considerarse como los pioneros de la IA, entre los que encontramos al impulsor del evento, John McCarthy, y también a Ray Solomonoff, Herbert A. Simon, Trenchard More, Allen Newell, Oliver Selfridge y Arthur Samuel, entre otros. Todos los investigadores allí presentes compartían el mismo interés por las teorías autómatas, el estudio de la inteligencia, las redes neuronales, etc. En definitiva, la principal inquietud de estas mentes brillantes era la inteligencia de los computadores.

Según la impresión de John McCarthy, ese encuentro no fue un éxito, debido a que cada mente brillante no intercambió verdaderamente ideas con las otras mentes. El propio McCarthy lo expresaba así: "Para mí fue una gran frustración (...) Tampoco hubo, por lo que yo pude ver, ningún intercambio auténtico de ideas" (McCorduck, 1991, pp. 95-96). Sin embargo, el encuentro de Dartmouth serviría como punto de partida para agrupar a muchos investigadores en torno a un campo de investigación prometedor en ese momento. Tanto es así que la Fundación Rockefeller fue la que se encargó de

financiar el encuentro impulsado por McCarthy, con el objetivo de estudiar el desarrollo de un nuevo lenguaje de programación que permitiera dotar de inteligencia a las máquinas. El lenguaje que germinó en ese encuentro, y que aparecería más tarde, sería el LISP (List Processing Language).

Ese verano de 1956 en el Dartmouth College sirvió para impulsar una comunidad científica con metas prometedoras, e incluso con un importante sentido de identidad que pasaría a la historia. También sirvió como impulso para el establecimiento de laboratorios de IA en varias universidades, concretamente en Stanford, bajo supervisión de McCarthy; en el MIT, con Marvin Minsky; en Carnegie Mellon, con Neweel y Simon; y bajo la supervisión de Donald Michie, en Edimburgo. Tanta importancia tuvo ese acontecimiento que esos laboratorios de ideas se mantienen hasta hoy día como piezas clave para la investigación en el campo de la IA. Además, ese encuentro tuvo una gran importancia en la historia de la computación porque significó el impulso para la invención del Logic Theorist, el primer programa de IA, creado por Newell y Simon.

Finalmente, aunque el espíritu inicial de la conferencia fuera apasionante en lo que respecta a los temas que se plantearon y a la posibilidad de encontrar numerosos avances, todo parece indicar que finalmente no se lograron muchas cosas, ya que ni siquiera se publicó el informe final que se había prometido desde un principio. No obstante, ese momento sirvió para acuñar una expresión que pasaría a la historia de la computación y que despertaría gran interés, la inteligencia artificial.

Presente y futuro de la IA

En la actualidad, la IA supera a la inteligencia humana en muchos ámbitos; por ejemplo, en el de los videojuegos, ya que existen ordenadores que se dedican a los juegos y son una clara muestra de victorias contra verdaderos expertos humanos en la materia. Podemos encontrar evidencias en las últimas décadas, como Backgammon, de Hans Berliner, Scrabble, Jeopardy!, Deep Blue, etc. En el caso concreto del intelecto sintético Deep Blue, es importante destacar que no existe una verdadera autonomía de la máquina, ya que cuenta con un antecedente de programación donde está presente la mano humana, aunque eso no implica que este ejemplo no haya supuesto un importante avance en el campo de la IA, como bien señala Garry Kasparov (2017). Estos ejemplos representan importantes logros en el campo de la IA.

Sin embargo, habrá personas que no se sientan tan impresionadas ante tales casos, pues nuestra capacidad de impresión va variando con el tiempo y depende del progreso tecnológico del momento.

La IA está presente en diversos campos en la actualidad, como en el reconocimiento óptico, utilizado para la clasificación de correos o la digitalización de antiguos documentos que no queremos que se pierdan. También en la traducción automática, por ejemplo, en el caso de Google, que aunque es imperfecta, ha supuesto un claro desarrollo. El reconocimiento facial se ha introducido en numerosos pasos fronterizos de Europa, como el aeropuerto de Ámsterdam-Schiphol, además de EE.UU. o Australia.

En el campo militar, la IA ha aportado numerosos avances; por ejemplo, en el despliegue a gran escala de robots que trabajan desactivando bombas y de drones autónomos letales. En realidad, la industria militar es una de las grandes beneficiadas en la actualidad del avance de la IA, dado que las naciones más desarrolladas mantienen una carrera militar en la que la IA es su eje principal.

Internet es también otro campo que se ha visto claramente beneficiado del desarrollo de la IA; por ejemplo, en software que se dedican al rastreo y vigilancia de correos electrónicos o los relacionados con las cuestiones de preferencia de compra, como Amazon. Las transacciones económicas que realizamos con nuestras tarjetas de créditos cuando hacemos una compra por Internet también están redirigidas por IA. No obstante, si de lo que estamos hablando es de Internet, el motor de búsqueda de Google es el mejor ejemplo para mostrar cómo la IA ha impregnado nuestras vidas virtuales.

Otro campo en el que existe una clara presencia de sistemas de IA es el sector financiero. El pionero a la hora de introducir IA en el campo financiero fue el Citibank, a comienzos de los años 80 del siglo XX, y posteriormente el Security Pacific National Bank, en el año 1987. Pero hay que decir que hoy en día las principales compañías inversoras hacen un uso generalizado de la IA. Ese uso, no obstante, es de diversos tipos: va desde simples intercambios financieros a complejas operaciones que se adaptan a las condiciones cambiantes del mercado. Además, existen numerosas aplicaciones que se basan en IA y que son para uso personal en el sector financiero, como Kasisto, Moneystream, Wallet.AI, etc., así como otras utilizadas en el sector en general para conceder préstamos y detener el fraude, como Lending Club, Affirm, Prosper Daily, ZestFinance, entre otras. En consecuencia, la automatización

está muy presente y un claro ejemplo es el Flash Crash de 2010, que, aunque no es un ejemplo de IA muy desarrollado, sí que nos sirve para entender cómo la tecnología está manejando grandes cantidades de dinero y haciendo transacciones económicas que comprometen las vidas de muchas personas.

El estadounidense Jerry Kaplan también nos muestra numerosos ejemplos en su obra *Abstenerse humanos* (2016), sobre todo con la automatización de varios puestos de trabajo. Un ejemplo muy ilustrativo, y de una empresa de la que posiblemente casi todo el mundo haya escuchado hablar, es el de Amazon. También Agrobot, una empresa agrícola de Huelva (España) que se dedica a la recolección de fresa, y que está sustituyendo la mano de obra humana por la maquinaria basada en la IA que desarrolla en su oficina de Oxnard, California. Estos son dos claros ejemplos de cómo la IA ya está inmersa en nuestra vida en diversos ámbitos.

