Traducción
Confiar en las máquinas frente a los humanos. Debemos entender la diferencia
Foro Económico Mundial

Versión original en: https://www.weforum.org/agenda/2021/01/trusting-machines-versus-humans-we-must-understand-the-difference/

Históricamente, los seres humanos tardan mucho tiempo en confiar en la última ola de tecnología de máquinas.

En escenarios que involucran daño físico, las personas tienden a ver las máquinas como más dañinas que los humanos que realizan las mismas acciones.

Es importante que combinemos nuestro interés en cómo deben comportarse las máquinas, bajo un entendimiento de cómo las juzgamos.

Recientemente, las máquinas de votación han sido objeto de controversias. Y, sin embargo, la aversión de la gente a las máquinas no es nada nuevo.

Hace unos 500 años, la imprenta estaba siendo demonizada como un dispositivo satánico. El equivalente actual, la inteligencia artificial, se critica habitualmente como fuente de desempleo y prejuicios.

Pero, ¿está justificada toda rabia?

Los académicos que estudian las reacciones de las personas a las máquinas están comenzando a aprender cuándo y por qué juzgamos a los humanos y las máquinas de manera diferente.

Imagínese un automóvil que se desvía para evitar la caída de un árbol y, al hacerlo, atropella a un peatón. ¿Las personas juzgan esta acción de manera diferente si creen que fue la acción de un automóvil autónomo y no la de un humano?

En mi último libro, How Humans Judge Machines, mis coautores y yo les pedimos a más de 6.000 estadounidenses que reaccionaran ante escenarios como este, utilizando la configuración de un ensayo clínico.

La mitad de nuestros sujetos vieron solo escenarios que involucraron acciones humanas, mientras que la otra mitad evaluó solo escenarios que involucraron acciones de máquinas. Esto nos permitió explorar cuándo y por qué las personas juzgan a los humanos y las máquinas de manera diferente.

Mala máquina, buen ser humano

En el accidente automovilístico antes mencionado, la gente juzga la acción del automóvil autónomo como más dañina e inmoral, a pesar de que la acción realizada por el humano fue exactamente la misma.

En otro escenario, consideramos un sistema de respuesta a emergencias que reacciona ante un tsunami. A algunas personas se les dijo que la ciudad fue evacuada con éxito. A otros se les dijo que el esfuerzo de evacuación fracasó.

Nuestros resultados mostraron que, en este caso, las máquinas también recibieron el extremo más corto del palo. De hecho, si el esfuerzo de rescate fallaba, la gente evaluaba negativamente la acción de la máquina y la del humano positivamente.

Los datos mostraron que las personas calificaron la acción de la máquina como significativamente más dañina y menos moral, y también informaron que querían contratar al humano, pero no a la máquina.

¿Las máquinas siempre obtienen la pajita más corta?

Durante mucho tiempo, los estudiosos han sabido que las personas tienen aversión a los algoritmos. Incluso cuando los algoritmos son mejores para pronosticar que los humanos, la gente tiende a elegir pronosticadores humanos. Este fenómeno se conoce como aversión a los algoritmos y puede resultar costoso en un mundo en el que las pequeñas diferencias en la precisión predictiva son importantes.

En un artículo reciente, Berkeley Dietvorst, Joseph Simmons y Cade Massey exploraron las aversiones a los algoritmos utilizando cinco experimentos mediante los cuales las personas podían vincular una recompensa monetaria a las predicciones hechas por ellos mismos, otra persona o un modelo.

Si bien existe la necesidad de que las máquinas sean transparentes, debe complementarse con el entendimiento de que la transparencia puede, en última instancia, predisponer a las personas contra las máquinas.

Máquinas injustas

Pero hay casos en los que las personas valoran más a las máquinas que a los humanos, aunque solo ligeramente. Estos son escenarios morales que involucran violaciones de la justicia y la lealtad, que también se perciben como altamente intencionales cuando las realiza un humano.

Considere un robot versus un humano, ambos escribiendo letras para un sello discográfico. Imagina que una investigación descubre que estas letras plagian el trabajo de artistas menos

conocidos. Cuando presentamos a las personas con este escenario, descubrimos que juzgaban la acción del ser humano como más dañina y menos moral que la de la máquina.

Obtuvimos resultados similares para otros escenarios que involucran equidad, como evaluaciones de recursos humanos sesgadas y sistemas de admisión universitaria.

A la gente ciertamente no le gustan los humanos o las máquinas sesgados, pero cuando probamos su repudio experimentalmente, las personas califican los sesgos humanos como un poco más dañinos y menos morales que los de las máquinas.

Estamos pasando de una era de imponer normas sobre el comportamiento de las máquinas a una de descubrir leyes que no nos dicen cómo deben comportarse las máquinas, sino cómo las juzgamos. Y la primera regla es poderosa y simple: la gente juzga a los humanos por sus intenciones y a las máquinas por sus resultados.

Entonces, ¿podemos confiar en las máquinas? ¿Incluso queremos hacerlo? Puede que no sea posible una respuesta general a preguntas tan atrevidas, pero la investigación actual está comenzando a darnos alguna orientación.

César A. Hidalgo es el autor de How Humans Judge Machines, un libro revisado por pares de MIT Press que se puede leer gratis en judgingmachines.com. Tiene una cátedra en el Instituto de Inteligencia Artificial y Natural (ANITI) de la Universidad de Toulouse, y nombramientos en la Universidad de Manchester y la Universidad de Harvard.