Technology Policy Council

Europe Technology Policy Committee

U.S. Technology Policy Committee

October 26, 2022

DECLARACIÓN DE PRINCIPIOS PARA SISTEMAS ALGORÍTMICOS RESPONSABLES[1, 2]

Los sistemas algorítmicos, a menudo basados en inteligencia artificial (IA),[3] se están utilizando cada vez más por administraciones públicas y empresas para tomar o recomendar decisiones que tienen efectos de gran alcance en las personas, las organizaciones y la sociedad. Muchas decisiones de empleo, crediticias, de acceso a la educación, de atención sanitaria e incluso en el ámbito de la justicia penal son tomadas por máquinas, a menudo sin una revisión sustancial adicional por parte del ser humano.

Si bien los sistemas algorítmicos son prometedores para lograr que la sociedad sea más equitativa, inclusiva y eficiente, sus resultados no surgen directamente de la automatización. Al igual que las decisiones tomadas por el ser humano, las decisiones tomadas por máquinas pueden también no respetar los derechos de las personas y dar lugar a una discriminación dañina y otros efectos negativos. Es imperativo, por lo tanto, que los sistemas algorítmicos cumplan plenamente con las normas legales, éticas y científicas establecidas y que los riesgos de su uso sean proporcionales a los problemas específicos que se quieran abordar.

Un algoritmo es un conjunto autónomo de operaciones sucesivas que se utiliza para realizar cálculos, procesamiento de datos y tareas de razonamiento automatizado. Muchos algoritmos de IA se basan en modelos estadísticos que son "enseñados" o "entrenados" a partir de conjuntos de datos mediante el uso de aprendizaje automático (Machine Learning o ML). Otros están basados en análisis: el descubrimiento, la interpretación y la comunicación de patrones significativos en los datos.

---

[1]Este documento actualiza y está basado en la declaración conjunta de 2017 de los comités de política tecnológica de ACM Europa y EE. UU. sobre transparencia algorítmica y rendición de cuentas. La elaboración de este documento fue liderada por los autores Ricardo Baeza-Yates y Jeanna Matthews. Hicieron importantes contribuciones Vijay Chidambaram, Simson Garfinkel, Carlos E. Jimenez-Gomez, Bran Knowles, Arnon Rosenthal, Ben Schneiderman, Stuart Shapiro y Alejandro Saucedo. Comentarios y otro tipo de apoyo fueron proporcionados por: Michel Beaudouin-La fon, Jean Camp, Cansu Canca, Brian Dean, Jeremy Epstein, Oliver Grau, Chris Hankin, Jim Hendler, Harry Hochheiser, Lorena Jaume-Palasi, Lorraine Kisselburgh, Marc Rotenberg, Gerhard Schimpf, Jonathan Smith, Gurkan Solmaz y Alec Yasinsac.

[2] Versión en español traducida por Carlos E. Jimenez-Gomez y revisada por Ricardo Baeza-Yates, Jeanna Matthews, y Manuel Pérez-Quiñones.

[3] AI como se utiliza aquí se refiere a los sistemas que emplean aprendizaje automático (*Machine Learning*), incluyendo el aprendizaje profundo (*Deep Learning*), el aprendizaje por refuerzo (*Reinforcement Learning*), la inferencia estadística u otros enfoques algorítmicos de este campo. Nuestras recomendaciones también se aplican a los sistemas algorítmicos de un modo más amplia, incluidos aquellos que no emplean Inteligencia Artificial conforme a esta definición.

Los algoritmos y otros mecanismos subyacentes utilizados por los sistemas de IA/ML para tomar decisiones específicas pueden ser opacos, haciendo las decisiones menos comprensibles y siendo más difícil determinar si sus resultados están sesgados o son erróneos.

Factores que hacen que estos sistemas sean opacos pueden ser:

- De información (los datos utilizados para entrenar modelos y generar resultados analíticos se utilizan sin el conocimiento o consentimiento explícito del interesado);
- Técnicos (el algoritmo puede no prestarse a una interpretación sencilla);
- Económicos (el coste de brindar transparencia puede ser excesivo);
- Competitivos (la transparencia puede comprometer secretos comerciales o permitir la manipulación de los límites de una decisión); y/o
- Sociales (la entrada no anonimizada puede violar las expectativas de privacidad).

