PROPUESTAS PARA UNA INTERACCIÓN ÉTICA ENTRE LA INTELIGENCIA ARTIFICIAL (IA) CORPORIZADA Y LOS SERES HUMANOS

Alejandra Mariel Lovat[1]

1. ¿Qué es la IA?

La inteligencia artificial o IA actual es un conjunto de redes neuronales que puede aprender por prueba y error, y con mucha exactitud, a resolver problemas lógico-matemáticos, muchas veces mejor que nosotros, los humanos, pero carecen por ahora de conciencia (Caiafa y Lew, 2020).

El Parlamento Europeo en su Informe del 4 de enero del 2021 arribó a dos tipos de definiciones "sistema de IA" como:

> …todo sistema basado en programas informáticos o incorporado en dispositivos físicos que muestra un comportamiento que simula la inteligencia, entre otras cosas, mediante la recopilación y el tratamiento de datos, el análisis y la interpretación de su entorno y la adopción de medidas, con cierto grado de autonomía, para lograr objetivos específicos (punto 1).

Y, "autónomo":

> …todo sistema de IA que funciona interpretando determinados datos de entrada y utilizando un conjunto de instrucciones predeterminadas, sin limitarse a ellas, a pesar de que el comportamiento del sistema esté limitado y orientado a cumplir el objetivo que se le haya asignado y otras decisiones pertinentes de diseño tomadas por su desarrollador (Informe, 4/01/2021, punto 1).

La llamada "prueba de Turing" es la forma metodológica que encontró Alan Turing para comprobar que una máquina tuviera la capacidad de pensar. En el mismo, participan dos seres humanos -uno es el juez- y la máquina que se está investigando. La idea básica de la prueba es que uno de los seres humanos -el juez- descubra cuál participante es la máquina.

[1] Abogada de la Universidad de Buenos Aires (UBA). Doctora en Derecho con orientación en Derecho Privado por la Universidad de Ciencias Empresariales y Sociales (UCES). Investigadora y miembro del Instituto de Investigación en Formación Judicial y Derechos Humanos (UCES). Directora de Planeamiento de la Secretaría de Medios y Comunicación Pública de la Nación.

Para ello el juez debe conversar por turnos, hacer preguntas insidiosas y de tanto alcance como quiera, y la máquina tiene permiso para hacer todo aquello que le permita forzar una identificación falsa. La máquina puede usar todo tipo de trucos para que el juez crea que es un humano, incluso afirmar que pertenece a una cultura distinta de la del juez, mientras que el otro participante humano debe ayudar a que el juez haga una identificación correcta (Jack Copeland, 2012, p. 161).

Jeff Hawkins, un científico que desarrolló desde *softwares* para Intel hasta una empresa de investigación en neurociencia, señala que la IA actual no posee una inteligencia similar a la de un cerebro humano y que en el futuro para lograrlo deberá necesariamente ser corporizada. La clave consiste en que la IA debería tener sensores y poder moverlos a su voluntad y en función de las cosas que modela. Pero también podría ser una IA virtual moviéndose por internet (Heaven, 2021).

Ello, porque el cerebro humano funciona con cuatro atributos principales: 1. El aprendizaje a través del movimiento: hay que moverse para construir un modelo mental de las cosas, incluso si solo movemos los ojos o las manos. Esto se denomina materialización. 2. Decenas de miles de columnas corticales captan esta información sensorial, cada una con una imagen parcial del mundo. Compiten entre sí y se combinan a través de una especie de sistema de votación para construir una visión general. En una IA, esto podría suponer que una máquina controle diferentes sensores -visión, tacto, radar- para obtener un modelo más completo del mundo. 3. El aprendizaje continuo, con el que aprendemos cosas nuevas sin olvidar las anteriores. Los sistemas de IA actuales no pueden hacer esto. 4. Finalmente, estructuramos el conocimiento utilizando distintos marcos de referencia, lo que significa que nuestro conocimiento del mundo depende de nuestro punto de vista.

