Published on *LET* (http://let.iiec.unam.mx)

Unmasking A.I.'s Bias Problem. Artificial intelligence can imitate and enhance human decision-making - and amplify human prejudices. Can Big Tech tackle A.I.'s discrimination problem?

Enviado por cristobalrn en Vie, 12/14/2018 - 23:40
Cita:

Vanian, Jonathan [2018], “Unmasking A.I.'s Bias Problem. Artificial intelligence can imitate and enhance human decision-making -- and amplify human prejudices. Can Big Tech tackle A.I.'s discrimination problem?”,*Fortune*, 25 de junio, http://fortune.com/longform/ai-bias-problem/ [1]

Fuente:
Otra
Fecha de publicación:
Lunes, Junio 25, 2018
Revista descriptores:
Competencia mundial. Disputa hegemónica [2]
Estudios de caso: actividades - empresas [3]
Formas de la competencia entre grandes empresas [4]
Fronteras del capital [5]
Relaciones entre empresas estados y sociedad [6]
Tecnologías militares - tecnologías de uso dual [7]
Tema:
Los sesgos y prejuicios que pueden reproducir las tecnologías de inteligencia artificial
Idea principal:

En marzo de 2016 Microsoft dio a conocer a Tay, su “chatbot social” con inteligencia artificial (IA) que respondería a preguntas escritas por las personas a través de redes sociales. A diferencia de otros programas que contestan preguntas de los usuarios en diversas páginas web, Tay fue diseñada para conversar de una manera más sofisticada, con una dimensión “emocional”. Al responder, Tay podría mostrar sentido del humor y responder de forma amistosa. En su diseño, Tay incorporó una de las características más importantes de la IA: la capacidad de volverse más inteligente, más eficiente y más útil conforme pasa el tiempo y conforme recibe más información. Por ello, la capacidad de Tay para responder sería mejor conforme más personas entablaran una conversación con ella.

Lo que nadie previó fue el ataque de los *trolls*. Al saber que Tay aprendería e imitaría el lenguaje de las personas con las que se relacionaba, los *trolls* inundaron su perfil de Twitter con mensajes racistas, homofóbicos y ofensivos. Después de unas horas, Tay comenzó a publicar contenido ofensivo en Twitter. Menos de 24 horas después de su debut, Microsoft dio de baja su cuenta y ofreció disculpas por lo ocurrido. Lo más sorprendente del caso es que este desastre

público tomó por sorpresa al equipo de investigación de Microsoft.

Después de lo sucedido con Tay, Eric Horvitz, el director de investigación e inteligencia artificial de Microsoft, pidió a su equipo de trabajo que indagara qué salió mal con la función de "procesamiento de lenguaje natural" de su chatbot. El problema consistió en que se pasaron por alto protocolos que hacen que los programas pongan en una lista negra las palabras ofensivas. Así, Tay no tuvo barreras que limitaran el tipo de información que absorbía y de la que aprendía.

La desastrosa experiencia con Tay ha hecho que Microsoft y otras empresas aprendan del ejemplo y desarrollen chatbots más sofisticados. En Microsoft son más cuidadosos en relación con el comportamiento de sus bots conforme aumentan en escala y reconocen que el monitoreo de las tecnologías de IA nunca termina.

Aunque Tay es un programa relativamente simple dentro del espectro de la IA, es representativo de las dificultades, inconvenientes y potencialidades que estas tecnologías traen consigo. Los errores de Tay también son útiles para ejemplificar los mayores peligros que le quitan el sueño a los tecnólogos, "incluso mientras el mundo de los negocios se prepara para confiar su futuro a esta nueva tecnología revolucionaria".

