Máquinas justas en el mundo digital

La inteligencia artificial plantea nuevos desafíos en la sociedad. Aspectos tales como la transparencia, los niveles de sesgo y la equidad son esenciales para garantizar su beneficio equitativo.

Desde hace varios años que la inteligencia artificial y los modelos de aprendizaje de máquinas ya no son el futuro, sino que son el presente; y es que aunque no nos demos cuenta, ya forman parte de nuestro día a día a través distintas cosas, tales como el algoritmo que nos recomienda películas o lo que nos muestra nuestro *feed* en redes sociales. Sin embargo, la aplicación de estas herramientas no siempre es igual. No es lo mismo tener un modelo que predice qué película uno querría ver, a uno que, por ejemplo, esté pensado para poder predecir el resultado de un juicio penal.

Mientras que un algoritmo de recomendación de películas puede cometer errores sin tener un gran impacto en nuestras vidas (es decir, nos sugiere una película que realmente no nos gusta), la situación cambia drásticamente cuando estos algoritmos se aplican en contextos con graves consecuencias para los individuos. Por ejemplo, en el sistema judicial, un algoritmo que intenta predecir los resultados de un juicio penal podría influir en la sentencia de una persona, potencialmente alterando su vida de manera significativa.

Los problemas éticos emergen cuando consideramos los sesgos presentes en los datos con los que estos algoritmos son entrenados. Si el sistema judicial ha tenido históricamente un sesgo hacia ciertos grupos demográficos, un algoritmo entrenado con estos datos puede perpetuar o, incluso, intensificar este sesgo.

Además, otro problema ético es la opacidad de estos modelos. Dado que estos funcionan como «cajas negras», puede ser difícil para los afectados, o incluso para los encargados de implementarlos, entender exactamente cómo el algoritmo llega a una determinada decisión. Esto dificulta la posibilidad de impugnar una decisión o de asegurar que la decisión se ha tomado de manera justa. El uso de algoritmos en estos contextos también plantea cuestiones de responsabilidad y consentimiento tales como: ¿quién es responsable cuando un algoritmo toma una decisión perjudicial? ¿Cómo podemos asegurar que las personas afectadas por estas decisiones hayan dado su consentimiento informado para este tipo de procesamiento de sus datos?

Nuevos desafíos

Estos son nuevos desafíos que la sociedad debe abordar ahora que el *machine learning* y sus aplicaciones ya están mucho más consolidadas, especialmente cuando se quieren aplicar en políticas u organismos públicos (Tabares-Soto et al, 2022: p. 184). Para el sector privado, el desafío también es similar: ¿cómo hacemos que estos modelos no generen perjuicios a quiénes son sometidos a esa suerte de “juicio” que estos mismos pueden tener?

A priori, el problema parece ser bastante obvio: le estamos derivando a una máquina una tarea que normalmente haría un humano, siendo que estas no razonan, solamente aplican probabilidad en base a lo previamente aprendido. Pero también hay otras cosas que considerar: quién (o quiénes) lo van a utilizar y para qué se quieren utilizar.

Así, la idea de tener “algoritmos éticos” nace para minimizar potenciales riesgos que puedan existir a la hora de aplicar estas herramientas. Aspectos tales como la transparencia, los niveles de sesgo y la equidad son claves para asegurar una correcta aplicación de estas herramientas dentro de la sociedad.

Cuando hablamos de transparencia, nos referimos a que deberíamos ser capaces de entender cómo un algoritmo toma sus decisiones. Esto implica que no deberíamos tener «cajas negras» que toman decisiones sin que podamos ver lo que hay dentro o por lo menos tener una cierta noción de cómo está funcionando el modelo. Por otro lado, el sesgo se refiere a cualquier tendencia sistemática en los resultados producidos por el algoritmo que favorece a ciertos grupos sobre otros. Por ejemplo, un algoritmo que siempre predice que las personas de un grupo demográfico específico tendrán un rendimiento más bajo en un trabajo, aunque esto no sea cierto, se dice que tiene un sesgo. La equidad, por último, se refiere a que todos los grupos de personas sean tratados de manera justa por el algoritmo. Esto significa que los resultados producidos por este no deberían discriminar injustamente a ningún grupo de personas.

Entonces, ¿qué clase de riesgos pueden haber? El ejemplo típico: si estamos usando alguna IA o un modelo de *machine learning* para diagnosticar a un paciente, es muy probable que existan varios casos en donde el modelo se equivoque en la predicción. No es un problema grande cuando el error es un falso positivo (es decir, cuando se le diagnostica a un paciente una enfermedad que en realidad no la tiene), pero sí puede ser grave cuando ocurre lo contrario.

En un estudio recientemente realizado sobre distintas bases de datos utilizadas para detectar el COVID-19 en imágenes de rayos X utilizando modelos de *Machine Learning* (Arias-Garzón et al, 2023), los investigadores encontraron que los desequilibrios en los datos y los sesgos en la edad y el sexo de los pacientes pueden hacer que el algoritmo no funcione bien, afectando la precisión y generalización de las predicciones.

Otro ejemplo puede ser el que fue mencionado anteriormente: si se utiliza para la toma de decisiones en un juicio puede terminar en que una persona inocente vaya a la cárcel. Incluso puede darse el caso de que un abogado cuente con una herramienta similar y que a partir de esta decida si le conviene o no defender a una persona. En este caso, el imputado puede verse perjudicado, porque el jurista podría optar por no defenderlo dado que puede ser un caso con altas posibilidades de perder.

