Justicia algorítmica

La justicia algorítmica estudia las propiedades y el impacto social de los algoritmos que utilizamos y que afectan directamente a las personas, garantizando que se respeten los derechos fundamentales. Para llevarla a la práctica es indispensable la colaboración de la informática, que aporta herramientas para cuantificar, vigilar y limitar sesgos y abusos producidos por los sistemas de inteligencia artificial y mejorar sus aplicaciones beneficiosas en nuestra sociedad.

[ILUSTRACIÓN: MUSHAKESA/ ISTOCK]

Antes de la invención de los ordenadores y de la inteligencia artificial, los algoritmos ya acompañaban nuestras vidas: por ejemplo, cuando seguimos una receta de cocina al pie de la letra, o cuando efectuamos operaciones matemáticas a mano. Los algoritmos consisten en una sucesión de pasos bien definidos y establecidos que nos permiten convertir un input a un output. Los ordenadores modernos son capaces de ejecutarlos, a partir de instrucciones expresadas en lenguajes de programación. Desde un inicio, los informáticos hemos estudiado las propiedades de los algoritmos que producimos de forma rigurosa y teórica, muchas veces usando herramientas matemáticas. Por ejemplo: ¿cuál es la eficiencia de un algoritmo que ordena correctamente listas de números? ¿Cuánta memoria necesita?

Con el paso del tiempo y, en particular, estos últimos años, los algoritmos se han ido sofisticando, empujados por dos motivos principales: por un lado, empezamos a desarrollar sistemas capaces de aprender directamente a partir de ejemplos, sin necesidad de programar explícitamente cada instrucción. En segundo lugar, pudimos empezar a recoger cantidades masivas de datos para su entrenamiento, lo que a su vez potenció su capacidad de generalización y precisión.

Este nivel de sofisticación ha llevado a un cambio de paradigma en la relación entre algoritmos y sociedad: hoy en día, en lugar de aplicarlos únicamente a objetos como grafos o matrices (por ejemplo, para encontrar la distancia más corta entre dos ciudades en Google Maps), también los usamos para tomar decisiones sobre personas. Y no cualquier tipo de decisiones: en muchos casos, la inteligencia artificial puede alterar nuestras vidas, por ejemplo, cuando se emplea para decidir si contratar a alguien o si conceder una hipoteca. También se ha intensificado su uso en ámbitos públicos, desde la educación al sistema jurídico o la asignación de recursos sociales. De forma similar, en algunos países se usan para alertar sobre casos de violencia de género o para decidir qué barrios requieren más vigilancia policial.

Decisiones sobre las personas

A medida que delegamos la toma de decisiones importantes—particularmente, en el ámbito social—a sistemas algorítmicos, ha surgido la necesidad de examinar nuevas propiedades más allá de las métricas "clásicas", como la eficiencia o el uso de memoria. Por ejemplo, ¿presenta el algoritmo una precisión significativamente menor en ciertos subgrupos de la población? Estos sesgos pueden originarse tanto por la presencia de patrones históricos discriminatorios en los datos usados para entrenar el modelo, como por la falta de criterios de optimización adecuados en el diseño del algoritmo, entre muchos otros motivos.

Si usamos algoritmos sesgados, podemos automatizar desigualdades existentes o excluir a individuos de

ciertas oportunidades injustamente. En el Reino Unido, por ejemplo, se retiró un sistema de asignación de notas para el acceso a la universidad porque penalizaba sistemáticamente a estudiantes procedentes de escuelas con menos recursos. En España, se han reportado casos de algoritmos que causan cierres erróneos de cuentas bancarias y decisiones arbitrarias sobre la probabilidad de reincidencia en las prisiones. En Estados Unidos, se han empleado modelos de evaluación de riesgo financiero que otorgan puntuaciones más bajas a personas que viven en ciertos códigos postales.

También debemos analizar cuán seguro está el algoritmo de sus propias decisiones. ¿Son estas consistentes o, por el contrario, arbitrarias? Puede ocurrir, por ejemplo, que a partir del mismo conjunto de datos, un modelo estime un alto riesgo de enfermedad para una persona, mientras que otro modelo prediga un riesgo bajo. ¿Qué deberíamos hacer entonces? Responder a esta pregunta implica pensar detenidamente qué significa que una decisión sea justa, explicable o revisable en el contexto de un sistema automatizado.

