Recomendaciones para una IA responsable

Ricardo Baeza-Yates[1]

Introducción

Este ensayo resume la mayoría de las propiedades de un sistema de software, en particular, los basados en inteligencia artificial (IA), y entrega recomendaciones para el análisis de algunas de ellas. No todas estas propiedades son necesarias para tener una IA responsable y algunas son más generales, pero todas ellas pueden generar algún impacto legal o ético.

Es importante señalar también que se habla de IA confiable o IA ética, pero ambos términos tienen problemas. El primero es tendencioso, pues sabemos que los sistemas de IA no son confiables todo el tiempo y, además, pone el peso del problema en el usuario y no en los creadores. Asimismo, la IA no puede ser ética pues la ética es humana y, por lo tanto, no es una característica de una tecnología. Dado que el tema de sistemas de IA responsable es un área en pleno desarrollo, todo lo presentado puede cambiar significativamente en el futuro, cuando ya no sea una tendencia, sino una necesidad. En particular, cuando hablamos de inteligencia artificial nos referimos a técnicas de aprendizaje automático. Pero en el futuro existirán sistemas híbridos que mezclan estas técnicas con bases de conocimiento y otras de inferencia más avanzadas, al igual que de comprensión semántica y razonamiento lógico.

[1] Instituto de IA Experiencial, Northeastern University Silicon Valley, EE. UU. Departamento de Ciencias de la Computación, Universidad de Chile, Santiago, Chile.

Propiedades de Sistemas de IA

A partir de un estudio bibliográfico y conocimiento previo, se recolectaron 30 propiedades relacionadas con datos y software, independientes de si ellas eran necesidades éticas o no, ya que en muchos casos esto depende de la aplicación específica. Estas propiedades se analizaron desde dos puntos de vista:

- **Aplicación**: referidas a los datos, los modelos/algoritmos principales de la aplicación, al sistema completo de IA o a la gobernanza de este mismo, sin incluir la gobernanza de los datos. La Tabla 1 muestra nuestro análisis inicial en estos cuatro aspectos.

- **Impacto**: son importantes para la justicia, para el gobierno, para los usuarios del sistema o para la sociedad en general. La Tabla 2 muestra nuestro análisis inicial de estos cuatro actores.

En ambas tablas hemos traducido del original en inglés, aunque algunas de ellas no tienen traducción directa como *accountability*, que es rendición de cuentas (esto es un ejemplo de sesgo semántico codificado en el lenguaje). Con respecto a este análisis debemos señalar:

- Respecto a la aplicación, hay que advertir que, en el caso de un sistema basado en aprendizaje automático supervisado, lo importante es el modelo que se usa. Sin embargo, un modelo por sí solo no resuelve nada y necesita una aplicación que lo procesa, interpreta y ejecuta. Por lo tanto, el modelo en producción también es un algoritmo.

- Respecto al impacto, aunque hay propiedades que tienen implicancias legales y éticas, la relevancia de ellas dependerá del uso particular que se dé al sistema de IA y del contexto de este. Por ejemplo, las implicancias legales son muy distintas para un país específico o un uso a nivel mundial, en el cual muchas más propiedades serán necesarias.

Propiedad	Datos	Modelo/ Algoritmo	Sistema	Gobernanza
Procedencia de los datos	✓			✓
Privacidad	✓		✓	✓
Control de calidad	✓		✓	✓
Trazabilidad	✓		✓	✓
Acceso y corrección	✓		✓	✓
Mantenimiento	✓	✓	✓	✓
Equidad y sesgos	✓	✓	✓	✓
Cumplimiento legal	✓	✓	✓	✓
Completitud		✓	✓	✓
Conciencia		✓	✓	✓
Eficiencia		✓	✓	
Validación y testeo		✓	✓	
Interpretabilidad		✓	✓	
Explicabilidad		✓	✓	
Accesibilidad			✓	
Rendición de cuentas			✓	✓
Responsabilidad			✓	✓
Integridad y confianza			✓	✓
Seguridad			✓	✓
Proporcionalidad			✓	✓
Interoperabilidad			✓	✓
Autonomía			✓	✓
Transparencia			✓	✓
Documentación			✓	✓
Beneficiosa			✓	✓
Resiliencia			✓	✓

Usabilidad	✓	✓
Sostenibilidad	✓	✓
Auditabilidad	✓	✓
Reproducibilidad	✓	

Tabla 1: Propiedades indicando su dependencia de distintas partes del sistema de IA.