Otro sector en el que la IA tendrá importante presencia es en el sector del automóvil. Google impulsó en el año 2008 el diseño de un vehículo autónomo, como nos señala Martin Ford. Google se empeñó en demostrarle al mundo que podía ser capaz de desarrollar un vehículo autónomo que condujera en mejores condiciones que los propios humanos, y por eso contrató a los mejores ingenieros de las carreras de DARPA, que es la Agencia de Investigación de Proyectos Avanzados de la Defensa. Es cierto que los resultados de las carreras de DARPA fueron mejorando considerablemente, y que ese fue el principal motivo por el que Google se interesó en sus ingenieros. La irrupción del automóvil sin motor puede propiciar la aparición de un nuevo paradigma de movilidad y de interactuación entre humanos y automóviles, por lo que estaríamos hablando de algo muy importante y que requiere de una profunda reflexión ética.

En los últimos años, hay un fuerte interés en el desarrollo de la IA que va en dos direcciones: una que tiene que ver con una teoría de la información que sea más sólida para el aprendizaje artificial, y otra hacia el desarrollo del aspecto práctico y comercial de varios sistemas de resolución de problemas concretos y de ámbitos específicos. A pesar de todo, no existe una postura unánime en torno al futuro de la IA. Incluso Nick Bostrom señala esta cuestión:

> Las opiniones de los expertos sobre el futuro de la IA varían enormemente. No hay acuerdo sobre la sucesión temporal de los acontecimientos ni sobre qué formas podría llegar a adoptar la IA. Las predicciones sobre el futuro

> desarrollo de la inteligencia artificial, señaló un estudio reciente, 'son tan firmes como diversas' (Bostrom, 2016, p. 19).

Bostrom aportó los resultados de una encuesta, realizada por el Future of Humanity Institute de la Universidad de Oxford, y del que es director fundador. Esos resultados evidenciaban que no existe un consenso claro en torno al futuro de la IA. Las encuestas giraban en torno a la pregunta de cuándo esperaban los expertos que se iba a alcanzar la inteligencia artificial de nivel humano. Los resultados se reflejan en la siguiente tabla:

¿Cuándo conseguiremos una inteligencia artificial de nivel humano?			
	10%	50%	90%
PT-AI	2023	2048	2080
AGI	2022	2040	2065
EETN	2020	2050	2093
TOP 100	2024	2050	2070
Combinados	2022	2040	2075

En esta tabla se muestran los resultados de cuatro encuestas diferentes, así como la combinación de los resultados. Los participantes de las encuestas se ven reflejados en el documento de Vincent Müller y Bostrom (Müller y Bostrom, 2014). A pesar de que las encuestas son predictivas, Bostrom defiende la postura de que es muy probable que la superinteligencia surja poco después del momento en el que la IA alcance un nivel humano.

Los avances en los diferentes campos que conforman el grueso de la IA no se dan de la misma forma, pues hay ámbitos que dependen a su vez de otros. Por ejemplo, el campo de la robótica no avanza al mismo ritmo que el del aprendizaje maquinal. Boston Dynamics es una empresa de ingeniería y robótica fundada en 1992 por Marc Raibert, exprofesor del MIT –siglas en inglés del Massachusetts Institute of Technology–, que ha experimentado numerosos avances en los últimos 25 años, por lo que el tiempo comprendido para desarrollar ciertos proyectos se podría considerar mayor que el requerido en el ámbito del aprendizaje maquinal para desarrollar sus proyectos. Los investigadores de la IA reconocen el papel relevante del aprendizaje para la inteligencia humana y se preguntan si es posible emular esa forma de aprendizaje en las computadoras. El aprendizaje maquinal tiene como objetivo la creación de programas que permitan la generalización de comportamientos

a partir de ejemplos que le son suministrados, y que por lo tanto generan un patrón de comportamiento.

Intelectuales del campo de la IA como Bostrom o Ray Kurzweil centran sus estudios en la proyección de la IA hacia el futuro. El primero habla de "superinteligencia" y el segundo de "singularidad", dedicando cada uno de ellos, incluso, una obra en particular para abordar este aspecto futurible. Estamos pues ante una marcha imparable, no únicamente de los robots, como señala Andrés Ortega (2016), sino de la inteligencia artificial; una marcha imparable que necesita ser pensada desde la filosofía, y concretamente desde la ética, pues va a plantear, en un futuro no tan lejano, numerosos e importantes desafíos. A este respecto, me gustaría rescatar lo expresado por Jorge Enrique Linares:

> En el mundo tecnológico, el individuo se enfrenta a una realidad: por un lado, experimenta la potenciación de la libertad individual mediante la tecnología; pero, por otro lado, percibe y sufre la fragmentación social y el aislamiento, los problemas ecológicos y políticos planetarios ante los cuales denota una creciente incapacidad para actuar solidariamente, para determinar criterios y valores universales, y para superar el relativismo o el escepticismo moral que neutraliza la responsabilidad ética (Linares, 2008, p. 38).

El diseño y producción de la IA sigue una lógica imparable, y también podría decirse que insaciable, lo que implica que nuestras relaciones con el mundo estén cambiando continuamente, ya sea por nuestra relación con las máquinas o por la relación entre las propias máquinas. Los productos de IA son cada vez más autónomos, por lo que debemos comenzar a pensar cuál es nuestro papel en el mundo y a re-pensar nuestra relación para-con la máquina.

Los estudios emprendidos por Bostrom lo llevan a pensar que es muy probable que, una vez que la IA alcance los niveles humanos de inteligencia, se produzca una explosión de superinteligencia, lo que implica que los intelectos sintéticos sean autónomos respecto de los programadores, y que puedan constituir y dar forma a su vez a otros intelectos. Por lo tanto, el fenómeno de la superinteligencia invita a una profunda reflexión ética, pues no estamos hablando de un tema baladí, sino de un tema que compromete de una manera importante a la humanidad.

La denominada "cinética" de una explosión de inteligencia por parte de Bostrom (2016), nos muestra cómo podría tener lugar la sincronización y velocidad de despegue en el momento en el que la IA alcanzara niveles

cognitivos humanos. Llegado el momento en el que la IA alcanzara los mismos niveles cognitivos que los humanos, existen diferentes caminos que se pueden divisar en el horizonte. Sin embargo, y atendiendo a la reflexión que suscita este texto, es importante partir de la consideración del fenómeno de la transición, debido a la existencia de una, supuesta, gran probabilidad de explosión de superinteligencia.