Incluso sistemas algorítmicos bien diseñados pueden producir resultados o errores inexplicables. Es posible que contengan errores o que los datos de entrenamiento utilizados no hayan sido los adecuados para el propósito buscado. Las condiciones de su uso también pueden haber cambiado, invalidando así las suposiciones en las que se basó el diseño de tales sistemas.

Es más, el simple uso de un conjunto de datos ampliamente representativo no garantiza que el sistema esté libre de sesgos. La forma en que se procesan los datos, el ciclo de retroalimentación del usuario y el modo en que se despliega el sistema pueden crear nuevas problemáticas. Para mitigar los riesgos de sesgo o inexactitud inherentes al uso de sistemas automatizados de toma de decisiones:

- Los implementadores y operadores de sistemas deberían adherirse a los mismos estándares en la selección de entradas o la arquitectura de sistemas, que a los que se somete al ser humano cuando toman decisiones equivalentes;
- Los desarrolladores de sistemas de IA deberían realizar evaluaciones de impacto exhaustivas previamente al despliegue de sistemas de IA;
- Los formuladores de políticas deberían ordenar que se utilicen registros de auditoría para lograr los más altos estándares de exactitud,[4] transparencia e imparcialidad; y
- Los operadores de sistemas de IA deberían ser responsables por las decisiones que toman utilizando el sistema, independientemente de si se utilizan herramientas algorítmicas.

Los siguientes principios instrumentales, coherentes con el Código Ético de la ACM,[5] tienen el propósito de fomentar una toma de decisiones basada en algoritmos que sea justa, precisa y beneficiosa.

[4] Usamos exactitud para traducir el término *accuracy* del inglés, pues es más fidedigno que precisión.

[5] El código ético y de conducta profesional de ACM está diseñado para inspirar y guiar la conducta ética de todos los profesionales de la computación o informática, incluidos los profesionales actuales y futuros, instructores, estudiantes, personas influyentes y cualquier persona que utilice las tecnologías de computación o informáticas. El código incluye principios formulados como declaraciones de responsabilidad, basados en la creencia de que el bien público es siempre la consideración principal. Cada principio se complementa con pautas, que brindan explicaciones para ayudar a los profesionales de la computación a comprender y aplicar el principio. Consulte https://www.acm.org/code-of-ethics.

October 26, 2022

PRINCIPIOS PARA SISTEMAS ALGORITMICOS RESPONSABLES

1. **Legitimidad y competencia:** Los diseñadores de sistemas algorítmicos deberían tener tanto las competencias de gestión como la autorización explícita para construir y desplegar dichos sistemas. Asimismo deberían tener amplia experiencia en el dominio de la aplicación, una base científica para el uso previsto de dichos sistemas, y ser ampliamente considerados como socialmente legitimados por las partes interesadas afectadas por el sistema.[6] Se deben realizar evaluaciones legales y éticas para confirmar que cualquier riesgo introducido por los sistemas será proporcional a los problemas que se aborden, y que todas las partes interesadas relevantes entienden los compromisos entre beneficio y daños.

2. **Minimización del daño:** Gerentes, diseñadores, desarrolladores, usuarios y otras partes interesadas en los sistemas algorítmicos deberían ser conscientes de los posibles errores y sesgos involucrados en su diseño, implementación y uso, así como el potencial daño que un sistema puede causar a los individuos y la sociedad. Las organizaciones deberían realizar periódicamente evaluaciones de impacto sobre los sistemas que utilizan, para determinar si el sistema podría causar daños, especialmente daños discriminatorios, así como para aplicar las mitigaciones adecuadas. Cuando sea posible, los sistemas deberían aprender de las mediciones de desempeño real, y no solamente de patrones de decisiones anteriores que pueden haber sido discriminatorias.

3. **Seguridad y privacidad:** el riesgo de actores maliciosos puede ser mitigado mediante la introducción en cada fase del ciclo de vida de los sistemas, de buenas prácticas en seguridad y privacidad, incluyendo controles sólidos para mitigar nuevas vulnerabilidades que surjan en el contexto de los sistemas algorítmicos.

4. **Transparencia:** se alienta a los desarrolladores de sistemas a documentar claramente la forma en que se seleccionaron los conjuntos de datos, las variables y los modelos específicos utilizados para el desarrollo, entrenamiento, validación y test, así como las medidas específicas que se usaron para garantizar la calidad de los datos y los resultados. Los sistemas deberían indicar su nivel de confianza en cada resultado, y el ser humano debería intervenir cuando la confianza sea baja. Los desarrolladores deberían asimismo documentar los enfoques que se utilizaron para explorar posibles sesgos. Para los sistemas con un impacto crítico en la vida y el bienestar, se deberían exigir procedimientos de verificación y validación independientes. El escrutinio público de los datos y modelos brinda la máxima oportunidad de corrección. Por lo tanto, y en interés público, los desarrolladores deberían facilitar la realización de pruebas por parte de terceros.[7]

---

[6] No se deben implementar proyectos sin una base científica clara (por ejemplo, inferir rasgos de personalidad a partir de imágenes faciales).