Lo cierto es que estos problemas pueden darse por muchos motivos, pero cuando notamos sesgos o errores sistemáticos en las predicciones, es porque estamos frente a un problema de datos. Esto es una obviedad: los modelos aprenden a partir de la información que le entregamos y si estos cuentan con sesgos, entonces las máquinas aprenderán de ellos.

Estrategias para minimizar

Lo bueno es que existen formas de reducir estos problemas. Uno de los primeros pasos al empezar un proyecto de este tipo es evaluar sus posibles efectos, tanto buenos como malos. Esto significa entender qué podría pasar al aplicar la IA o el ML a un problema en particular. Para esto, podemos usar diversas herramientas y metodologías.

Un método sencillo para evaluar el impacto de un sistema automatizado es a través de cuestionarios. Por ejemplo, el gobierno de Canadá ha creado un cuestionario[1] que nos ayuda a entender cómo un sistema de decisión automatizada puede afectar diferentes aspectos de una situación.

También existen propuestas desde la academia para entender mejor los posibles sesgos en estos sistemas. Investigadores de las Universidades de Chicago y Carnegie Mellon sugieren que debemos distinguir entre los sistemas que pueden causar daño a las personas —modelos «punitivos»— y los que están diseñados para ayudar —modelos «asistenciales» (Saleiro et al, 2018: p. 5)—. Entender esta diferencia es clave para saber si lo que nos interesa es proteger a ciertos grupos o bien asegurarse que la mayoría pueda verse beneficiada por la aplicación de esta herramienta.

Además, hay formas de mejorar estos sistemas en la fase de recolección de datos, aunque este enfoque

puede ser difícil de lograr debido a la cantidad de información necesaria. Otra opción es ajustar u «optimizar» los estos sistemas antes o después de que se hayan entrenado con los datos. Sin embargo, esto puede reducir la eficacia general del sistema, por lo que se debe tener cuidado.

Por último, es esencial mantener una buena documentación del proceso. Esto significa tener registros claros de nuestros datos: dónde se obtuvieron, cómo se recolectaron y cómo se procesaron. También es útil mantener una documentación clara de cómo funciona el sistema. Google, por ejemplo, ha creado una herramienta llamada Model Cards[2] que ayuda a entender las características de un modelo de IA. Y existen otras herramientas que, sin entrar en tecnicismos, nos permiten entender de manera más intuitiva cómo el sistema llega a sus predicciones. Esta información es vital para que todos podamos entender, confiar y, si es necesario, cuestionar las decisiones tomadas por estos sistemas.

Las máquinas no son humanos

De todas maneras, también es necesario recordar que las máquinas no tienen la habilidad para replicar muchas de las capacidades humanas. Por ejemplo, en el campo del derecho, un abogado debe combinar el pensamiento abstracto con habilidades para resolver problemas en situaciones de alta incertidumbre, tanto en el ámbito legal como factual; algo que evidentemente un modelo de clasificación no necesariamente puede hacer todavía (Surden, 2014: p. 87).

Dicho de otra forma, los modelos no pueden “razonar” de la misma manera que un ser humano. Incluso los modelos de lenguaje de gran tamaño —tales como ChatGPT o Bard— que tan populares se han vuelto en el último tiempo, no son capaces de hacerlo porque detrás de esa «caja negra» solo hay probabilidades y cálculos entre vectores y matrices.

Finalmente, es importante recordar que la responsabilidad de la aplicación ética de la IA y el ML recae en nosotros, los humanos. Mientras continuamos avanzando en la era de la inteligencia artificial cada uno de nosotros, ya sea como investigadores, desarrolladores, legisladores o simplemente como ciudadanos, tenemos un papel importante en la conformación de una sociedad digital justa y equitativa.

De hecho, hay otras dimensiones a tener en cuenta: el cómo estos se insertan en nuestra sociedad (por ejemplo, el impacto en nuestros trabajos o su aceptación por parte de las personas) o cómo estos son regulados por las autoridades, especialmente cuando hablamos de privacidad de la información y la toma de responsabilidades, entre otros (Wirtz et al., 2019: p. 12). Así, la construcción de algoritmos éticos y responsables es un compromiso que debemos asumir juntos para aprovechar los beneficios de la IA y el ML, minimizando al mismo tiempo sus riesgos y desafíos.

Arias-Garzón, D., Tabares-Soto, R., Bernal-Salcedo, J., y Ruz, G.A.: «Biases associated with database structure for COVID-19 detection in X-ray images» en *Scientific Reports,* 2023. Disponible en: https://doi.org/10.1038/s41598-023-30174-1

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K.T., y Ghani, R.: «Aequitas: A Bias and Fairness Audit Toolkit» en *Cornell University,* 2019. Disponible en: https://arxiv.org/abs/1811.05577

Surden, H. :“Machine Learning and Law” en *Washington Law Review,* 2014. Disponible en: https://digitalcommons.law.uw.edu/wlr/vol89/iss1/5/

Wirtz, B.W., Weyerer, J.C. y Geyer, C.: «Artificial Intelligence and the Public Sector—Applications and Challenges» en *International Journal of Public Administration,* 2019. Disponible en:

https://www.scinapse.io/papers/2884686506

Tabares-Soto, R., Bernal-Salcedo, J., García-Arias, Z.N., Ortega-Bolaños, R., Hermosilla, M.P., Arteaga-Arteaga, H.B., y Ruz, G.A.: "Analysis of Ethical Development for Public Policies in the Acquisition of AI-Based Systems". En: Fudge, T.P. Exploring Ethical Problems in Today's Technological World. Hershey, PA, USA: IGI Global, 2022.