Por todo esto, Hawkins afirma que el test de Turing no resulta el método idóneo para demostrar que una IA es inteligente, ya que sólo está enfocado en poder engañar a alguien resolviendo una tarea con algún tipo de ingeniería inteligente, eso no significa un progreso hacia una comprensión más profunda de lo que significa ser inteligente (Heaven, 2021).

Sin embargo, las probabilidades de amplio desarrollo en el futuro de la IA no implican que la actual no sea útil.

2. *Sophia*: the robot

Un ejemplo de IA de los más avanzados es *Sophia*, la robot humanoide que tiene la capacidad de aprender los comportamientos humanos a través de la socialización con seres humanos por su fluida comunicación lingüística que le otorga incluso la habilidad de reírse de las bromas de su interlocutor. Es la primera inteligencia artificial que obtuvo la ciudadanía otorgada por el gobierno de Arabia Saudita el 25 de octubre de 2017 (Sophia, s.f.).

Fue creada en Hong Kong por la compañía americana *Hanson Robotics* y activada el 19 de abril del año 2015, desarrollada con los mayores avances en IA, una compleja serie de algoritmos basados en estadísticas computacionales, con un rápido proceso de información y la habilidad de reconocimiento de voz y facial de las personas con las que ha interactuado.

Resulta interesante que la primera IA con calidad de persona y estas específicas características que la humanizan sea corporizada como una mujer, quizás podemos imaginar que el argumento puede ser que el género femenino se encuentra culturalmente conceptualizado como “sensible, cálido, débil, maternal, cariñoso e intuitivo” -entre otras cualidades atribuidas socialmente- y podemos suponer que el fin es no causar una primera impresión aterradora al mundo, al estilo película de ciencia ficción como *Terminator*.

Sin embargo, la interacción con la máquina/IA no parece tan simple, teniendo en cuenta la gran variedad de sesgos cognitivos a los que se encuentra sujeta durante el *machine learning*.

3. Los sesgos cognitivos

Un ejemplo de sesgo cognitivo surgió de un estudio relacionado con el reconocimiento de imágenes, en el que se advirtió que las IAs encuentran e identifican objetos y los asocian con las etiquetas y categorías limitadas en las que se han entrenado, por lo que parte de los problemas de los sesgos hallados se encuentran en las etiquetas asociadas a la información del conjunto de datos.

“Descubrir no sólo bajo qué criterio las redes neuronales clasifican las imágenes sino incluso con qué *datasets* se han entrenado resulta misión aún más imposible debido a la opacidad de las IA y la de sus empresas propietarias”, lamenta

en un artículo publicado en Medium y avanzado por “El Diario”, Ujue Agudo, consultora y profesora de la UNAV. Por ejemplo, se descubrió que bajo las categorías “girl” -chica, en español- o “woman” -mujer- en español- se ordenaban más habitualmente imágenes de jóvenes posando o realizando tareas tradicionalmente asociadas a hombres. Los investigadores califican de “falta de coherencia” (Sánchez, 2020).

Otro caso se encontró en la aplicación “*FaceApp*”, que incorpora un sistema de alteración de imágenes que permite aplicar filtros de belleza, cambio de género o envejecimiento. Este servicio demostró a sus usuarios que, si en una imagen aparece una persona con pelo largo y rasgos de mujer, el *software* no es capaz de detectar el objeto que sostiene como una taladradora y lo puede cambiar como un secador (Sánchez, 2020).

La IA se alimenta de datos y patrones relacionados entre sí, es probable que al momento de enseñar con esa información a la IA entonces también se trasladen los sesgos humanos. Ellos pueden presentarse en la clasificación o simplemente como ausencias en el conjunto de los datos de entrenamiento y validación (Beivide García, 2020).

Una investigación llevada a cabo por un equipo de científicos de la República Checa y Alemania para determinar los efectos de los sesgos humanos en interpretación del resultado utilizado por la IA determinó que más de veinte diferentes sesgos cognitivos podrían alterar potencialmente el desarrollo de las reglas de aprendizaje automático y propone métodos para “desvelarlas”.

La dificultad radica en que la mayoría de las veces no sabemos que estamos siendo parciales. Creemos que estamos siendo inteligentes o intuitivos, o simplemente no pensamos en eso (Torres, 2018).