[7] Por ejemplo, proporcionando acceso e interfaces de programación (APIs) para este propósito y eliminando las cláusulas de términos de servicio que desalientan la publicación de resultados.

5. **Interpretabilidad y explicabilidad:** Se anima a los gestores de sistemas algorítmicos a producir información tanto sobre los procesos que siguen los algoritmos empleados (interpretabilidad) como sobre las decisiones concretas que toman (explicabilidad). La explicabilidad puede ser tan importante como la exactitud, especialmente en contextos de políticas públicas o en cualquier entorno en el que haya preocupaciones sobre cómo los algoritmos podrían tener sesgos para beneficiar a un grupo en perjuicio de otro sin ser identificado. Es importante distinguir entre explicaciones y racionalizaciones posteriores que no reflejan la evidencia o el proceso de toma de decisiones utilizado para llegar a la conclusión explicada.

6. **Mantenibilidad:** se debería recopilar evidencia de la solidez de todos los sistemas algorítmicos a lo largo de sus ciclos de vida, incluida la documentación de los requisitos del sistema, el diseño o la implementación de cambios, los casos de prueba y los resultados, y un registro de los errores encontrados y corregidos.[8] El mantenimiento adecuado puede requerir tener que volver a entrenar a los sistemas con nuevos datos de entrenamiento y/o reemplazar los modelos empleados.

7. **Auditabilidad e Impugnación:** los reguladores deberían fomentar la adopción de mecanismos que permitan a individuos y grupos cuestionar los resultados y buscar reparación por efectos adversos resultantes de decisiones algorítmicamente informadas. Los gerentes deberían asegurarse de que datos, modelos, algoritmos y decisiones se registran para que puedan ser auditados y los resultados replicados en casos en los que se sospecha o se alega daño. Las estrategias de auditoría deberían hacerse públicas para permitir que personas, organizaciones de interés público e investigadores, puedan revisarlas y recomendar mejoras.

8. **Rendición de cuentas y responsabilidad:** los organismos públicos y privados deberían rendir cuentas por las decisiones tomadas por los algoritmos que utilizan, incluso si no es factible explicar en detalle cómo esos algoritmos produjeron los resultados. Dichos organismos deberían ser responsables de los sistemas completos tal como se implementan en sus contextos específicos, no sólo de las partes individuales que componen un sistema determinado. Cuando se detecten problemas en los sistemas automatizados, las organizaciones responsables de implementar dichos sistemas deberían documentar las acciones específicas que tomarán para resolver el problema y bajo qué circunstancias el uso de dichas tecnologías debería ser suspendido o cancelado.

9. **Limitación de los impactos medioambientales:** los sistemas algorítmicos deben diseñarse para informar estimaciones de los impactos medioambientales, incluidas las emisiones de carbono tanto de la fase de aprendizaje como de las operaciones computacionales de su uso. Los sistemas de IA deben diseñarse para garantizar que sus emisiones de carbono sean razonables dado el grado de exactitud requerido por el contexto en el que se implementan.

[8] De lo contrario, el sistema puede devenir menos apropiado a medida que las entradas se desvíen de las previstas originalmente, o si las condiciones subyacentes del mundo real cambian (por ejemplo, los sistemas de reconocimiento facial se utilizan en un grupo demográfico más amplio o diferente al que estaba presente en los datos de entrenamiento).

APLICACIÓN DE LOS PRINCIPIOS: GOBERNANZA Y COSTE-BENEFICIO

El primer principio de legitimidad y competencia debe considerarse antes de implementar un sistema algorítmico. Es decir, el organismo donde se despliega debería tener un proceso de gobernanza claro para decidir cuándo diseñar e implantar un sistema algorítmico.

El segundo principio de minimización del daño, especialmente el daño discriminatorio, es un valor central de la ética y por esa razón también informa otros principios. Este y los principios restantes deberían abordarse en la medida necesaria durante todas y cada una de las fases del desarrollo e implantación del sistema para minimizar los daños potenciales. Estos principios son más importantes para los sistemas algorítmicos que afectan directamente a las personas y donde el ser humano tiene pocas oportunidades de intervenir.