"Singularidad" es otro término que se utiliza en el campo de la IA para referirse al sistema supertinteligente que es capaz de perfeccionarse a sí mismo y crear otros sistemas, incluso más inteligentes que él, siguiendo un crecimiento exponencial. El máximo exponente de la singularidad es el estadounidense Raymond Kurzweil, que considera que la singularidad puede perfeccionarse a sí misma, teniendo como horizonte la constitución de todo un universo basado en una entidad global inteligente. El estadounidense afirma que cuando un intelecto sintético supere a la inteligencia humana el progreso será mucho más acelerado. Kurzweil, al igual que Hans Moravec (1988), está convencido de que durante la primera mitad del siglo XXI las máquinas excederán la inteligencia humana. Según este especialista de la IA, el crecimiento de la IA será exponencial, al igual que sostenía anteriormente Yudkowsky:

> Representa la fase casi vertical del crecimiento exponencial que tiene lugar cuando el ritmo es tan extremadamente alto que la tecnología parece expandirse a una velocidad infinita, pese a que, desde la perspectiva matemática, no hay discontinuidad ni ruptura y los ritmos de crecimiento siguen siendo finitos, aunque extraordinariamente grandes. Pero desde nuestro limitado marco actual este evento inminente parece una ruptura aguda y brusca en la continuidad del progreso (Kurzweil, 2017, p. 26).

En este sentido, la singularidad está cerca y va a implicar un cambio de paradigma en varios campos, que Kurzweil menciona en su obra (2017, pp. 27-33). A finales de este siglo se espera que la mayor parte de la inteligencia no sea biológica, pero eso no significa que nos encontremos frente al principio del fin de la inteligencia biológica, ni mucho menos. Su propuesta para evitar nuestra anulación, pues es muy probable que quedemos a merced de las máquinas, es nuestra fusión con la máquina, aunque él habla propiamente de "enlace íntimo" (Kurzweil, 2017, p. 33).

Las propuestas de Bostrom y de Kurzweil tienen mucho en común, mientras que el primero habla de superinteligencia, el segundo habla de singularidad. Ambos pensadores están convencidos de que los intelectos sintéticos

experimentarán un aumento exponencial que llegará durante este siglo, influyendo así en numerosos ámbitos humanos hasta su dominio.

El principio de responsabilidad de Hans Jonas

En el año 1979, Hans Jonas publica *El principio de responsabilidad*. *Ensayo de una ética para la civilización tecnológica*, una obra en la que plantea la necesidad de responder a los retos que la civilización actual nos presenta por medio de la tecnología. Esta obra se publica en un momento en el que la conciencia ecológica comienza a abrirse camino. Aunque esta obra está orientada a la formación de la conciencia ecológica, considero que las ideas que en ella se exponen son de suma valía para aplicarlas también a los desafíos futuros de la IA. El objetivo principal de *El principio de responsabilidad* es despertar la conciencia para asegurar la esencia de la humanidad en el futuro y la supervivencia de la naturaleza, mediante un cambio ético radical con la aplicación del principio de responsabilidad.

El sonambulismo tecnológico (Winner, 2008) nos ha instalado en la creencia de que la tecnología solucionará todos nuestros problemas en el futuro, al igual que nos está facilitando muchas actividades en el presente, por lo que, cuanto más se desarrolle la tecnología, mayor beneficio tendrá la humanidad. Esta es una visión enteramente optimista, con claros tintes utópicos en la tecnociencia promovida por un gran número de científicos, ingenieros, inventores, etc. Los colectivos que defienden este optimismo, vinculado con el utopismo tecnológico, defienden, en cierta medida, una neutralidad axiológica, la ausencia de cualquier implicación moral en las creaciones tecnológicas y científicas.

La problematización de las creaciones anteriormente mencionadas no tiene nada que ver con aspectos morales, sino más bien con aspectos relativos a la utilización y aplicación que caerían en el terreno de lo estrictamente técnico. Son precisamente los efectos comprobados por la tecnología a lo largo del siglo XX los que despiertan en Jonas el interés de rechazar esa supuesta imparcialidad axiológica para despertar el *deber ser.* La producción de conocimiento en el terreno de la tecnociencia tiene un claro compromiso económico con la intención de beneficiar el lucro de ciertos sectores, y no precisamente de garantizar el bien común de la humanidad.

La promesa ilustrada sobre la técnica de promover un mundo mejor, mediante el uso, dominio y sometimiento de la naturaleza, ha tenido graves

consecuencias en el siglo XX. Esa técnica, que llegó para mejorar nuestras vidas, se ha convertido en una clara amenaza para la naturaleza en todas sus dimensiones, tanto interna como externa al hombre. La esencia humana está experimentando importantes transformaciones, lo que pone en juego la capacidad del hombre para pensar el futuro y proponer alternativas.

La producción y diseño tecnocientífico suceden cada vez con más velocidad, los ciclos de revolución tecnológica son cada vez más cortos, lo que invita a una mayor rapidez a la hora de reflexionar sobre las implicaciones que puedan llegar a existir. La lógica de producción tecnocientífica ha conducido al hombre a un cierto nihilismo, que desemboca en inmovilismo, como si no fuera posible cuestionarse nada que tuviera que ver con el progreso tecnológico. El hombre se enfrenta a un escenario sin precedentes, un escenario de novedosos experimentos y sobre los que no se esperan previsibles consecuencias, se encuentra desorientado, no tiene respuestas. Las teorías éticas habidas hasta ahora no sirven para cuestionar el carácter ético del progreso tecnocientífico, según Jonas porque las éticas existentes son antropocéntricas, y no han visto al hombre como objeto de la *techne* transformadora (Jonas, 1995, p. 29).

Las teorías éticas tradicionales se muestran insuficientes, porque no son capaces de dar una respuesta novedosa ante este escenario. La posibilidad del mal ya no recae sobre las relaciones intersubjetivas entre los hombres, se encuentra en otro nivel, se ha extendido hacia toda la biosfera, lo que aumenta y hace más complejas las implicaciones éticas. Por lo tanto, ante este vacío ético (Jonas, 1995, pp. 58-59) es imperioso proponer una nueva ética desde la que pensar la nueva realidad, poniendo en tela de juicio el avance tecnológico.

Jonas advierte que la tecnología contemporánea se ha convertido en un elemento de arrastre que configura el mundo. Junto con otros pensadores como Jacques Ellul (2003, 2004) considera que la técnica responde a un curso y a una lógica autónoma que deriva en una tendencia amenazadora:

> [J]unto a la magnitud y a la ambivalencia, otro rasgo de carácter del síndrome tecnológico que tiene una importancia ética propia: el elemento cuasi-forzoso de su avance, que por así decirlo hipostatiza nuestras propias formas de poder en una especie de fuerza autónoma de la que nosotros, los que la ejercemos, nos volvemos paradójicamente súbditos (Jonas, 2001, pp.38-39).

El poder de la tecnología es inconmensurable e impredecible, las consecuencias de la acción humana pueden implicar un alto porcentaje de riesgo. En

este sentido, para Jonas, de lo que se trata es de partir desde cierto velo de ignorancia para llegar a un saber propio de la ética de una dimensión nueva, una ética vinculada con el saber predictivo para poder acompañar y guiar la vigilancia a que le debe ser sometido el poder de la tecnociencia. Esto es sumamente importante, porque el progreso tecnocientífico, que ha tenido importantes beneficios para nuestras vidas, también ha ido acompañado de forma paralela de graves efectos. La ética a la que se refiere Jonas es la ética de la responsabilidad, entendida en su obra como el principio de responsabilidad, que nada tiene que ver con una esperada reciprocidad en el futuro.