Las tripas de la programación

Varias líneas de investigación actuales en informática se centran precisamente en definir matemáticamente, analizar y cuantificar nociones como el sesgo y la estimación de la incertidumbre algorítmica. Nuevas técnicas, como el "multicalibraje", nos permiten construir algoritmos que alcanzan alta precisión en todos los subgrupos de una colección de grupos de la población de interés simultáneamente. También hace posible procesar un algoritmo que presenta sesgos para corregirlos. Esto último se ha llevado a la práctica para reducir el error que se comete cuando se entrenan algoritmos con datos médicos que provienen de un estudio clínico que no cuenta con suficiente representación de individuos de grupos minoritarios. Otros campos relacionados de la informática estudian cómo estimar la certeza del algoritmo para las propias predicciones que realiza.

Siempre deberíamos poder asegurar lo siguiente: si una decisión se considera injusta o ilegal cuando la toma un ser humano, también deberíamos poder identificarla como tal cuando la toma un ordenador, y poder actuar en consecuencia.

Si una decisión se considera injusta o ilegal cuando la toma un ser humano, también deberíamos poder identificarla como tal cuando la toma un ordenador, y poder actuar en consecuencia

Aunque la motivación de estos problemas suele tener un origen social, nos lleva a profundizar nuestra comprensión de los sistemas algorítmicos que nos rodean, así como el nivel de rigor que exigimos en su análisis. Así, los avances informáticos, más allá de su motivación social, conllevan también desarrollos teóricos significativos. Por ejemplo, los avances en "multicalibraje" nos han permitido mejorar teoremas matemáticos en el campo de la complejidad computacional que no se habían progresado desde los años 1990.

El mundo que vemos

Hoy en día, los algoritmos no solo toman decisiones sobre nuestro presente y futuro, sino que también configuran la percepción que tenemos de nuestro entorno, así como las opciones y la información que creemos tener a nuestro alcance. Determinan las ofertas de trabajo que vemos en LinkedIn, los anuncios y publicaciones sugeridas en Instagram, las noticias que leemos en redes sociales, los productos que compramos en Amazon.

Toda esta información nos llega ya filtrada, priorizada y organizada por criterios que no siempre entendemos y que pueden ser potencialmente perjudiciales (por ejemplo, diseñados para maximizar nuestra adicción al contenido digital). Al mismo tiempo, pueden condicionar profundamente nuestra percepción del mundo y elecciones cotidianas. Por lo tanto, resulta crucial estudiar y comprender cómo se forman y evolucionan estos

rankings, redes y plataformas.

En este sentido, investigadores de la Universidad de Harvard han establecido un instituto para estudiar cómo evitar casos de discriminación algorítmica en plataformas de contratación como LinkedIn, mientras un equipo de la Universidad de Princeton colabora con trabajadores de plataformas como Uber o DoorDash para analizar los algoritmos opacos que se usan para determinar sus pagos y horarios. Otros expertos se enfocan en el efecto causal que tienen las predicciones de un algoritmo en el comportamiento futuro de sus usuarios, cómo los usuarios pueden juntarse para alterar el algoritmo de la plataforma que se aplica sobre ellos, cómo evaluar el efecto de los anuncios políticos personalizados de Facebook durante campañas electorales o formas para integrar conocimiento humano de un experto en las decisiones que toma el algoritmo de manera efectiva.

Por otra parte, se están revelando fenómenos como la colusión algorítmica –algoritmos que se comportan de forma coordinada para limitar la competencia en el entorno comercial–, donde sistemas autónomos de fijación de precio aprenden a mantener precios artificialmente altos para los productos que se venden, perjudicando a los consumidores.

