TRANSPARENCY AND EXPLAINABILITY THEMATIC AREA NARRATIVE IN ENGLISH ARABIC FRENCH PORTUGUESE AND SPANISH

Rachel Adams , Kelly Stone

Rachel Adams , Kelly Stone

IDRC GRANT / SUBVENTION DU CRDI : - GLOBAL INDEX ON RESPONSIBLE ARTIFICIAL INTELLIGENCE

Índice Global sobre IA Responsable

Dimensión: Gobernanza responsable de la IA

Subdimensión: Estado de derecho

Área temática: Transparencia y explicabilidad

Definiciones

La transparencia se define como "la cualidad de que [un proceso, gestión o decisión] se lleve a cabo abiertamente, sin secretos, para que las personas puedan asumir que [las acciones] son justas y honestas".[1] La transparencia ha sido identificada por el Consejo de Derechos Humanos de las Naciones Unidas (CDHNU) como uno de los cinco principios clave de buena gobernanza, junto con la responsabilidad, la rendición de cuentas, la participación y la capacidad de respuesta a las necesidades de la población.

En el contexto de la IA, la transparencia se define como la medida en la que el "funcionamiento interno" de un sistema de IA es abierto y accesible, lo que invariablemente requiere que se ofrezcan explicaciones "fáciles de entender" sobre los modelos algorítmicos, los datos que los orientan y la lógica de su utilización. La transparencia también debe existir en el propio uso de un sistema de IA, ya sea por parte de un agente público o privado, y en particular dirigido a las personas que puedan verse afectadas por el uso de este sistema. Para ello, la transparencia pretende garantizar que las personas puedan desarrollar una comprensión básica de los sistemas de IA, incluidas las razones para su desarrollo, formación, despliegue e impacto.

Por lo tanto, el principio de la transparencia implica el concepto de la explicabilidad, surgido en el ámbito del discurso ético sobre la IA, y que requiere que se proporcionen informaciones que puedan ser entendidas por una persona que no sea técnica – en concreto sobre el proceso que conduce a los resultados de los sistemas de IA. Se puede definir la explicabilidad como "la cualidad que permite que las personas entiendan cómo funciona un sistema en concreto y/o cómo se ha logrado un cierto resultado, proporcionando información sensata y fácil de comprender".[2] El principio de explicabilidad también está estrechamente relacionado con el concepto de interpretabilidad, que se refiere a la "capacidad de explicar los factores y la lógica que

[1] Traducción no oficial
[2] Traducción no oficial

conducen a un resultado, en términos comprensibles para un ser humano".[3]

La transparencia y la explicabilidad, como principios fundamentales de una gobernanza responsable de la IA, requieren que la información sobre los procesos, prácticas y decisiones, sea abierta, accesible y fácilmente comprendida por un amplio conjunto de partes interesadas, incluidas las entidades gubernamentales, empresas del sector privado, organizaciones de la sociedad civil e instituciones académicas.

Justificaciones

Los sistemas y herramientas de IA evolucionan continuamente impregnando, cada vez más, el panorama digital actual y la sociedad en general. La dependencia de las tecnologías de IA para desempeñar distintas funciones, con y sin el conocimiento de las personas, como hacer predicciones y recomendaciones o tomar decisiones, ha reforzado la necesidad de aumentar los niveles de transparencia sobre el uso de estos sistemas y de explicar su funcionamiento y su impacto. Comprender la lógica que subyace a las decisiones generadas por la IA, el momento y el propósito de desplegar sistemas de IA, así como los procesos y los datos de entrenamiento que les dan forma, se convierte en algo crucial debido a los riesgos potenciales que presentan estas tecnologías emergentes.

La Recomendación de la UNESCO sobre la Ética de la IA (Recomendación) describe la transparencia como un “requisito previo esencial para garantizar el respeto, la protección y la promoción de los derechos humanos”. La recomendación también reconoce que la falta de transparencia puede “mermar la posibilidad de impugnar eficazmente las decisiones basadas en resultados producidos por los sistemas de IA”, que pueden vulnerar el derecho a un juicio imparcial y a un recurso efectivo, limitando los ámbitos en los que estos sistemas pueden utilizarse legalmente. La recomendación también refiere que un mayor grado de transparencia puede “permitir a las personas comprender cómo se implementa cada etapa de un sistema de IA”, lo que puede proporcionar información sobre los factores que influyen en una predicción o decisión específicas, y sobre la existencia o no de garantías adecuadas (como medidas de seguridad o de equidad).

Además, la Organización para la Cooperación y el Desarrollo Económicos (OCDE) considera que el principio de la transparencia requiere la divulgación del uso de sistemas de IA, en concreto, cuando las herramientas de IA hacen predicciones o recomendaciones, o cuando el usuario interactúa directamente con tecnologías de IA. La transparencia requiere que las personas "comprendan cómo se desarrolla, entrena, funciona e instala un sistema de IA",[4] proporcionando acceso a información importante, que sea fácil de entender para que los consumidores puedan tomar decisiones informadas. Además, la OCDE también señala que un elemento crucial de la transparencia es la creación de mecanismos que faciliten diálogos claros entre las distintas partes interesadas y la existencia de entidades o instituciones dedicadas a promover el conocimiento general de la IA, respondiendo a las inquietudes del público sobre su uso y aumentando la confianza y aceptación de su aplicación responsable.