En este sentido, no se justifica poner en peligro la existencia de la humanidad, ya sea en el presente o en el futuro, bajo cualquier pretexto cientificista, o como Jonas señala, "nunca es lícito apostar, en las apuestas de la acción, la existencia o la esencia del hombre en su totalidad" (1995, p. 80). El imperativo de preservación de la propia existencia de la humanidad, en términos colectivos, prima en la ética jonasiana. Pero no solo se trata de preservar la existencia de la humanidad en términos colectivos, sino también la esencia del hombre, pues los experimentos tecnocientíficos la ponen en tela de juicio. El abanico de posibilidades que nos presenta el poder inconmensurable de la tecnociencia debe ser considerado desde la responsabilidad con miras al futuro, de no abusar de la capacidad que podemos llegar a tener, de no perjudicarnos bajo el precepto de mejorarnos.

El filósofo alemán se refiere a aquella responsabilidad que va más allá de los actos y sus consecuencias directas, es decir, *ex post facto*, necesariamente orientada hacia la ampliación de su horizonte hacia el futuro. Una responsabilidad que tenga que ver con una potestad justificada que encuentra su justificación en el compromiso, en algo que se le confía la garantía de protección. Es un concepto moral de responsabilidad que tienen en cuenta los fines. En una mesa redonda, celebrada en un simposio en el Hotel Schloss Fuschl de Austria en 1981, Jonas pronunció las siguientes palabras:

> Ahora bien, he dedicado algún esfuerzo para distinguir entre dos conceptos completamente distintos de responsabilidad; el concepto puramente formal, por así decirlo jurídico de la responsabilidad: que cada uno es responsable de lo que hace y se le puede responsabilizar de lo que ha hecho si se le tiene a mano. Esto mismo no es un principio de la acción moral, sino sólo de la responsabilización moral posterior por lo hecho. Cuando el sujeto de la responsabilización moral ya no está ahí, no hay por así decirlo nada que hacer. Pero hay que distinguir de esto un concepto completamente distinto de la

> responsabilidad, el que acabo de ilustrar en particular en la relación padre hijo, y es la responsabilidad por lo que hay que hacer: no pues la responsabilidad por los actos cometidos, sino estar obligado por la responsabilidad a hacer algo, porque se es responsable de una cosa. Pero se es responsable de la cosa porque la cosas están en el ámbito del propio poder, es decir, depende de la propia acción (…) la humanidad, y por tanto cada miembro de la humanidad, cada individuo concreto, tiene de hecho una obligación trascendente o metafísica de que también en el futuro haya en la tierra hombres, encarnaciones de este género humano –y en condiciones de existir–, que aún permitan hacer realidad la idea de ser humano (1997, p. 188).

Un cientificismo cerrado y ciego ante sus consecuencias, una fe ciega en un progreso indefinido, solo puede ser cuestionado con el poder de la imaginación, dándole más importancia al *malum*, pues para pensar las consecuencias catastróficas de algo, para el surgimiento del sentido de responsabilidad, es más fácil partir del lado negativo que del lado positivo, o del *bonum*. Como señala el alemán: "nos resulta más fácil el conocimiento del malum que el conocimiento del bonum; el primero es un conocimiento más evidente, más apremiante, está menos expuesto a la diversidad de criterios y, sobre todo, no es algo buscado" (Jonas, 1995, p. 65). El temor es planteado desde una anticipación cognitiva, pues cumple una función *heurística* que nos sirve para descubrir el bien y buscar mecanismos para su conservación. El poder cognitivo recae sobre la imaginación y sobre el sentimiento, porque nos anticipamos para conocer y a la misma vez nos conmovemos, algo fundamental para asumir responsabilidad.

La *heurística del temor* es el primer paso del planteamiento del filósofo alemán, y se encuentra en un estadio anterior a la "ética orientada al futuro", que promueve la representación de los efectos que se plantean en el futuro remoto, y se construye a partir de una representación imaginante que nos mueve. Es importante destacar que la propuesta de Jonas tiene una función profética de la catástrofe, pues poseía vastos conocimientos propios de la tradición judaica.

En definitiva, cuando Jonas nos habla de responsabilidad, se refiere a ese comportamiento altruista que no encuentra su justificación en la reciprocidad y lo vincula con el deber. Con un deber para con la existencia y la esencia del hombre, un deber que exige un compromiso real con el futuro de la humanidad. El deber se convierte en el primer comportamiento humano colectivo no solo encaminado a resguardar al hombre, sino también al conjunto de la

naturaleza. Como señala Josep M. Esquirol, a partir de la lectura de Jonas, "tanta responsabilidad como poder técnico" (2011, p. 111).

La responsabilidad ante los desafíos del futuro en el campo de la IA

Una vez expuestas algunas de las ideas planteadas por Jonas en su obra *El principio de responsabilidad*, es importante considerarlas y reinterpretarlas a la luz de nuestro tiempo y dentro del contexto del desarrollo de la IA. Los desafíos que se nos presentan tienen un importante impacto en diferentes ámbitos. En este apartado, me gustaría reflexionar sobre algunos de esos desafíos desde la óptica de Jonas. Entre los desafíos nos encontramos: el profesional y económico, el de mejoramiento humano y el militar y de seguridad, entre otros, aunque aquí solo comentaré esos tres.

a) Profesional y económico

La tecnología está transformando considerablemente el ámbito de las profesiones y es el principal detonante de dicha transformación. Cada vez hay más empresas que dedican sus investigaciones a la creación de una gran variedad de sistemas, máquinas, herramientas, etc., que tienen como finalidad la recopilación y almacenamiento de gran cantidad de conocimiento y habilidades que tradicionalmente han sido características de los seres humanos. Esta transformación que ha impulsado la tecnología, ha tenido principalmente dos impactos, como señalan los ingleses Richard Susskind y Daniel Susskind (2016, p. 109): la automatización y la innovación.

Cuando los ingleses hablan de "automatización", se refieren a aquel fenómeno que se levanta sobre el objetivo de lograr eficacia y ahorrar costes, lo que tradicionalmente conocemos como optimización. Se podría considerar que la automatización ha estado vinculada tradicionalmente con la mejora de aquellos sistemas que solían ser manuales, y eso no es menos cierto. Sin embargo, en la actualidad, cuando se habla de automatización, se hace más hincapié en la incorporación de la tecnología en la profesión, que en la mejora de los aspectos manuales. Aquellos trabajos que se caracterizan por la monotonía y por la rutina, son los que suelen encontrarse en el foco de la automatización, aunque también otros que no son tan monótonos y que requieren de ciertas habilidades persuasivas, como ciertas estrategias de marketing elaboradas por algunas empresas importantes a partir de cuestionarios en la web.