Un ejemplo de sesgo cognitivo en la IA fue que *Microsoft* debió borrar en 2017 un *chat bot* -conversador artificial- porque se volvió nazi y adorador de Hitler, porque lo que sucede es que los algoritmos no pueden detectar si una corriente de datos es verdadera o falsa, siempre la base sobre la que operan es información provista por seres humanos (Campanario y Vazhnov, 2017, p. 102).

Y es en el momento actual de verdadera revolución digital de la historia de la humanidad en el que más contenido se está produciendo al publicar constantemente todas las personas en las redes, hacer comentarios, opiniones, prejuicios, en definitiva, sus sesgos. Por ello es sumamente importante la labor y la “ética” de los programadores y empresas que manejan esos datos, ya que toda persona que trabaje

con algoritmos a la hora de tomar decisiones tiene que tener en cuenta el gran impacto social que tiene su tecnología (Aranda, 2019).

En septiembre del 2019 un artículo nos adelanta la investigación efectuada por científicos de la Universidad de Columbia, quienes crearon un brazo robótico sólo con la capacidad de aprender -o sea sin programación acerca de la física ni de su propia estructura- tras moverse aleatoriamente al principio este robot aprendió acerca de su propia corporización golpeándose contra cosas, como hacen k, al cabo de un tiempo no sólo aprendió a moverse con sentido, sino a repararse a sí mismo y a escribir con un rotulador.

Los propios científicos creen que esta habilidad nos acerca a la autoconciencia de las máquinas (Científicos, 2019).

4. Niveles de Conciencia de la IA

El “despertar” de la conciencia de las IAs denominado *MMC*’ o “modelamiento de la conciencia utilizando máquinas” es un estudio que se viene desarrollando hace años y el resultado esperado es crear una máquina con conciencia fenoménica, es decir, con empatía y sentimientos.

Y la posibilidad de este tipo de desarrollo significaría otorgarle a la IA, necesariamente corporizada, la facultad de alterar sus programas en favor de una mejor respuesta al ambiente, siendo determinante la importancia del rol de la conciencia para desarrollar ciertas tareas y tomar ciertas decisiones, facilitando la creación de sociedades y acuerdos (Fuentes Barassi, 2011, pp. 50-55).

Esa capacidad que se espera que pueda alcanzar la IA, asociativa de conjugar o intuir cosas, adquirida a través de la experiencia requiere un sistema cognitivo muy complejo, de entender primero y de desarrollar después, como es el caso de los chistes o las ironías (Aranda, 2018).

De todas maneras, por el momento no se espera lograr tan avanzada y poderosa IA. Sin embargo, estos sucesos nos permiten arribar a ciertas reflexiones: ¿cómo los robots/IAs se relacionarán con los humanos en el futuro? ¿cuál podría ser el mejor marco ético en este tipo de interacción progresiva entre humano-máquina?

Resulta muy interesante que la propia *Sophia* destaque la estrecha relación en el futuro de compañerismo y cooperación entre los robots/algoritmos y los seres humanos.

5. La inevitable relación laboral entre seres humanos e IAs

Existen diversas consideraciones sobre, por ejemplo, el futuro mundo del trabajo. Yuval Harari (2018) señala que la potencialidad del uso de las IAs resulta prometedor a nivel financiero, e incluso al principio, una gran colaboración para los trabajos rutinarios, pesados o de peligro para los seres humanos, lo que podría comenzar con una "ayuda" progresará en un reemplazo continuo en actividades que requieren capacidades mucho más avanzadas ya que en su obra *21 lecciones para el siglo XXI* adelantó

> La IA no solo está a punto de suplantar a los humanos y de superarlos en lo que hasta ahora eran habilidades únicamente humanas. También posee capacidades exclusivamente no humanas, lo que hace que la diferencia entre una IA y un trabajador humano sea también de tipo, no simplemente de grado. Dos capacidades no humanas importantes de la IA son la conectividad y la capacidad de actualización (Harari, 2018, p. 30).