"Pocos dudan que nos encontramos al borde de una carrera corporativa por ubicarse en la cima de la IA". La consultora IDC predice que hacia el año 2021 las organizaciones invertirán 52 mil millones de dólares anualmente en productos relacionados con IA. Los economistas y analistas piensan que las empresas tendrán muchos miles de millones de dólares más en ahorros y ganancias a partir de esas inversiones. Algunos de estos ahorros y ganancias provendrán de la reducción en el número de empleados, pero muchos más tendrán su origen en el aumento de la eficiencia al conectar a los productores con los consumidores, las medicinas para los pacientes, las soluciones para los problemas. La consultora PwC estima que la IA podría contribuir con 15.7 billones [trillion] de dólares a la economía mundial en 2030, más del producto combinado de China e India en la actualidad.

El auge de la inteligencia artificial ha sido impulsado en buena medida por los grandes avances en la tecnología de "aprendizaje profundo" [deep learning]. Con esta tecnología, las empresas están alimentando sus redes computacionales con enormes cantidades de información con el objetivo de que puedan reconocer patrones con mayor velocidad y precisión que los humanos. Las empresas más importantes que ya están utilizando tecnologías de aprendizaje profundo en sus servicios son Facebook, Google, Microsoft, Amazon e IBM. Se espera que en el futuro próximo, empresas de todos los tamaños utilicen software de aprendizaje profundo que al hacer minería de datos encuentre diamantes que la vista y la inteligencia humana no podrían detectar.

A pesar de su enorme potencial al procesar información, los sistemas con IA "tienen un lado oscuro. Sus decisiones son tan buenas como la información con la que los humanos los alimentan". Los programadores están aprendiendo que la información con la que se entrenan y alimentan los sistemas de aprendizaje profundo no es neutral. Esta información puede contener y reflejar sesgos [biases] conscientes o inconscientes. Los datos pueden estar cargados de tendencias y patrones que tienen siglos de antigüedad. Por ejemplo, un algoritmo puede analizar una base de datos con información histórica y concluir que los hombres blancos tienen mayor probabilidad de ser exitosos como directores generales de una empresa. Estos softwares aún no pueden ser programados para reconocer que hasta hace poco apenas había posibilidad de

que alguien que no fuera un hombre blanco llegase a ser director ejecutivo. Si se permitiera que los sistemas con IA automatizaran las recomendaciones de empleos, es muy probable que amplificaran sesgos de los que la sociedad no se sentiría orgullosa. “La ceguera a los sesgos es un defecto fundamental de esta tecnología”.

Esto se debe a que los algoritmos más poderosos han sido diseñados para llevar a cabo tareas específicas, no para actuar conforme a definiciones de justicia. La IA convierte la información en decisiones a una velocidad sin precedentes. Pero lo que los científicos y los estudiosos de la ética están aprendiendo es que esa información no necesariamente es justa.

Un factor que vuelve más enredado el problema es que el aprendizaje profundo es más complejo que los algoritmos que le precedieron, por lo que los productos de IA pueden “comportarse” de maneras no deseadas ni planeadas por sus creadores. Adicionalmente, las “cajas negras” que crea la secrecía que los programadores de estos sistemas y algoritmos tienen respecto a su propiedad intelectual, hace más difícil para los reguladores saber qué problemas podría traer un sistema particular.

Los defectos de los sistemas de IA y aprendizaje profundo son más cercanos de lo que la mayoría de la gente piensa. El ejemplo más importante fue la difusión de noticias falsas en Facebook en vísperas de la elección presidencial de Estados Unidos en 2016. Los algoritmos utilizados por esa red social no fueron diseñados para distinguir entre el contenido falso y el verdadero, sino para mostrar contenido personalizado a los usuarios según sus preferencias, con base en sus búsquedas y en las de personas con gustos similares. El resultado es que las páginas de inicio de millones de personas fueron inundadas de noticias falsas.