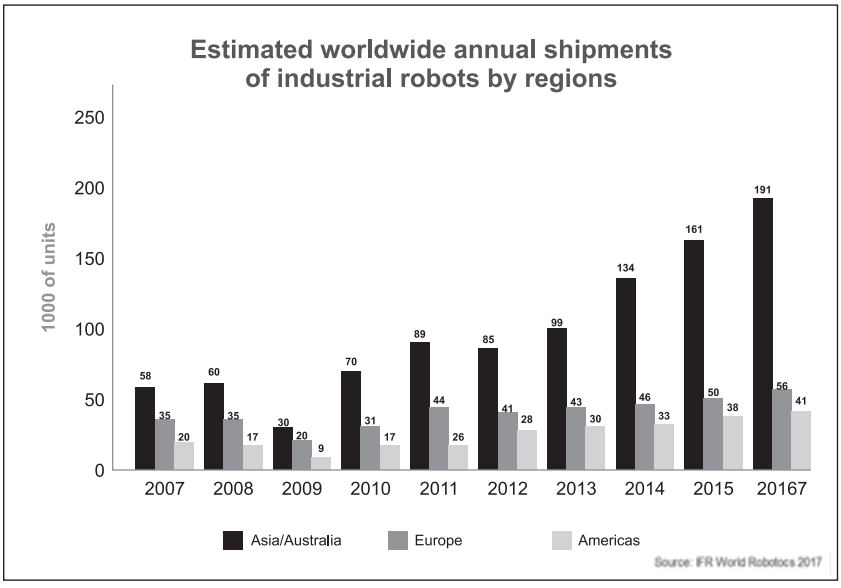
Viremos ahora hacia el otro impacto que ha tenido la tecnología, la innovación. Tiene que ver con aquellos servicios novedosos que se han ido prestando en ciertos espacios, y que no necesariamente tienen que ver con la sustitución de una tarea rutinaria, que tradicionalmente desempeñaba un humano. La innovación tecnológica puede brindar una serie de conocimientos prácticos para abrir el abanico de posibilidades de actuación. Por ejemplo, hay Apps para los Smartphone que nos sugieren un menú según nuestras preferencias de gusto y de disponibilidad económica, sin necesidad de movernos del escritorio de nuestra oficina, como es el caso de Adomicilioya, que contó en el año 2016 con 130.000 descargas, según el diario El Universo[2]. Investigadores en el campo del impacto de la tecnología sobre las profesiones, como son la familia Susskind, se muestran optimistas ante el fenómeno de la innovación, pues aseguran que facilitará la producción de nuevos conocimientos (Susskind y Susskind, 2016, p. 112).

Detrás de todo el diseño de tecnología para diversos campos profesionales, se encuentra la IA. Es muy común, como nuestra familiaridad con los sistemas de IA es creciente conforme pasa el tiempo, que en ocasiones no identifiquemos estos sistemas y que por lo tanto los pasemos por alto. La IA ha penetrado en muchos aspectos de nuestra vida como las computadoras, la banca *on line*, las compras por Internet, etc. El crecimiento exponencial de variados sistemas de IA ha permitido ampliar el espacio de incidencia de la tecnología en las profesiones.

La IA promueve una automatización de las profesiones mucho mayor. Como nos muestra Kaplan (2017, p. 123) hay trabajos que requieren de la utilización de la visión para la identificación de determinados objetos y para la ubicación de esos objetos en un lugar concreto, por ejemplo, en la estantería de un almacén. Tradicionalmente, los empleos que se han visto amenazados por la automatización han sido los más rutinarios; sin embargo, la IA ha extendido el campo de actuación de la automatización hasta campos profesionales que anteriormente no se habían visto afectados por ese fenómeno, lo que sugiere la necesidad de una profunda reflexión sobre este nuevo alcance de la automatización potenciada por la IA.

2 https://www.eluniverso.com/vida-estilo/2016/04/27/nota/5546724/aplicacion-movil-facilita-pedidos-comida-domicilio. Consultado el 31 de enero de 2018.

Los robots se abren camino. La competencia entre los trabajadores y los "falsos trabajadores", en términos de Kaplan (2016), está servida. Los robots ya funcionan a mucha más velocidad que los seres humanos en muchas tareas profesionales, un claro ejemplo es el robot construido por *Industrial Perception*, que lograr mover una caja por segundo, mientras que un humano tarda 6 segundos. Además, el robot no se cansa, ni sufre lesiones, y, al menos por este momento, no tiene derechos laborales, algo que lo hace muy atractivo para muchas empresas. El robot de *Industrial Perception*, ya está suponiendo una clara amenaza para la estabilidad de muchos empleos que son rutinarios. Un dato muy importante es el presentado en un informe de la Federación Internacional de Robótica (2017), en el que se muestra que la venta de robots ha aumentado un 16% en el año 2016, con respecto al año anterior, lo que supone un importante avance si tenemos en cuenta que desde el año 2012 el crecimiento ha sido muy considerable. El gráfico que se muestra a continuación refleja claramente esa tendencia a la alza:



Las reacciones en lo que respecta al futuro de la robótica en el campo profesional son muy variadas. Hay quienes se sitúan en el terreno del optimismo, planteando un argumentario de beneficios gracias al aumento de la automatización. En cambio, en frente, se encuentran los que no son tan optimistas, sino más bien pesimistas, aunque unos más que otros, pero que tienen en común el planteamiento sobre la necesidad de reflexionar filosóficamente

acerca de las implicaciones que podría conllevar un aumento de automatización en el futuro. Esta contraposición de ideas se escenificó de manera notable en la Cumbre de las Ideas, celebrada en Puebla en el año 2016, y que contó con Matt Ridley, Ronald Arkin y James Bessen, del lado de los tecnooptimistas, y con Nick Bostrom, Tyler Cowen y Martin Ford, del lado de los tecnopesimistas.

Desde la óptica de Jonas, podríamos partir de una anticipación cognitiva, mediante la imaginación, para lograr un acercamiento a las catastróficas consecuencias que podría tener un impacto a gran escala de la IA en el campo profesional. Pero es importante destacar que no por imaginar un impacto negativo eso debería implicar una oposición rotunda a la presencia de IA en el campo profesional, ya que, como se ha comentado anteriormente, la heurística del temor es una condición necesaria para aplicar una ética orientada al futuro. Sería importante que las instituciones públicas y privadas discutieran las implicaciones que tendría la introducción de IA en cada campo profesional, por medio de la realización de estudios especializados. Esta sería una tarea ardua, aunque necesaria, para llevar a cabo una primera medición de las implicaciones concretas que podrían existir en cada ámbito profesional. En este sentido, y recogiendo el testigo de lo expuesto por Jonas, con el fin de asumir responsabilidad orientada al futuro, y no comprometer la esencia del hombre en todas sus dimensiones, así como la supervivencia de la humanidad y la naturaleza, tendríamos que buscar las estrategias para rechazar y reducir el impacto de determinados intelectos sintéticos.