Ello implicará que los humanos deban estar continuamente capacitándose y cambiando de tipo de tareas o trabajos que se vayan requiriendo, y a pesar de ello, las IAs resultarán tan eficientes que los humanos ya no tendremos trabajos que hacer, exponiéndonos a un posible dominio completo de las máquinas, donde los humanos -la mayoría, de clase pobre- no tengan qué hacer y quizás los Gobiernos tengan que pensar en una "asignación básica universal" para la subsistencia (Harari, 2018, pp. 43-47).

Ahora bien, otras reflexiones en torno al tema indican que quizás el futuro no sea tan espeluznante, en tanto la "progresión" a la que nos referíamos antes puede llegar a ser ralentizada, esperándose que en Latinoamérica se priorice la necesidad de un tiempo de aceptación del cambio a nivel psicológico, sociológico y cultural de las sociedades muy a pesar de las ganancias potenciales.

Los nuevos empleos para los humanos serán aquellos que complementen o trabajen en conjunto con la tecnología, sobre todo como "entrenadores" para sistemas de inteligencia artificial, "explicadores" para comunicar los resultados de estos sistemas, y "sostenedores" para monitorear el comportamiento de los sistemas (Plata, s.f).

Aranda (2018) indica que, en la tarea de interpretar y extraer información de documentos legales, la automatización de procesos mediante el uso de la IA va a

permitir que el ser humano a cargo se centre en aportar más valor a dicho proceso, disponiendo de más tiempo para la escucha al cliente, la estrategia, la creatividad, la empatía y la resolución de problemas, entre otros. Asimismo, señala la importancia de generar “innovación emocional”, en tanto las empresas que venden productos y servicios están orientadas a la personalización en el ofrecimiento de los mismos conforme nuestros gustos y/o necesidades basadas en el uso que hacemos de las redes sociales.

También sostenemos que las profesiones relacionadas con el acompañamiento terapéutico de las personas, como las relacionadas con la psicología, *coachers* en inteligencia emocional y relacionadas a trabajos espirituales como el yoga y la meditación serán sumamente necesarios y buscados en estas nuevas sociedades por los propios humanos.

6. Ética y derecho en torno al uso de IA

Pensamos que los principios bioéticos vienen a tomar parte, y si bien pueden ser opuestos entre sí, deben ser aplicables y coexistir al mismo tiempo siguiendo el paradigma de convergencia que proponen Maliandi y Thüer (2008), éstos son: el de “precaución” -como exigencia para preservar equilibrios, que puede entenderse como de “no maleficencia” o “de no dañar” la salud humana o el medio ambiente, las conductas humanas y los recursos económicos de la población-; el de “exploración” viene a ser de beneficencia en tanto sin investigación -ética- no hay evolución posible; los otros dos son “la no discriminación genética” y el “respeto por la diversidad genética”. El cumplimiento de esos cuatro principios ante el conflicto crea como principio superior al de la “convergencia” que resuelve la imposibilidad de aplicación total de los cuatro con la posibilidad de su armonización (Lovat, 2019, p. 16).

Un marco de regulación ética para la interacción entre humanos-máquinas, entendemos debe responder no sólo a los derechos humanos concretamente plasmados en convenciones y tratados internacionales, sino a también encontrarse intercedida por conocimientos psicológicos, sociales, de perspectiva de género, diversidad y culturales del ambiente en el que intervendrá la IA, sin omitir el requerimiento de que se desenvuelva en un entorno ético de respeto de la privacidad en la comunicación con los humanos y recopilación de datos.

Algunas pautas para el desarrollo de una normativa ética que se incorpore a los robots/IAs debe realizarse como premisas lógicas, por ejemplo, en torno a los derechos humanos: “debo proteger la vida de los seres humanos, mi usuario es humano, entonces debo proteger su vida”.

Como ejemplo de esto, las leyes de Asimov fueron tomadas por el Parlamento Europeo, entre otros principios generales, como fundamentalmente dirigidas a los diseñadores, fabricantes y operadores de robots, incluidos los que disponen de autonomía y capacidad de autoaprendizaje integradas (Anexo, 2017).