Las propiedades en negrita son las 7 inicialmente definidas por la *Association for Computing Machinery* (ACM) para lograr la *transparencia* de los sistemas de IA [1], aunque transparencia no es uno de los principios definidos y solo aparece en el título. La ACM es la asociación de profesionales de computación más grande del mundo, con más de 100 mil miembros, y por ende, es un referente de primer nivel.

Propiedad	**Justicia**	**Gobierno**	**Usuarios**	**Sociedad**
Procedencia de los datos	✓	✓	✓	✓
Privacidad	✓	✓	✓	✓
Control de calidad			✓	✓
Trazabilidad				
Acceso y corrección				
Mantenimiento				
Equidad y sesgos	✓	✓	✓	✓
Cumplimiento legal	✓	✓	✓	✓
Completitud			✓	✓
Conciencia			✓	✓
Eficiencia			✓	✓
Validación y pruebas				
Interpretabilidad				
Explicabilidad	✓	✓	✓	✓
Accesibilidad	✓	✓	✓	✓

Rendición de cuentas	✓	✓	✓	✓
Responsabilidad	✓	✓	✓	✓
Integridad y confianza	✓	✓	✓	✓
Seguridad	✓	✓	✓	✓
Proporcionalidad	✓		✓	✓
Interoperabilidad			✓	
Autonomía			✓	
Transparencia			✓	✓
Documentación			✓	✓
Beneficiosa			✓	✓
Resiliencia			✓	✓
Usabilidad			✓	✓
Sostenibilidad	✓	✓		✓
Auditabilidad	✓	✓		
Reproducibilidad	?	?		

Tabla 2: Propiedades dependiendo de a quién impactan o a quién les importa.

En estas tablas podemos ver que hay propiedades de más alto nivel que otras. También vemos que algunas propiedades son transversales en ambas tablas (es decir, la fila correspondiente está completa) y las hemos ordenado verticalmente de modo de que se vean los grupos de propiedades que tienen aspectos y/o actores similares. Sobre esta base, proponemos una estructura jerárquica de estas propiedades en la Tabla 3, en la que hemos destacado en negrita las propiedades relacionadas con la ética (si es una propiedad principal, todas las propiedades secundarias también lo son). Igualmente, hemos incluido aquí el término original en inglés.

Propiedad Principal	Propiedades Secundarias	Notas
Conciencia (*Awareness*)	Validez ética y legal Validez científica Autonomía (*autonomy*) Integridad (*integrity*)	Legitimidad e identidad del sistema
Proveniencia de datos (*Data provenance*)	**Control de calidad** (*quality assurance*) **Equidad y sesgo** (*equity & bias*) Trazabilidad (*traceability*) Acceso y corrección (*access & redress*)	Representan el ciclo de vida de los datos
Robustez (*Robustness*)	**Control de calidad** (*quality assurance*) Adaptabilidad (*adaptability*) Escalabilidad (*scalability*) Extensibilidad (*extensibility*) Interoperabilidad (*interoperability*)	Representan la completitud del sistema
Usabilidad (*Usability*)	Eficiencia (*efficiency*) Accesibilidad (*accessibility*) Resiliencia (*resilience*) Reproducibilidad (*reproducibility*)	Permiten la satisfacción del usuario
Transparencia (*Transparency*)	Validación y testeo *(validation & testing*) Documentación (*documentation*) Interpretabilidad (*interpretability*) Explicabilidad (*explainability*) Auditabilidad (*auditability*)	Permiten la transparencia del sistema
Responsabilidad (*Responsibility*)	Conformidad legal (*Legal compliance)* Rendición de cuentas (*Accountability)* Proporcionalidad (*Proportionality*) Privacidad (*Privacy)* Seguridad (*Security & safety*) Integridad y confianza (*Trustworthy*) Mantenibilidad (*Maintenance*) Sostenibilidad (*Sustainability)* Beneficiosa (*Beneficial/wellbeing)*	Permiten que el sistema cumpla con principios éticos y normas legales