Es muy probable que determinados intelectos sintéticos pongan en riesgo la esencia del hombre en determinadas profesiones, mediante una profunda transformación o sustitución completa de su función. La supervivencia de la humanidad también puede estar cuestionada desde el aspecto económico, pues como ya se ha dicho anteriormente, la estabilidad del principal sustento familiar, que es el trabajo, está cuestionada en las próximas décadas.

Así pues, si se quiere garantizar un futuro de estabilidad para la humanidad, y no queremos atentar contra la esencia de determinadas profesiones, como la del médico o la del maestro, por decir algunas, es importante pensar y re-pensar, desde la ética de la responsabilidad, alternativas para reducir el impacto de la IA en el campo profesional. Las alternativas tendrían que ser discutidas y propuestas desde los sectores responsables y los afectados, planteando medidas de diversa índole, que podrían ir desde el rechazo de

determinada IA en algunos sectores, hasta la implementación total, pasando por la progresividad, etc.

b) Mejoramiento humano

El cambio de escenario que está por venir es más que inminente. El paradigma de la medicina tradicional, fundamentado en aspectos terapéuticos, que tiene por objeto el "reparar", está más que cuestionado, de ahí que broten nuevas ideas en lo relativo a un nuevo paradigma. En este caso, el paradigma que entra en escena es el de "perfeccionamiento".

El enfoque del paradigma médico tradicional, el terapéutico, está fundamentado en una larga tradición judeocristiana, que hace que la reacción espontánea sea la de considerar la naturaleza humana como un elemento que tiene que ver con la "eternidad" y con lo "intangible", de ahí que no sea posible en ningún caso mejorar, sino curar o reparar. Esto se debe principalmente a que la tradición judeocristiana ha plasmado en el espíritu tradicional de Occidente una idea de oposición a la alteración o modificación de la propia naturaleza humana. Pues bien, el transhumanismo se posiciona en contra de la postura tradicional judeocristiana, ya que se levanta sobre la máxima de un perfeccionamiento ilimitado y en desafío a la muerte y el envejecimiento, porque la ciencia y la tecnología nos brindan las herramientas necesarias para esos fines.

También es cierto que aunque el transhumanismo se posicione de forma contraria a la tradición judeocristiana, asimismo deriva, en cierta medida, de la tradición del humanismo clásico, además de otras tradiciones en las que no cabe detenerse por falta de espacio en este momento, representada desde Pico della Mirandola, hasta Kant, pasando por Francis Bacon o La Mettrie, quienes hacen hincapié en la idea de la perfectibilidad infinita del ser humano que no está encerrada en una naturaleza humana determinada e intangible, que es la que defiende la tradición judeocristiana.

El horizonte de posibilidades que nos presenta la realidad en la que vivimos es muy difícil de esclarecer con los medios que poseemos en la actualidad; sin embargo, las limitaciones humanas no deben frenar el interés del ser humano por querer ir más allá de lo meramente cognoscible en la actualidad. El transhumanismo nos presenta una serie de razones desde las cuales impulsar los anhelos en forma de superación de las limitaciones humanas, entre las que se encuentra: el aumento de la esperanza de vida, el aumento de nuestra

capacidad intelectual, el fortalecimiento de nuestros organismos corporales, la potenciación de nuestros sentidos y facultades y el cultivo de nuestro bienestar (Bostrom, 2008, pp. 5-7). Para Bostrom, son varios los motivos que nos llevarían a hacer una defensa razonable del mejoramiento humano por medio de la ciencia y la tecnología.

No es de extrañar que sea el sueco Nick Bostrom (2016) el que haga las nuevas aportaciones, pues, en su obra *Superinteligencia: caminos, peligros, estrategias*, avisará de los desafíos éticos que implica el desarrollo de la Inteligencia Artificial, por cierto, estrechamente vinculada con el transhumanismo. Aunque no todo el reconocimiento sobre el campo transhumanista se lo debería llevar Bostrom, sino también Julian Savulescu (2017), con quien comparte una importante obra en este campo. En este sentido, el transhumanismo llega en un momento en el que es necesario plantear nuevas utopías, y, por eso, su "estrategia" es muy atractiva, como señala Antonio Diéguez:

> Cuando tantas promesas hechas por otras utopías han dejado de ser creídas, el transhumanismo se presenta con promesas renovadas, no mucho más irrealizables que las de las viejas utopías, pero sí más potentes, deslumbrantes y atractivas. Una parte importante de su fuerza está precisamente en que ya no encuentran una competencia respetable, excepto desde el lado –también renovado– de las religiones (2017, pp. 20-21).

Esta búsqueda de transformación humana se erige primeramente sobre enfoques de nivel biotecnológico, impulsados principalmente por el Doctor Aubrey de Grey y Michael Rae (2013), que son unos referentes en este campo, al promover la utilización de los avances tecnocientíficos en el campo de la medicina y la biología. Este enfoque establece una importante relación con la IA, a partir del concepto "singularidad tecnológica", acuñado por Vernor Vinge (2013) y Ray Kurzweil (2017), planteando el traspaso de la frontera entre una inteligencia biológica y una inteligencia tecnológica. Para establecer el vínculo entre el transhumanismo y la IA, me gustaría hacer referencia a lo sostenido por Jairo Andrés Villalba:

> Dichos mecanismos en los que se basa la inteligencia artificial, permiten identificar un punto de correlación tecnocientífico frente a lo denominado "transhumanismo", pues ambos convergen en conceptos referenciales tales como bio (vida), info (información), cogno (conocimiento) y nano (simplicidad), equivalentes a la biotecnología, la información tecnológica, la ciencia cognitiva y la nanotecnología –elementos básicos de la convergencia NBIC–, cuyo propósito busca "reconocer el grado de correlación y amplitud de las

> máquinas en un contexto específicamente alternativo al servicio del desarrollo humano""(Villalba Gómez, 2016, p. 139)

La singularidad tecnológica identifica la IA como el medio para la búsqueda de la mejora de las capacidades humanas utilizando factores y variables tecnológicas. El desarrollo de la IA no se encuentra al margen de los avances tecnocientíficos en el campo de las últimas exploraciones de las redes neurales y otros campos relacionados, impulsados por proyectos como *Blue Brain Project* (2015), liderado por la École Polytechnique Fédérale de Lausana, Suiza, *Human Brain Proyect* de la Comisión Europea (2010-2020) o el proyecto BRAIN –siglas en inglés de *Brain Research through Advancing Innovative Neurotechnologies*– del Departamento de Salud y Servicio Humano de Estados Unidos de Norteamérica (2025). Estos proyectos tratan de mejorar los conocimientos en programación de IA para construir mapas neuronales más especializados y así llevar a cabo una modificación y mejoramiento del humano.