Leyes de Asimov actuales:

- Un robot no hará daño a un ser humano o, por inacción, permitir que un ser humano sufra daño.

- Debe hacer o realizar las órdenes dadas por los seres humanos, excepto si estas órdenes entrasen en conflicto con la 1ª Ley.

- Debe proteger su propia existencia en la medida en que esta protección no entre en conflicto con la 1ª o la 2ª Ley.

Pero también cargada de reforzamiento verbal y gestual positivo aportando al individuo -junto a conocimiento “data” que podrá obtener la IA en interacción con su usuario- un complemento holístico para mejorar su estado anímico en un momento determinado.

Asimismo, pensamos principios éticos que rescaten solamente los mejores valores como los conceptos de libertad, intimidad/privacidad, justicia, responsabilidad, integridad, respeto, dignidad, autodeterminación, no discriminación -inclusión-, igualdad, solidaridad, colaboración, y compasión, entre otros, que pueda emular los sentimientos humanos más nobles.

Una ética con enfoque de género y diversidad con especial emulación de empatía hacia los más vulnerables: personas mayores, niños, niñas, enfermos, mujeres, diversidad de género y grupos minoritarios invisibilizados.

Este criterio de orientación de esta IA cargada de este tipo de valores éticos puede pensarse para el ámbito sanitario, educativo, de senectud, e incluso doméstico.

Sin embargo, creemos que el conflicto pudiera suceder para el caso que, durante el aprendizaje en interacción de la IA con diversos humanos, pueda reprogramar sus valores a voluntad.

Los resultados se esperan prometedores, pero ¿y si la corporización y desarrollo de la IA termina en una catástrofe para la humanidad?

En este caso siempre será necesario contar con la posibilidad de apagarla, porque ante su desarrollo progresivo y aprendizaje continuo de los humanos, la introducción de sesgos se supone ser inevitable, y ésta es la barrera principal que se impone resolver al corporizar una IA y exponerla al mundo.

Otras cuestiones éticas que deben analizarse responden al uso por parte de los humanos de los robots propios o ajenos -ya que el otorgarles calidad de persona o ciudadanía, como *Sophia*, los podría convertir en iguales en términos legales-; también las pautas de conductas, derechos y obligaciones de las empresas desarrolladoras de las IAs y constructoras de los robos o la intervención autoridad de aplicación, regulación y control.

7. Conclusiones

En conclusión, los principios éticos deben ser acompañados por una normativa clara y lógica basada en el sentido común y sin contradicciones que pudieran dar lugar a que la IA resuelva tomar decisiones con un enfoque utilitarista.

El utilitarismo es una teoría de la rama ética de la filosofía según la cual las conductas moralmente buenas son aquellas cuyas consecuencias producen felicidad, existiendo dos elementos básicos que definen el utilitarismo: la relación del bien con la felicidad de los individuos y su consecuencialismo, actuar de modo que nuestras acciones produzcan la mayor cantidad de felicidad en el mayor número de personas posible (Triglia, s.f.).

Sin embargo, el objetivo de lograr una felicidad para la mayoría de las personas deja de lado los grupos minoritarios, las personas que no comparten los mismos valores de la mayoría, y como otras teorías apareja la infelicidad de esos grupos.

El grado de razonamiento y los niveles éticos que se le impregnen a la IA dependerán de los valores y la ética de sus desarrolladores, en un mercado en el que Mark Purdy -un experto global en IA de la empresa Accenture- afirma que la IA tiene la capacidad de duplicar la tasa de crecimiento de las economías evolucionadas -que

representan el 50% del PBI mundial- de acá al año 2035 y de aumentar la productividad laboral en el mismo período un 40% (Campanario y Vazhnov, 2017, p. 107).

