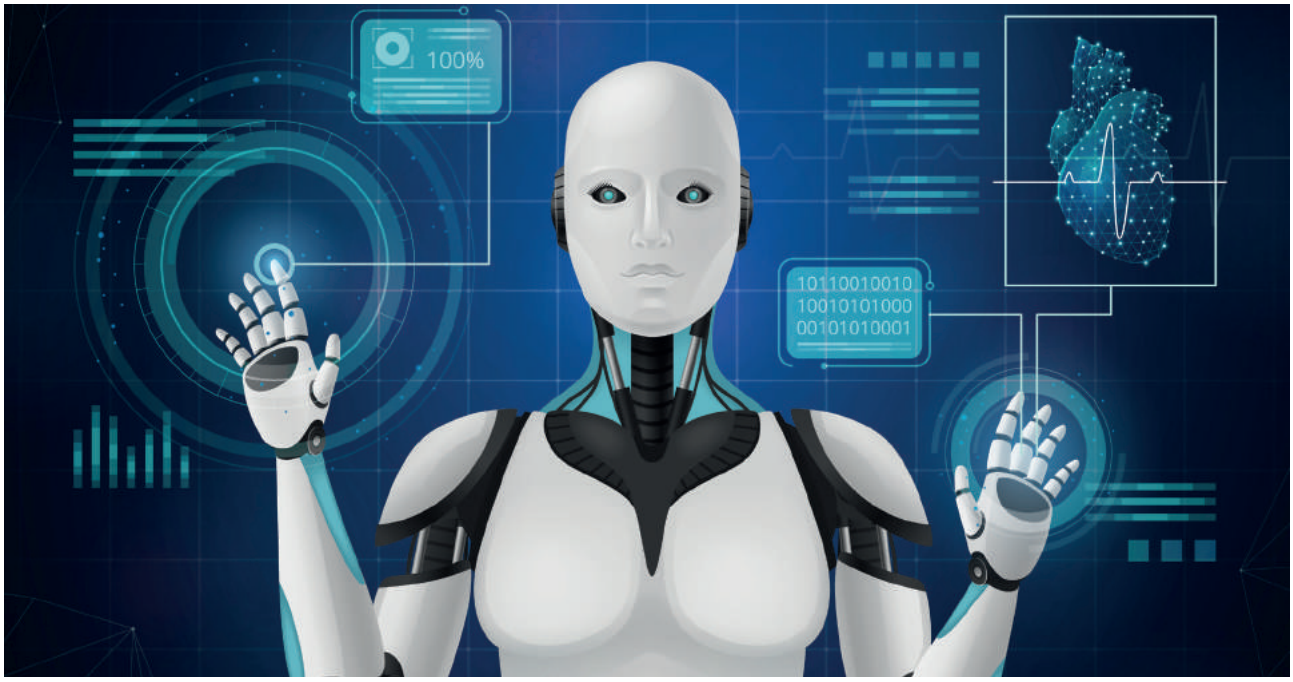
Consideraciones y recomendaciones éticas en Inteligencia Artificial (IA)

POR: RAFAEL A. GONZÁLEZ*



Hace mucho tiempo que los experimentos morales han servido de base para explorar filosóficamente las dificultades y los modelos de decisión en escenarios de dilemas morales y éticos. Esto se hace a través de escenarios ficticios que pongan a pensar a la gente sobre qué haría en cada caso.

El problema del tranvía, por ejemplo, ya es bien conocido. La decisión consiste en elegir si accionar o no una palanca para desviar un tranvía que se sabe terminará por arrollar a cinco trabajadores en la vía. El dilema moral radica en que, al hacerlo, se arrollará a otra persona que está en la vía alterna.

Existen múltiples variaciones de este dilema. Puede ser que, en lugar de una palanca y una persona en la vía alterna, lo que pueda salvar a los trabajadores sea empujar a una persona en la vía. Hay escenarios incluso más truculentos, como el del "bebé llorón", en que su padre debe decidir si sofocarlo para proteger a un grupo de civiles escondidos de un ejército enemigo y que sin duda los masacrarían si oyen el llanto. Esencialmente, estos experimentos nos presentan el mismo dilema: unas vidas por otras[1].