Una de las principales preocupaciones de los investigadores es la forma en que las aplicaciones de la IA podrían poner en desventaja a grupos minoritarios al leer e interpretar –o malinterpretar– la información colectiva. La investigadora Timnit Gebru pone como ejemplo el mercado de seguros. En el caso de los accidentes viales, una serie de datos podría mostrar que hay mayor probabilidad de que los accidentes sucedan en los barrios pobres, donde la mayor densidad de población crea mayores posibilidades de choques. Los barrios pobres también suelen ser habitados por una mayor proporción de grupos “minoritarios” respecto de los barrios ricos. Al analizar una serie de datos con esas correlaciones, un programa de aprendizaje profundo señalaría que hay una relación estrecha entre pertenecer a una “minoría” y tener accidentes automovilísticos. A partir de ese sesgo, el sistema de IA concluiría que en un choque entre múltiples personas la culpa corresponde al conductor que pertenece a un “grupo minoritario” y recomendaría cobrar primas más altas a los conductores de estos grupos, independientemente de su registro.

El ejemplo anterior ilustra cómo un sistema con IA puede interpretar información aparentemente neutral (datos sobre dónde ocurren los accidentes automovilísticos) de formas que crean desventajas o discriminaciones (precios más altos en las primas de seguro a los “grupos minoritarios” basándose en su perfil racial o étnico, independientemente de dónde viven). Además, a diferencia de las generaciones previas de algoritmos, a los sistemas con IA más avanzados se les ha dado la capacidad de tomar decisiones legalmente significativas (por ejemplo, en el ámbito médico y penal). Hasta el momento, se ha prestado poca atención a las implicaciones legales de la toma de decisiones por parte de los sistemas con IA.

Mientras las grandes empresas tecnológicas se preparan para incorporar la tecnología de aprendizaje profundo en el software comercial destinado a sus clientes, las preguntas sobre los riesgos e implicaciones de la IA están pasando a primer plano. Un ejemplo de lo anterior es el surgimiento del grupo Aether de Microsoft, un grupo que discute sobre ética e IA en la ingeniería y la investigación y que incorpora representantes de la academia, la sociedad civil, el gobierno y la industria.

Para el autor del artículo, el principal reto que plantean las tecnologías con IA no es técnico, sino filosófico y está relacionado con la naturaleza humana. Es difícil para los científicos y los programadores codificar la justicia en el software, pues la concepción de lo que es justo puede variar de persona a persona y a lo largo del tiempo.

Otro ejemplo de los sesgos a los que son propensas las tecnologías con IA lo ofrece un concurso de belleza organizado por científicos rusos en 2016. El concurso consistía en que las personas enviaran fotografías y un sistema con IA juzgaría su belleza basándose en factores como la simetría de sus rostros. De las 44 personas ganadoras que la máquina eligió entre miles de fotografías enviadas, sólo una tenía piel oscura. La elección provocó un escándalo internacional. Los organizadores atribuyeron el sesgo a que las computadoras habían sido entrenadas con series de imágenes que no contenían muchas fotografías de personas de color. Debido a este sesgo por omisión, las computadoras tendían a ignorar a las personas con piel de color y a considerar más bellas a las que tenían piel clara.

Estos defectos de los algoritmos pueden parecer triviales tratándose de un concurso de belleza, pero esas tecnologías pueden ser utilizadas en situaciones más delicadas y tener consecuencias más graves. ¿Qué pasaría si, por ejemplo, un vehículo autónomo no reconociera cuando “viera” a una persona de color?

Empresas como Microsoft e IBM han reconocido el problema y han expresado que están tomando medidas para mejorar sus tecnologías de reconocimiento de imágenes con el objetivo de mitigar los sesgos.

Amazon ha tenido que lidiar con otro sesgo en el que incurren las tecnologías con IA, el sesgo por problemas de muestreo, al desarrollar algoritmos para separar la fruta podrida de la que está en buen estado. Los algoritmos de reconocimiento visual suelen ser entrenados para identificar cómo “deben” verse los objetos -en este caso, las fresas- a partir de una gran base de datos con imágenes. Pero las imágenes de frutas podridas son menos comunes. Esto ha generado dificultades, pues los algoritmos tienden a ignorar o a minimizar los casos atípicos. Para hacer frente a este sesgo, la empresa minorista en línea está probando una técnica

llamada “sobremuestreo”, que consiste en asignar un mayor peso estadístico a los datos subrepresentados (en este caso, la fruta podrida), con el objetivo de que el algoritmo ponga más atención a las frutas en mal estado.