ChatGPT puede modificar sus respuestas en función del género u otras características inferidas del usuario; por ejemplo, sugiriendo oficios estereotipados o de menor salario según el género

Todo este tipo de problemas se está multiplicando con los avances de la inteligencia artificial generativa. Se ha demostrado repetidamente cómo ChatGPT puede modificar sus respuestas en función del género u otras características inferidas del usuario; por ejemplo, sugiriendo oficios estereotipados o de menor salario según el género de la persona a la que se supone que va dirigida la respuesta. A causa del gran número de datos que necesitan modelos como ChatGPT para ser entrenados, otro riesgo que se está expandiendo de forma dramática es la recolección de datos personales, que conlleva nuevos y crecientes problemas de privacidad de datos. Nos faltan garantías claras sobre cómo se almacenan y se procesan nuestros datos, además de avances tecnológicos para asegurar que todo este proceso puede realmente mantener el anonimato, ya que a nivel técnico es muy complicado asegurar que un modelo no revele datos privados a adversarios, tal y como demuestran recientes investigaciones.

Informática para llevar la teoría a la práctica

En resumen, cuando hablamos de justicia algorítmica nos referimos al estudio de las propiedades y del impacto social de los algoritmos, ya sea porque deciden sobre el precio de los alimentos que compramos, sobre el acceso a servicios esenciales, sobre oportunidades laborales o sobre libertades individuales.

La justicia algorítmica se refiere al estudio de las propiedades y del impacto social de los algoritmos

Naturalmente, se trata de un problema interdisciplinar que no es posible resolver únicamente desde el campo de la informática. Desde hace años, numerosos juristas, filósofos, expertos en regulación, entre otros, trabajan en este campo.

No obstante, la investigación que se realiza desde el propio campo de la informática es fundamental: no podemos llevar estos principios éticos o legales de la teoría a la práctica si no contamos con formas de cuantificarlos matemáticamente y de analizarlos directamente en los sistemas algorítmicos. Los problemas sociales que conllevan ciertos usos de la inteligencia artificial no provienen solo de una mala aplicación de la tecnología, sino que muchas veces es la tecnología misma que no está suficientemente desarrollada. Necesitamos pararnos a pensar detenidamente sobre qué consecuencias tendrá un sistema algorítmico una

vez lo soltemos en nuestra sociedad.

Muchos nos hemos tomado esta responsabilidad con la seriedad que merece: en los últimos años, se han consolidado congresos y revistas académicas dedicadas exclusivamente al estudio de estas temáticas, y la mayoría de las principales empresas tecnológicas cuentan hoy con equipos especializados encargados de realizar evaluaciones y auditorías algorítmicas de manera interna. Debemos asegurar que la presencia de estas iniciativas sea robusta en el campo académico –donde la investigación debería permanecer libre de intereses financieros–, así como en el sector público –donde a veces se implementan algoritmos de forma apresurada, mal diseñada y sin suficiente evaluación previa–.

Como ciudadanos, es crucial que desafiemos la narrativa muy común hoy en día que presenta a la inteligencia artificial como un sistema mágico, infranqueable e indescifrable. Esta visión permite a las grandes empresas y otras instituciones usar sistemas algorítmicos que nos afectan directamente sin tener que asumir la responsabilidad que conlleva ejercer tal poder sobre nuestras vidas.

Como ocurre con cualquier otra tecnología, debemos exigir el máximo nivel de rigor científico, de modo que las decisiones automatizadas siempre se tomen de forma responsable y segura. Solo de esta forma podremos integrar la inteligencia artificial en nuestra sociedad sin renunciar a los principios fundamentales de justicia y transparencia.

Alur, R., Raghavan, M., & Shah, D. *"Human expertise in algorithmic prediction"*. En: *Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J.* & *C. Zhang (eds). Advances in Neural Information Processing Systems. NeurIPS Proceedings*, 2024. Vol. *37.*

Barocas, S., Hardt, M., & Narayanan, A. (2023): *Fairness and machine learning: Limitations and opportunities*. Massachusetts, The MIT press.

Baumann, J., & Mendler-Dünner, C. *Algorithmic Collective Action in Recommender Systems: Promoting Songs by Reordering Playlists*. En: *The Thirty-eighth Annual Conference on Neural Information Processing Systems. NeurIPS Proceedings, 2024*. Disponible en: https://proceedings.neurips.cc/paper_files/paper/2024/file/d79792543133425ff79513c147dc8881-Paper-Conference.pdf

Hébert-Johnson, U., Kim, M., Reingold, O., & Rothblum, G. *Multicalibration: Calibration for the (computationally-identifiable) masses.* En: *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018. Vol. 80. Disponble en: http://proceedings.mlr.press/v80/hebert-johnson18a/hebert-johnson18a.pdf