En términos filosóficos, se puede hablar de decisiones orientadas desde el utilitarismo o desde la deontología. El enfoque utilitario es racional y busca

maximizar la utilidad de la decisión, con lo cual salvar cinco vidas a cambio de una, resulta ser lo moralmente correcto. El enfoque deontológico en cambio, propone un imperativo moral, según el cual ciertas acciones son simplemente malas, independientemente de las consecuencias, con lo cual seguir estrictamente la ley de "no matarás" implicaría no accionar la palanca en el caso del tranvía.

Naturalmente, son dilemas porque no son tan sencillos. Desde la deontología, accionar la palanca es directamente conducir a la muerte de una persona. Pero, ¿la pasividad no podría considerarse como una responsabilidad tácita en la muerte de cinco? Desde el utilitarismo, ¿valen todas las vidas lo mismo? Si son cinco desconocidos a cambio de la madre, o si son cinco asesinos a cambio de un médico que salva vidas, ¿valen lo mismo?

Pero lo que nos convoca es todavía más difícil. ¿Qué pasaría si esa decisión es tomada autónomamente por una Inteligencia Artificial? El enfoque deontológico podría ilustrarse con la primera Ley de la Robótica de la novela "Yo, Robot" de Asimov: "Un robot no debe dañar a un ser humano o, por su inacción dejar que un ser humano sufra daño". Ante el dilema del tranvía, esto suena a corto circuito. Y si la alternativa es un algoritmo que maximice la utilidad, de todas maneras, habría que definir si toda vida humana es equivalente, ya que para los humanos típicamente no es así y tendemos a dar más a valor a quienes pertenecen a nuestro grupo (familia, nación, grupo étnico).

De hecho, los experimentos morales ya hacen parte del arsenal en el estudio de la ética en la IA conduciendo a escenarios de ciencia ficción donde esta tecnología podría decidir atacar a la humanidad para protegerla de sí misma.

Pandemia, prejuicios y robots asesinos

El tema es que esto ya no es ciencia ficción. Todos hemos visto en vivo y en directo la toma de decisiones morales en el caso de la pandemia del Covid-19. Si, por ejemplo, un gobierno impone el aislamiento para salvar vidas, podría estar generando pérdida de empleos, problemas sicológicos y falta de atención ante otras condiciones médicas.

En la distribución de vacunas ocurre lo mismo, ¿primero los viejos o los jóvenes; primero los médicos o las personas con condiciones preexistentes de alto riesgo; inmigrantes o nacionales; países ricos o países pobres? Estas decisiones se han ido tomando en muchos casos con el apoyo de modelos epidemiológicos, simulaciones y modelos predictivos de aprendizaje de máquina.

> *"Ser cuidadosos con los supuestos y premisas, implícitos en los modelos, particularmente cuando se transfieren de un campo de aplicación a otro"*

No obstante, pareciera ser que el timón moral sigue en manos de unos humanos que muchas veces lo hacen desde consideraciones políticas (utilitarismo de votos) o económicas (básicamente los primeros países en vacunarse son los más ricos). Pero el hecho es que ya no hacen falta experimentos artificiales comparando un puñado de vidas contra otros: el Covid ha cobrado ya más de dos millones de vidas.

El año pasado, Minciencias lanzó la convocatoria "Mincienciatón", lanzada para dar respuesta al Covid-19, así como para generar capacidades de respuesta ante emergencias similares. El Centro de Excelencia y Apropiación en Big Data y Data Analytics, Alianza CAOBA, propuso justamente la elaboración de modelos de IA, aprendizaje de máquina y analítica de grandes volúmenes de datos, para analizar información histórica o generar escenarios futuros. Uno de los ejercicios fue una revisión sistemática de todos los modelos que a la fecha se hubieran desarrollado para Coronavirus (no solo el último SARS-CoV-2).

Encontramos cerca de 200 artículos que aplicaban docenas de técnicas o algoritmos diferentes (muchas veces varios en el mismo artículo). Quisiera decir que fue una sorpresa el que solo 20 de los artículos tenían alguna discusión de implicaciones éticas que, de hecho, se reducía a una breve mención de que el proyecto había surtido un proceso de aprobación por algún comité. Por lo demás, no parece ser una preocupación prioritaria.