Tabla 3: Agrupación jerárquica de propiedades.

Puede observarse que el control de calidad se ha separado en dos para aplicarlo tanto a los datos como al sistema. Otras propiedades están ya incluidas dentro de las anteriores. Por ejemplo,

la protección de datos es el resultado de la privacidad y la seguridad. Del mismo modo, la eficacia del sistema o el uso racional de recursos es parte de la sostenibilidad. Otras propiedades relacionadas con los datos están también en responsabilidad. Por ejemplo, la recolección mínima y el almacenamiento de datos por un tiempo acotado es parte de la proporcionalidad. Quiero recordar que interpretabilidad se refiere a entender cómo el sistema llega a una decisión, mientras que explicabilidad significa que el sistema debe poder explicar una respuesta específica a un usuario dado. Esta agrupación en el original en inglés [4], fue uno de los insumos usados para generar los nuevos principios para sistemas algorítmicos responsables de la ACM publicados en octubre de 2022 [5] y en particular el primer principio de legitimidad y competencia.

Gobernanza

La gobernanza de un sistema de IA incluye los datos y sus metadatos, los modelos y todos los procesos para entrenarlos, validarlos y evaluarlos, el software en producción con todos los mecanismos de registro y control subyacentes, y toda la información necesaria para mantener y evaluar las propiedades antes descritas.

Hay pocas propuestas concretas para la gobernanza. Una de ellas es la de Ben Shneiderman [12], que la divide en tres partes. La ingeniería de software como tal, que es a la que se refiere el párrafo anterior, que concierne al equipo de desarrollo. Por encima de ella está el diseño organizacional, que es la gobernanza de la institución misma. Finalmente, agrega un tercer nivel, que es el de certificación externa, el cual incluye regulación gubernamental y auditorías de software.

Sin embargo, en la práctica, es mejor ver la gobernanza en forma temporal respecto a los principios instrumentales que se definan. Si usamos los 9 nuevos principios de la ACM, tenemos el diagrama de la Figura 1 [6]. En negrita se destacan las herramientas principales que ayudan a la gobernanza. Si la gobernanza es adecuada, evitaremos la frecuencia de la última herramienta: auditoría algorítmica.

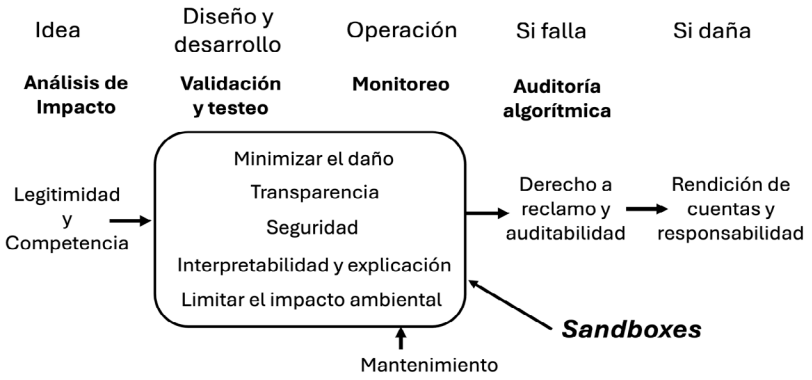

Figura 1: Gobernanza temporal usando los principios de la ACM.