El grado de transparencia exigido a un sistema algorítmico debería ser consistente con el impacto del sistema. Recomendamos identificar niveles de impacto de manera que se apliquen mayores requisitos de transparencia a los sistemas con mayores niveles de impacto (por ejemplo, sistemas con riesgo para la vida humana o sistemas en áreas reguladas como contratación, vivienda, crédito y asignación de recursos públicos). De manera similar, el nivel de mantenimiento requerido debería ser proporcional al impacto del sistema.

Los profesionales responsables de aplicar estos principios deben decidir sobre el equilibrio coste-beneficio necesario en función de su conocimiento del dominio y la consulta con las partes interesadas. Ejemplos de tales compensaciones incluyen:

- Las soluciones deberían ser proporcionales al problema que se está resolviendo, incluso si eso afecta a la complejidad o costo (por ejemplo, rechazar el uso de videovigilancia pública para una simple tarea de predicción).
- Se debería considerar una amplia variedad de métricas de rendimiento, pudiéndose ponderar de modo diferente según el dominio de la aplicación. Por ejemplo, en algunas aplicaciones en el área de la salud, los efectos de los falsos negativos pueden ser mucho peores que los falsos positivos, mientras que, en la justicia penal las consecuencias de los falsos positivos (por ejemplo, encarcelar a una persona inocente) puede ser mucho peor que los falsos negativos. La más deseable configuración operacional del sistema rara vez es la que tiene la máxima exactitud.
- Inquietudes sobre la privacidad, la protección de secretos comerciales o la revelación de resultados analíticos que podrían permitir que actores malintencionados manipulen al sistema, pueden justificar la restricción de acceso a individuos cualificados, pero no deberían usarse para justificar la limitación al escrutinio por parte de terceros, o para eximir a los desarrolladores de las obligaciones de reconocer y corregir los errores.
- La transparencia debe ir acompañada de procesos de rendición de cuentas que permitan a las partes interesadas afectadas por un sistema algorítmico buscar una reparación significativa por los daños causados. La transparencia no debería utilizarse para legitimar un sistema o transferir la responsabilidad a terceros.

- Cuando el impacto de un sistema es alto, puede ser preferible un sistema más explicable. En muchos casos, no existe compensación entre explicabilidad y la exactitud. En algunos contextos, sin embargo, las explicaciones incorrectas pueden ser incluso peores que la ausencia de explicación (por ejemplo, en los sistemas de salud, un síntoma puede corresponder a muchas enfermedades posibles, no sólo una).

La política pública es importante. Es difícil esperar que las fuerzas del mercado incentiven a las empresas privadas a equilibrar las compensaciones que implican riesgos para las personas y la sociedad cuando los intereses de dichas empresas son diferentes. Por lo tanto, las políticas públicas son necesarias para exigir, o al menos fomentar, evaluaciones de impacto y niveles de explicabilidad y auditabilidad para diferentes clases de sistemas. Las políticas públicas que aclaran dónde se registran los registros de auditoría y quién tiene acceso a ellos alentarán a diseñadores y desarrolladores a considerar el análisis modal de fallos y efectos, y aumentarán la confianza de los usuarios, las partes interesadas y los organismos de supervisión.

Las recomendaciones anteriores se centran en el diseño responsable,[9] desarrollo y utilización de sistemas algorítmicos. La responsabilidad debe ser determinada por la ley y las políticas públicas. El poder cada vez mayor de los sistemas algorítmicos y su uso en aplicaciones relevantes y críticas para la vida implica que se debe tener sumo cuidado al usarlos. Estos nueve principios instrumentales están destinados a ser una inspiración para iniciar debates, comenzar investigaciones y desarrollar métodos de gobernanza para brindar beneficios a una amplia gama de usuarios, al mismo tiempo que promueven la fiabilidad, la seguridad y la responsabilidad. En definitiva, es el contexto específico el que define el diseño y uso correctos de un sistema algorítmico en colaboración con representantes de todas las partes interesadas afectadas.

[9] Se insta a los diseñadores y desarrolladores a producir suficientes evidencias de la fiabilidad de un sistema para que pueda usarse de manera responsable, en lugar de poner la responsabilidad sobre el usuario para que confíe en los sistemas sin evidencia suficiente (por ejemplo, como en la IA confiable o fiable).