Además de los ejemplos asociados a la pandemia, también se han evidenciado en los últimos años otros múltiples casos reales, particularmente centrados en sistemas de recomendación. Cuando el algoritmo nos recomienda qué serie ver en Netflix o qué vídeo ver en YouTube esto tiene implicaciones en la cultura nacional, en la polarización política o en la manipulación social.

Cuando un algoritmo sugiere qué empleado contratar dentro de un conjunto de candidatos, hemos visto en acción los prejuicios que han aprendido de nuestra propia historia, resultando en decisiones machistas, racistas o clasistas. Lo mismo ocurre con los sistemas de apoyo a la identificación de sospechosos de un crimen o a la admisión de estudiantes a una universidad.

Quizá uno de los temas más críticos y de mayor interés creciente sea el de las armas autónomas. Hace algunos años se inició una campaña global para abolir los "robots asesinos"[2]. Esta campaña está avalada por el Secretario General de la ONU, más de dos docenas de recipientes del Premio Nobel, expertos y académicos reconocidos en IA, así como 30 países, incluyendo a Colombia.

Los argumentos más relevantes son la prevención de una nueva carrera armamentista, el hecho que desplegar ataques sería más rápido y por ende más prevalente, y el que lo consideran inmoral. Esto último básicamente es consecuencia de que no habría participación ni responsabilidad humana en las decisiones, lo cual, crucialmente, eliminaría la "compasión humana".

No obstante, cualquier progreso que pueda haber hecho esta campaña parece haber sido eliminado por completo con la reciente recomendación de un panel de expertos constituido para el Senado de los EE.UU. El denominado *National Security Commission on Artificial Intelligence* ha dicho todo lo contrario que la campaña para abolir los robots asesinos. Para este panel, liderado por el ex-CEO de Google, Erik Schmidt, es un imperativo moral para los EE.UU. continuar con el desarrollo de armas autónomas.

> *"Experimentos morales ya hacen parte del arsenal en el estudio de la ética en la IA, conduciendo a escenarios de ciencia ficción, donde esta tecnología podría decidir atacar a la humanidad para protegerla de sí misma"*

Por una parte, y este es el argumento principal (aparentemente desde una ética utilitarista), los robots se equivocan menos y el resultado entonces serían menos víctimas inocentes; aunque no es claro si es la IA la que decide quién es culpable o inocente. Por otra parte, ante la carrera armamentista se encojen de hombros: si no lo hacemos nosotros, lo harán los rusos y los chinos.

Así que debería ser claro que hace rato que las implicaciones éticas de la IA no son ciencia ficción. Si bien, por lo menos el debate y el interés se han despertado, también es cierto que pareciera que no hayan dado lugar a soluciones, sino más bien a una creciente angustia existencial. Que es inevitable, dicen. Que si la inteligencia artificial se equivoca es nuestra culpa, son nuestros datos, nuestros prejuicios. Que, si la decisión es errada, en todo caso no lo será tanto como la decisión humana. La salida cínica.

Algunas recomendaciones

Lo cierto es que hay muchas cosas que sí podemos hacer. Motivados por las decisiones políticas en medio de la pandemia, un grupo de expertos recientemente ha recogido una serie de recomendaciones que no solo aplican para modelos que usen IA, sino simulaciones, modelos epidemiológicos o analítica de datos en general.

El manifiesto, publicado en la revista *Nature*, sostiene que, para garantizar efectos prosociales, los modelos, en realidad los modeladores, deberían seguir cinco recomendaciones:

- Supuestos. Ser cuidadosos con los supuestos y premisas, implícitos en los modelos, particularmente cuando se transfieren de un campo de aplicación a otro.
- Parsimonia. Manejar adecuadamente la complejidad, ya que demasiadas variables pueden hacer que el modelo sea más ajustado a los datos de entrenamiento, pero usualmente el precio que hay que pagar es una menor precisión en las predicciones (que es donde radica la utilidad del modelo). Es el famoso principio de la Navaja de Ockham, que debería estar en el fondo de pantalla de todo ingeniero.
- Transparencia. Reconocer que las decisiones de diseño no están libres de valores. Los constructos, las hipótesis, incluso las herramientas que se decida emplear, vienen cargadas de prejuicios, preferencias y orientaciones disciplinares, entre otras. Como mínimo, esto implica ser completamente transparentes y rigurosos en la declaración y documentación de dichas decisiones.
- Cuentas y cuentos. Ser conscientes del impacto que pueden generar los modelos y del riesgo que, al presentarnos números, gráficas y predicciones limpias y ordenadas, nos pueden convencer que son la verdad revelada. Esto nos lleva de escenarios donde podemos tener aproximaciones válidas, a precisiones equivocadas. Se trata, pues, de no remplazar la narrativa por el número, sino de que cada cuenta vaya acompañada por su cuento que la explique, justifique y matice.
- Ignorancia. Reconocer que ningún modelo es perfecto, que siempre, inevitablemente, captura solo una porción del mundo y corremos el riesgo de que el modelo oculte nuestra ignorancia y, al hacerlo, nos presente soluciones que no se corresponden con la realidad. Es recordar que Herbert Simon (Premio Nobel y pionero de la inteligencia artificial) demostró nuestra racionalidad limitada y, en consecuencia, nos invitó a deshacernos de la idea que hay modelos de optimización para todo; que en la mayoría de situaciones complejas, a lo sumo podemos apuntarle a la satisfacción ("*satisficing solutions*").

Estas recomendaciones son generales y requieren del compromiso de las organizaciones, agremiaciones y academia, pero hay otras cosas que se pueden hacer a nivel técnico e individual.

Cada vez cobra más fuerza la noción de "justicia algorítmica" (y términos relacionados como "*fair machine learning*", "*fair classifiers*" o "*fair AI*") que ofrece soluciones que contribuyen a disminuir la discriminación.

Por ejemplo, la presencia de variables de confusión (típicamente por no tener claras las relaciones de causalidad en los modelos), es un factor crucial en el prejuicio de los modelos de aprendizaje de máquina, para lo cual ya hay varias aproximaciones matemáticas y algorítmicas que pueden reducir la probabilidad de que se presenten. Una discusión general de esto se puede encontrar en el popular "Book of Why" de Judea Pearl, aunque existen muchas contribuciones específicas según cada dominio de aplicación[3].

Por el lado individual hay un creciente movimiento que nos permite hacer parte de diversas plataformas de activismo social que puedan identificar, denunciar, sensibilizar o incluso aportar datos para reducir el sesgo algorítmico o, en general, para abordar la dimensión ética de la IA. Esto incluye campañas como la que mencionamos contra los robots asesinos, o grupos para cuestiones específicas de género, raza, edad, inclusión financiera y acceso a crédito, defensa de falsos positivos en el campo policial, entre otros.

> *"Hay un creciente movimiento que permite hacer parte de diversas plataformas de activismo social que pueden identificar o incluso aportar datos para reducir el sesgo algorítmico y abordar la dimensión ética de la IA"*

También se ha visto cada vez mayor participación de los empleados en los usos, clientes y aproximaciones de las empresas tecnológicas. Ya ha habido ampliamente publicitadas protestas (*walkouts*) de empleados de Google y Microsoft en contra de algunos contratos de defensa militar, llevando a que de hecho las empresas reversen o no renueven dichos contratos.

De todas maneras, habrá preguntas que no podemos resolver. O mejor, preguntas para las cuáles nuestra tarea de discernimiento debe ser continua y no habrá respuesta definitiva. Si queremos mayor justicia y mayor ética incorporada en los algoritmos, por ejemplo, habría que preguntarse si esto se hará basándonos en los valores de un país o de otro, o si se hará con nuestros valores actuales que quizá resulten obsoletos en veinte años.

¿Deberán los algoritmos entonces evolucionar éticamente como nosotros; deberán consistentemente soportar una ética mínima; o deberán obedecer a los humanos y quedarse en el rol de ayudante o consejero?

* Rafael A. González, PhD, Profesor Titular de la Facultad de Ingeniería, Pontificia Universidad Javeriana

1 Para consultar, e incluso "jugar" a decidir entre in sinnúmero de variantes, ver por ejemplo https://www.moralmachine.net/

2 https://www.stopkillerrobots.org/

3 Por ejemplo, en Zhao, Q., Adeli, E., & Pohl, K. M. (2020). Training confounder-free deep learning models for medical applications. Nature Communications, 11(1). https://doi.org/10.1038/s41467-020-19784-9 se presenta en detalle una estrategia para el caso de aplicaciones médicas.