Control de Calidad

Un modelo basado en aprendizaje automático supervisado debe ser robusto. Esto incluye una serie de análisis que no siempre se realizan, el que se resume a continuación. Para ello supondremos que en ciertos casos hay una noción de complejidad del problema o al menos una medida sustituta (*surrogate, proxy*) de ella; si no, habría que usar la valoración de expertos del área de aplicación.

Calidad de los Datos

Cobertura de los datos de entrenamiento: ¿tenemos suficientes datos en todo el espacio de posibles conjuntos de entrada? Debemos recordar que la complejidad de cada instancia es distinta, nunca es uniforme. Por lo tanto, el número de datos deberá ser al menos proporcional a esta complejidad. Si no es posible conocer la complejidad, una forma de analizar esto es escoger las variables más importantes de la entrada y dividir su rango de valores en dos o más partes de acuerdo con el problema. Por ejemplo, si uno usa la edad, podría dividir en niños y adolescentes, adultos hasta los 65 años y mayores de 65 años. ¿Tenemos una cantidad de datos en cada rango

proporcional a la población en ese rango de acuerdo con el censo del país en que se usa la aplicación?

Análisis de sesgos de datos: ¿analizó posibles sesgos en los datos? Si tiene datos demográficos, estos pueden incluir sesgo de género, edad, etnia, nivel económico, nivel educativo, etc. En algunos casos los sesgos pueden ser justificados dado el problema a resolver, pero habrá que verificar que sea el esperado. Por ejemplo, si se espera que haya más representación masculina, no debiera estar sobrerrepresentado. Es decir, hay que decidir cuál es la distribución neutral para el problema dado y eso no siempre es simple. Es fundamental tener en cuenta que en muchos casos se trata de una decisión social, no una decisión de los desarrolladores, así que debe ser validada con expertos del área de aplicación. Por otro lado, hay sesgos que pueden ser desconocidos. Si usamos el mismo ejemplo de la edad anterior, podemos hacer muestras al azar del mismo tamaño en esos tres rangos. ¿Son las distribuciones similares? Si no lo son, ¿hay una justificación para que no lo sean o hay un sesgo que no esperábamos?

Calidad del Modelo

Análisis de sesgos algorítmicos: el modelo mismo puede generar sesgos no previstos. Un buen ejemplo es el caso de Deliveroo [10], una aplicación para repartir comida, la cual al intentar maximizar la ganancia económica, asignaba menos trabajo a los repartidores que no podían o no querían trabajar a las horas de más pedidos (la cena). En este caso el modelo no consideró que debía distribuir equitativamente los pedidos, atendiendo a las limitaciones personales válidas de los repartidores (atención de niños o personas dependientes, horas no hábiles, etc.).

Otra fuente de sesgos algorítmicos es el ciclo de realimentación sistema-usuario debido a su uso. Por ejemplo, los sistemas de recomendación solo exponen al usuario a un número limitado de alternativas, generando un sesgo de popularidad en el cual los ítems más frecuentes tienen ventajas. También la posición en la pantalla de cada ítem genera un sesgo y el orden de los resultados genera

otro sesgo (sesgo de *ranking*). A esto hay que agregarle el problema de la burbuja (*filter bubble*) o de la cámara de eco, referido al conocimiento parcial de las preferencias de los usuarios (que afecta a la personalización de la experiencia). Todo esto se agrava con los sesgos cognitivos de los usuarios y sesgos de segundo orden producto del uso de los resultados de un sistema para alimentar otro sistema (por ejemplo, usar el resultado de los buscadores para generar contenido nuevo en la Web).

Punto de operación: el modelo debe ajustarse al objetivo del problema y esto no necesariamente significa usarlo en el punto de mayor exactitud (*accuracy*). De hecho, el mejor modelo es el que maximiza la exactitud en el punto de operación, el que no es necesariamente el mejor posible en todos los puntos de operación posibles (curva precisión/exactitud o similar). Por ejemplo, en aplicaciones médicas es mejor tener más falsos positivos (es decir, personas que no están enfermas que luego de una visita al médico son descartadas) que falsos negativos (personas que están enfermas pero que nunca se enterarán). Un ejemplo de este tipo de análisis se presenta en predicción de dislexia [11].