Identificaciones

[3] Traducción no oficial
[4] Traducción no oficial

Esta área temática recoge las medidas adoptadas por los diferentes países para promover y garantizar un nivel suficiente de transparencia y explicabilidad en el desarrollo, uso y despliegue de tecnologías de IA. En concreto, se debe tener en cuenta: (1) los ***marcos jurídicos*** relativos a las normas de transparencia y explicabilidad en el uso, desarrollo y despliegue de la IA, (2) las ***acciones gubernamentales*** para promover la transparencia y la explicabilidad en el uso de sistemas de IA, y (3) ***los agentes no estatales*** que trabajan para promover, permitir o aplicar la transparencia y la explicabilidad en todas las fases del ciclo de vida de la IA.

Los *marcos jurídicos* del país pueden adoptar la forma de leyes, reglamentos, políticas (incluso por sector y/o departamento) y/o directrices. Las *acciones gubernamentales* pueden incluir proyectos de ley o de políticas, la creación de organismos de supervisión para controlar el cumplimiento o la aplicación de las políticas destinadas a hacer más transparentes y explicables los sistemas de IA. Los *Agentes No Estatales* (ANE) pueden ser organizaciones no gubernamentales (ONG), pero también empresas multinacionales, organizaciones militares privadas, medios de comunicación, grupos étnicos organizados, instituciones académicas, grupos de presión, sindicatos o movimientos sociales que contribuyan a reforzar la transparencia y la explicabilidad de los sistemas de IA.

Ejemplos

Marcos jurídicos

Australia ha establecido una serie de Principios Éticos de la IA para orientar el desarrollo y el uso de la inteligencia artificial de forma responsable y ética. Los principios proporcionan un marco para ayudar a particulares y organizaciones a garantizar que las tecnologías de IA se utilicen de forma acorde con los valores y las expectativas sociales de Australia. Uno de los principios fundamentales es el de la Transparencia y la Explicabilidad según el cual “debe haber transparencia y divulgación responsable para que las personas puedan entender cuándo se están viendo afectadas significativamente por la IA y puedan saber cuándo un sistema de IA está interactuando con ellas”. Más concretamente, los principios establecen que la divulgación responsable debe hacerse de manera oportuna y “proporcionar justificaciones razonables de los resultados de los sistemas de IA”, lo que incluye necesariamente información que ayude a las personas a entender los resultados, incluidos los factores contemplados en los procesos de toma de decisiones. Estos principios pretenden orientar el uso responsable y ético de la tecnología de IA, en diversos sectores de Australia, fomentando la confianza, la equidad y los beneficios sociales.

Acciones gubernamentales

En abril de 2021, la Comisión Europea publicó el Reglamento del Parlamento Europeo y del Consejo que establece normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial). La Ley de IA clasifica los diferentes tipos de tecnologías en función de tres niveles de riesgo (inaceptable, alto riesgo y bajo o mínimo riesgo) imponiendo, a continuación, determinadas obligaciones a los proveedores de IA para atenuar posibles daños. El Título IV impone obligaciones específicas de transparencia a los sistemas clasificados como de alto riesgo, incluidos los que: (1) interactúan con seres humanos, (2) se utilizan para detectar emociones o determinar la asociación a

categorías (sociales) a partir de datos biométricos, (3) generan o manipulan contenido («ultrafalsificaciones»). Por ejemplo, cuando las personas interactúan con un sistema de IA o cuando se reconocen sus emociones o características por medios automatizados, el Título IV exige que se informe a las personas que utilizan la IA de que están siendo sometidas a dicha tecnología y que se les *explique* su funcionamiento. Además, el Título IV exige a los operadores de sistemas de IA que revelen cuándo se está generando contenido a través de medios automatizados, a menos que entre dentro de una de las excepciones previstas en la legislación, para garantizar que las personas puedan elegir con conocimiento de causa su nivel de implicación.

Aunque la Ley de IA aún no ha entrado en vigor, se están realizando significativos esfuerzos para garantizar que tanto el Consejo Europeo (que representa a los 27 Estados Miembros de la UE), como el Parlamento Europeo llegan a un acuerdo sobre una versión común del texto. El 6 de diciembre de 2022, el Consejo adoptó su posición común ("orientación general") sobre la Ley de IA, que le permitirá entablar negociaciones con el Parlamento Europeo antes de llegar a un acuerdo. Según informes recientes, se espera que la Ley de IA se promulgue a finales de 2023, momento en el que todos los proveedores de sistemas de IA tendrán que cumplirla, en un plazo de 24 a 36 meses.

Agentes no estatales

Credo AI, una de las diversas empresas especializadas en plataformas de software de gobernanza, riesgos y cumplimiento (GRC), que permite a las empresas gestionar de forma responsable los activos de IA, ofrece una innovadora "Plataforma de Gobernanza de IA" para ayudar a las empresas a "analizar, auditar y gestionar los riesgos"[5] que plantea el aprendizaje automático y los sistemas de IA. En noviembre de 2022, la empresa presentó una nueva funcionalidad de evaluación e información para favorecer aún más el cumplimiento, por parte de las empresas, de las obligaciones normativas y las demandas de los consumidores, en materia "de gobernanza, información y divulgación"[6] en torno al desarrollo y uso de la IA. Esta funcionalidad presta especial atención a la explicabilidad, así como a los aspectos de equidad, solidez, seguridad y privacidad.

[5] Traducción no oficial
[6] Traducción no oficial