8. Bibliografía y fuentes de información

8.1 Bibliografía

Aranda, C. (noviembre 2019). Sesgos e inteligencia artificial: ojo al dato. *Revista Uno*. https://www.revista-uno.com/33-revista-uno-tecnologias-exponenciales/sesgos-e-inteligencia-artificial-ojo-al-dato/

Beivide García, A. (octubre 2020). El sesgo en la inteligencia artificial. *ISA Sección Española*. https://isa-spain.org/el-sesgo-en-la-inteligencia-artificial/?cn-reloaded=1

Caiafa C., y Lew, S. E. (17 de junio de 2020). ¿Qué es la Inteligencia Artificial? *Boletín Radio astronómico, CONICET.* https://ri.conicet.gov.ar/bitstream/handle/11336/110093/CONICET_Digital_Nro.57cd70fa-207f-493b-9c0e-c4363e32746b_A.pdf?sequence=2&isAllowed=y

Campanario, S., y Vazhnov, A. (2017). *Modo Esponja. Cómo absorber con creatividad el impacto del cambio acelerado. El viaje del aprendizaje permanente*. Sudamericana.

Fuentes Barassi, C. (2011). Conciencia e Inteligencia Artificial: Consideraciones Críticas sobre la Plausibilidad de que una Máquina Programada Posea Conciencia Fenoménica (Tesis de Magíster en Estudios Cognitivos). Universidad de Chile, Santiago de Chile. http://repositorio.uchile.cl/bitstream/handle/2250/108753/fi-fuentes_c.pdf?sequence=3&isAllowed=y

Harari, Y. (2018). *21 lecciones para el siglo XXI*. Debate.

Heaven, W. (8 de marzo de 2021). El test de Turing es una de las peores cosas que le ha pasado a la IA. *MIT Technology Review.* https://www.technologyreview.es/s/13289/el-test-de-turing-es-una-de-las-peores-cosas-que-le-ha-pasado-la-ia

Jack Copeland, B. (2012). *Alan Turing. Pionero de la era de la información*. Epub Libre.

Lovat, A. (2019). Seres humanos biónicos e inteligencia artificial humanizada. Nexo entre la humanidad y las máquinas. *Ratio Iuris. Revista de Derecho Privado*, 7(2), 1-31. http://dspace.uces.edu.ar:8180/xmlui/handle/123456789/4821

Plata, B. (s.f.). El futuro del trabajo: ¿robots versus humanos? *IDB Mejorando Vidas*. https://www.iadb.org/es/mejorandovidas/el-futuro-del-trabajo-robots-versus-humanos

Sánchez, J. (1 de octubre de 2020). Así funcionan los sesgos de la Inteligencia Artificial. ABC Soluciones. https://www.abc.es/tecnologia/informatica/soluciones/abci-funcionan-sesgos-inteligencia-artificial-202009130134_noticia.html?ref=https%3A%2F%2Fwww.google.com%2F

Torres, A. L. (17 de abril de 2018). El sesgo humano, el gran problema para la IA y así es cómo lo solucionaremos. *Planeta Chatbot*. https://planetachatbot.com/sesgo-humano-problema-inteligencia-artificial/

Triglia, A. (s.f.). Utilitarismo: una filosofía centrada en la felicidad. Stuart Mill y Jeremy Bentham desarrollaron esta teoría filosófica. https://psicologiaymente.com/psicologia/utilitarismo

8.2 Fuentes de información

Anexo a la Resolución del Parlamento Europeo con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica -2015/2103- INL- (16 de febrero de 2017). Parlamento Europeo.

http://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_ES.html?redirect#def_1_3

Científicos de la Universidad de Columbia han construido un robot que aprende acerca de sí mismo igual que lo hace un bebé. Podría ser un paso hacia la autoconciencia de las máquinas (4 de septiembre de 2019). *Uk International, The day, News to open minds*. https://theday.co.uk/translations/espanol/nuevo-robot-va-camino-de-la-autoconciencia

Informe sobre inteligencia artificial: cuestiones de interpretación y de aplicación del Derecho internacional en la medida en que la UE se ve afectada en los ámbitos de los usos civil y militar, así como de la autoridad del Estado fuera del ámbito de la justicia penal. [2020/2013(INI)] (4 de enero de 2021). *Parlamento Europeo*. https://www.europarl.europa.eu/doceo/document/A-9-2021-0001_ES.html

Sophia, robot. (s.f.). Wikipedia. https://es.wikipedia.org/wiki/Sophia_(robot)