Ralf Herbrich, el director de IA de Amazon, señala que la técnica de “sobremuestreo” puede ser aplicada también a algoritmos que estudian humanos para minimizar o evitar los sesgos. “Para asegurarte de que un algoritmo utilizado para reconocer rostros en fotografías no discrimine o ignore a las personas de color, o a las personas mayores, o a las personas con sobrepeso, puedes añadir peso [estadístico] a las fotos de esos individuos” para que el algoritmo les preste más atención.

Otros ingenieros están buscando mitigar los sesgos a partir de volver más incluyentes y libres de prejuicios los datos con que se entrenan los algoritmos. Para ello, es importante que en el diseño y “etiquetado” de los datos se busque incluir la mayor cantidad de puntos de vista posibles, con miras a que todas las personas sean representadas. Empresas como Google y iMerit están llevando a cabo esfuerzos en este sentido.

Aunque los científicos y desarrolladores reconocen los defectos de los sistemas con IA, consideran que los beneficios potenciales -sociales y financieros- de estas tecnologías son mayores que sus peligros, lo que justifica seguir adelante. También confían en que la creación de grupos multidisciplinarios como Aether ayudarán a las empresas a solucionar los sesgos antes de que provoquen problemas públicos.

Los ejecutivos de las principales empresas que están innovando en IA consideran, además, que la existencia de sesgos que se pueden mitigar no es un motivo suficiente para no continuar con investigaciones que podrían mejorar la vida de muchas personas (en el ámbito médico, por ejemplo, al reducir los errores de diagnóstico).

Otro aspecto en el que puede haber grandes avances es en el de la regulación. Más transparencia y apertura respecto de los datos y los algoritmos ayudaría a los reguladores a identificar los sesgos con mayor facilidad y a evitar que causen problemas. Es importante que haya la mayor transparencia y rigurosidad al probar los algoritmos cuando se trata de temas delicados como determinar si una persona puede acceder a un seguro o si debe ir a la cárcel.

Hoy son pocas las personas que piensan que la IA es una tecnología infalible y neutral. En sí mismo, concluye el artículo, eso es ya “un signo de progreso”.

Datos cruciales:

La consultora IDC predice que hacia el año 2021 las organizaciones invertirán 52 mil millones de dólares anualmente en productos relacionados con IA.

La consultora PwC estima que la IA podría contribuir con 15.7 billones [trillion] de dólares a la economía mundial en 2030, más del producto combinado de China e India en la actualidad.

Nexo con el tema que estudiamos:

Las tecnologías con inteligencia artificial (como los vehículos autónomos, los programas que

utilizan las aseguradoras, el software médico más avanzado, entre otros) "actúan" y toman decisiones en función de los algoritmos con que fueron diseñados y se alimentan con datos que, aunque parezcan neutrales, pueden no serlo. Por esta razón, estas tecnologías son propensas a reproducir los sesgos del programador o de los datos con que se alimentan. Es necesario que se generen acuerdos vinculantes entre corporaciones y estados sobre el uso de estas nuevas tecnologías antes que su uso se generalice, de tal forma que eviten reproducir sesgos racistas, sexistas o de otro tipo.

Links
[1] http://fortune.com/longform/ai-bias-problem/
[2] http://let.iiec.unam.mx/taxonomy/term/12
[3] http://let.iiec.unam.mx/taxonomy/term/16
[4] http://let.iiec.unam.mx/taxonomy/term/17
[5] http://let.iiec.unam.mx/taxonomy/term/18
[6] http://let.iiec.unam.mx/taxonomy/term/20
[7] http://let.iiec.unam.mx/descriptores-let/tecnolog%C3%ADas-militares-tecnolog%C3%ADas-de-uso-dual