Nos enfrentamos a un importante desafío bioético relacionado con el desarrollo de la tecnología por medio de la IA y su influencia sobre los aspectos biotecnológicos planteados por el transhumanismo, que como ya sabemos promueven un desafío al envejecimiento y a la muerte humana. En este sentido, y siguiendo la línea de lo propuesto por Jonas, eso implicaría un importante riesgo, pues se estaría cuestionando la propia esencia del hombre al desafiar el envejecimiento y la muerte. No obstante, eso no es nada nuevo, pues los avances biotecnológicos, gracias a la IA, han permitido evitar numerosas enfermedades en los últimos años. Lo que verdaderamente nos compromete es una transformación de la esencia del hombre en profundidad, al estilo que la plantea Kurzweil (2017), cuando habla de un "trascender" lo biológico hacia lo tecnológico. Ahí es donde sería fundamental establecer una importante discusión para ver qué implicaciones existen y diseñar estrategias para obtener el mayor beneficio posible y el menor impacto deseable para la humanidad desde una perspectiva de responsabilidad.

c) Militar y de seguridad

El aprovechamiento y el desarrollo de la robótica y la IA en el campo militar es asombroso. Politólogos del Brookings Institution, que es un centro de

estudios fundado en el año 1916 en Washington D. C., como Peter W. Singer[3], sostienen que las guerras del futuro las librarán máquinas en la totalidad, que seguirán profundizando el distanciamiento en el campo de batalla, como ya se viene haciendo a lo largo de la historia con cañones, fusiles, aviones, misiles de largo alcance, etc., lo que implica un distanciamiento moral con respecto a los enemigos y pone de relieve algunas cuestiones éticas que tendrían que ser discutidas en el campo de batalla debido a la "deshumanización".

La tecnología de la IA ha llegado para quedarse en el campo militar y tenemos numerosos ejemplos de eso en la actualidad. En el año 2001, el diario *El País*[4] nos informaba acerca de la tecnología artificial que estaba desarrollando la agencia DARPA en el ejército estadounidense. Existen numerosas posibilidades de la aplicación de la IA en el campo militar, y de eso es muy consciente el Secretario de Defensa James Mattis, que en una visita reciente a Amazon, Google y otras compañías de Silicon Valley, se comprometió a seguir desarrollando la IA en el campo militar. La IA puede usarse en sistemas de entrenamiento con el proporcionamiento de enemigos impredecibles, el aumento de tropas, como por ejemplo con el Big Dog[5], la automatización del combate, la optimización en la identificación de objetivos, etc.

Eso evidencia que existe un importante impulso militar detrás de la IA, y ese impulso está orientado hacia la autonomía de los robots, como el caso del CIWS, que podría utilizarse en otros mecanismos y con otros fines, o los drones, por lo que surgen cuestiones éticas que tendríamos que plantear. Entre las cuestiones éticas –que surgen a voz de pronto y que son planteadas, por ejemplo, por un informe elaborado por Patrick Lin, George Bekey y Keith Abney (2008), para el Departamento de la Marina de los Estados Unidos– se encuentran:

- ¿Podrán los robots autónomos seguir las pautas establecidas por las leyes de guerra y las reglas de enfrentamiento, tal y como se especifican en las convenciones de Ginebra?

---

[3] Es interesante consultar sus reflexiones en materia de futuro tecnológico en el campo militar a través del siguiente enlace: https://www.brookings.edu/experts/peter-w-singer/. Consultado el 1 de diciembre de 2017.

[4] https://elpais.com/diario/2001/10/18/ciberpais/1003372528_850215.html. Consultado el 1 de diciembre de 2017.

[5] https://www.bostondynamics.com/bigdog. Consultado el 30 de noviembre de 2017.

- ¿Sabrán los robots diferenciar entre el personal civil y el militar?
- ¿Reconocerán a un soldado herido y se abstendrán de disparar?

Ronald Arkin, profesor del Instituto de Tecnología de Georgia, aborda algunas de estas preguntas a partir de estudios técnicos, en la investigación *Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/ Reactive Robot Architecture*.

En este sentido, nos enfrentamos a un importante cambio de visión de cómo se concibe el campo militar y de seguridad, que debe ser reflexionado desde la ética y siempre teniendo en cuenta el principio de responsabilidad para-con la humanidad y para-con el futuro. Si queremos garantizar los derechos y la dignidad de los seres humanos, como dicta el Derecho Internacional Humanitario (DIH), es importante que se adquiera responsabilidad, siguiendo la línea de Jonas, y comenzar a pensar alternativas desde lo político y lo jurídico a partir de la incorporación de la IA en el campo militar. La COMEST –siglas en ingles de la Comisión Mundial de Ética del Conocimiento Científico y la Tecnología– ha reafirmado su compromiso con el mandato de la UNESCO –siglas en inglés de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura– para promover y construir la paz, pues el objetivo principal es partir de análisis críticos y éticos sobre el uso de las tecnologías robóticas.

No obstante, y esta cuestión debe ser tenida muy en cuenta para llevar a cabo la aplicación del principio de responsabilidad a este campo, es fundamental que antes pongamos de relieve los puntos más importantes, con el fin centrar la atención en aquellos aspectos que requieran de una mayor asunción de responsabilidad. Así pues, para poner de relieve aquellos aspectos que requieren de una mayor atención, sería importante llevar a cabo una tarea de deconstrucción, como señala Justin Joque en su reciente obra *Deconstruction machines*.

El uso de los drones se está normalizando en los ejércitos. Existen varias definiciones de drones; sin embargo, creo que la más acertada y amplia es la propuesta por Chamayou, que lo define como "un vehículo de tierra, mar o aire, controlado a distancia o de forma automática" (1995, p. 11). La COMEST reconoce que el uso de drones nos presenta un nuevo abanico de peligros morales, pues nos sitúa en un escenario de completa desconexión física entre el operador y el dron, lo que podría suponer que fuera interpretado casi

como un juego (UNESCO, 2017, p. 22). La ONU (Organización de Naciones Unidas), ha mostrado su preocupación por el nuevo escenario al que nos conduce la tecnología en el campo militar, pues ha surgido inquietud por la efectividad que puede tener el DIH, a la luz del uso de tecnologías nunca antes conocidas desde el progreso de la IA. Quizás, una primera forma de asumir responsabilidad para estar a la altura de los nuevos tiempos, sería revisar el DIH a la luz de las nuevas tecnologías de IA empleadas en el campo militar.

Surgen importantes debates ante este fenómeno, por ejemplo, lo que tiene que ver con el reconocimiento de objetivos. Ante la pregunta anteriormente planteada, sobre cómo sería posible que un dron cumpliera con el mandato de la Convención de Ginebra de 1949, e hiciera una distinción entre un militar y un civil o personal sanitario, se plantean dudas. El reconocimiento de un objetivo no solo se reduce a factores de reconocimiento visual. El importante progreso de reconocimiento visual que ha experimentado la IA en la última década es asombroso; sin embargo, en la toma de decisiones militares, no solo se tienen en cuenta factores de reconocimiento visual, sino que existen una serie de condicionantes mucho más amplios y que, al menos por el momento, los drones no están preparados para manejar.

Siguiendo el hilo del principio de responsabilidad de Jonas, me gustaría rescatar lo defendido por la COMEST, que afirma que la responsabilidad del uso de cualquier sistema de robótica recae sobre el operador, es decir, sobre el ser humano que controla ese sistema. En este sentido, es importante no evadir responsabilidad, pues es el hombre el que decide utilizar este tipo de artefactos en el campo de batalla, y asume todas las posibles consecuencias que ello implica. Los ejércitos están obligados a asumir responsabilidad a este respecto, y a buscar las estrategias y formas necesarias para que el uso de la robótica no viole lo establecido en el terreno legal, pero tampoco en el moral.

Conclusiones

Como se ha podido comprobar a lo largo de estas páginas, nos enfrentamos a importantes desafíos, fruto del desarrollo tecnológico. La tecnología nos está planteando escenarios que eran desconocidos para nosotros. Sin embargo, se podría sostener que el hecho de que la humanidad se tenga que enfrentar a nuevos escenarios, como consecuencia de algún acontecimiento, no es algo reciente, y en eso podríamos estar de acuerdo. Pero en lo que no podríamos estar de acuerdo es que el nivel de complejidad es más alto que otras veces,

pues la IA nos está conduciendo hacia situaciones que presentan riesgos significativos. En este sentido, es fundamental someter el impacto de la IA en nuestras vidas a un profundo debate académico, político, social, en el que todos aquellos sectores que se vean involucrados realicen un importante ejercicio de reflexión.

La sugerencia de someter a debate público las repercusiones de la IA en nuestras vidas nace de la necesidad de imaginar alternativas y tiene que ver con la asunción de la responsabilidad. Como humanidad, nos enfrentamos a importantes desafíos, siendo necesario que discutamos desde el presente alternativas que garanticen un futuro de armonía. No se trata de comenzar la casa por el tejado, sino de dar los primeros pasos con humildad, pero con decisión, comenzando desde abajo, desde los espacios más cercanos, desde los centros de estudio, desde los centros de trabajo, desde los círculos barriales, a discutir este gran acontecimiento que condicionará y comprometerá nuestra existencia en diversos campos.

Finalizo este trabajo considerando que la asunción de responsabilidad es muy importante y fundamentalmente necesaria para poder elaborar alternativas frente a los desafíos. La falta de reflexividad en el presente nos puede conducir a consecuencias inesperadas en el futuro. En este sentido, un ejercicio de cautela, en el que el principio de responsabilidad sirva como principio rector en todos los campos mencionados en este trabajo, puede permitir el inicio del camino de forjar la conciencia necesaria en este tiempo de gran poderío tecnológico y de consecuencias inesperadas.

Referencias

Arkin, Ronald. (2011). *Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture.* Disponible en: https://www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf

Blue Brain Proyect. (2015). École Polytechnique Fédérale de Lausana. Disponible en: http://bluebrain.epfl.ch/

Bostrom, N. (2008). Why I Want to be a Posthuman When I Grow Up. *Medical Enhancement and Posthumanity* (pp. 107-137). Gordijn, B. & Chadwick, B. (Eds.). Oxford: Springer.

Bostrom, N. (2016). *Superinteligencia, caminos, peligros, estrategias*. Madrid: Teell Editorial.

Bostrom, N. & Savulescu, J. (2017). *Mejoramiento humano*. Madrid: Teell Editorial.

Brain research through advancing innovative neurotechnologies-BRAIN Proyect. (2014). Departament of health & human services. Disponible en https://braininitiative.nih.gov/pdf/BRAIN2025_508C.pdf

Chamayou, G. (2015). *A Theory of the Drone*. New York: The New Press.

De Gray, A. & Rae, M. (2013). *El fin del envejecimiento: los avances que podrían revertir el envejecimiento humano durante nuestra vida*. Berlín: Lola Books.

Diéguez, A. (2017). *Transhumanismo. La búsqueda tecnológica del mejoramiento humano.* Barcelona: Herder.

Esquirol, J. (2011). *Los filósofos contemporáneos y la técnica. De Ortega a Sloterdijk*. Barcelona: Gedisa.

Federacional Internacional de Robótica. (2017). *The Impact of Robots on Productivity, Employment and Jobs*. Disponible en https://ifr.org/img/office/IFR_The_Impact_of_Robots_on_Employment.pdf

Human Brain Proyect. (2010-2020). Comisión Europea. Recuperado de https://www.humanbrainproyect.eu/

Jonas, H. (1995). *El principio de responsabilidad. Ensayo de una ética para la civilización tecnológica*. Barcelona: Herder.

Jonas, H. (1997). *Técnica, medicina y ética: sobre la práctica del principio de responsabilidad*. Barcelona: Paidós.

Jonas, H. (2001). *Más cerca del perverso fin y otros diálogos y ensayos*. Madrid: Los libros de la Catarata.

Joque, J. (2018). *Deconstruction machines: writing in the age of cyberwar*. Minneapolis: University of Minnesota Press.

Kaplan, J. (2016). *Abstenerse humanos. Guía para la riqueza y el trabajo en la era de la inteligencia artificial.* España: Teell Editorial.

Kaplan, J. (2016). *Inteligencia artificial. Lo que todo el mundo debe saber*. España: Teell Editorial.

Kasparov, G. (2017). *Deep thinking. Where Machine Intelligence Ends*. London: John Murray.

Kurzweil, R. (2017). *La singularidad está cerca*. Berlín: Lola Books.

Lin, P.; Bekey, G. & Abney K. (2012). *Robot Ethics: The Ethical and Social Implications of Robotics.* Massachusetts: MIT.

Linares, J. (2008). *Ética y mundo tecnológico*. México: Fondo de Cultura Económica.

McCorduck, P. (1991). *Máquinas que piensan*. Barcelona: Tecnos.

Müller, V. & Bostrom, N. (2014). Future progress in artificial intelligence: A Survey of Expert Opinion. Müller, V. (Ed.). *Fundamental Issues of Artificial intelligence*. (pp. 51-68). Berlin: Springer.

Nilsson, N. (2001). *Inteligencia artificial: una nueva síntesis*. Madrid: McGraw-Hill.

Ortega, A. (2016). *La imparable marcha de los robots.* Madrid: Alianza Editorial.

Moravec, H. (1988). *El hombre tecnológico. El futuro de la robótica y la inteligencia humana*. Madrid: Temas de hoy.

UNESCO. (2017). *Report of COMEST on robotics ethics*. París: UNESCO.

Villalba G. & Jairo A. (2016). Problemas bioéticos emergentes de la inteligencia artificial. *Revista Diversitas-Perspectivas en Psicología*, 12 (1), pp. 137-147.

Vinge, V. (2013). *Un fuego sobre el abismo*. Madrid: La factoría de las ideas.

Winner, L. (2008). *La ballena y el reactor*. Barcelona: Gedisa.