Calidad de los Resultados

Análisis de sensibilidad: un modelo debe ser robusto. Esto significa que pequeños cambios en la entrada causan pequeños cambios en la salida. Viceversa, grandes cambios en las variables principales de la entrada deberían producir grandes cambios en la salida (para esto debemos realizar antes un análisis del impacto de las variables o características de la entrada, *feature analysis*). En general, modelos más sensibles tendrán mayores niveles de error.

Análisis de error: las medidas promedio como la exactitud no muestran como el modelo se comporta para distintas instancias del problema. Es muy probable que el error sea insignificante cuando la instancia es fácil y mucho mayor cuando la instancia sea compleja. Este segundo caso es importante investigarlo pues aquí están los errores con mayor impacto en la operación del sistema y que son los

que al final pueden afectar a personas específicas, grupos minoritarios o incluso a una gran parte de la sociedad (por ejemplo, mujeres).

Interpretabilidad y explicabilidad

No hay una definición única para estos dos términos, pero hay consenso en que un modelo de aprendizaje automático es *interpretable* si un ser humano puede comprender cómo el modelo toma una decisión. Esto implica conocer el proceso de cómo el modelo llega a un resultado. Por otro lado, un modelo es *explicable* si un ser humano puede comprender por qué se tomó una decisión específica. Esto implica conocer qué atributos o variables influyen en el resultado y en qué medida. Más aún, esta explicación debería poder darse en lenguaje natural.

En el caso de modelos opacos, como aprendizaje profundo, donde los atributos son generados por el sistema, es más difícil dar una explicación. En otros casos más transparentes, es necesario realizar el mismo análisis de características mencionado para la sensibilidad (*feature analysis*), el que depende del método de aprendizaje automático utilizado.

La facilidad para dar una explicación depende, como es esperable, de la complejidad del problema. Más complejo, más difícil será explicarlo. Es probable, pero no es siempre cierto, que, si el modelo es más complejo, probablemente la exactitud de la respuesta será menor. Por esta razón las técnicas para generar explicaciones dependen del método de aprendizaje automático usado. En el caso de que sean genéricas, estas pueden ser locales (explican un resultado) o globales (intentan explicar el modelo completo). En [9] se presenta una taxonomía de modelos interpretables y una revisión extensa de todos los trabajos de explicabilidad a la fecha.

Conclusiones

Para una exposición más completa de todos los sesgos posibles en un sistema, recomendamos los sesgos de la Web [2]. Gran parte del problema de sesgos viene de categorías sociales que no tienen justificación científica, tal como se muestra en el documental Coded Bias [3]. Los ejemplos más clásicos son raza [8] o preferencia sexual [7], respecto de los cuales la biología nos enseña que estamos simplificando un espacio multidimensional complejo en unas pocas categorías arbitrarias. Por esta misma razón, cuando usamos interseccionalidad (intersección de dos o más categorías), intentamos recuperar los casos más complejos, pero al mismo tiempo, en cierto sentido, validamos las categorías. La solución, por supuesto, es eliminar este tipo de categorizaciones, pero con esto tenemos la paradoja de que, si no las conocemos, no podemos medir los daños que han provocado sus sesgos en los datos.

En las consideraciones éticas, suponiendo que ya están cubiertos todos los requerimientos legales, la primera pregunta que debemos hacernos es: ¿usarías tu producto si pertenecieras al grupo objetivo de tu aplicación? Si la respuesta es no, ya tenemos un problema ético. Si la respuesta es sí, es de esperar que no sea porque sabemos que en nuestro caso nos favorece, ya que si no entonces habría otro problema ético. El resto de las preguntas dependerá de los resultados de los análisis mencionados en este ensayo.

Referencias bibliográficas

Association for Computing Machinery. Algorithmic Transparency and Accountability. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf. 1/12/2017.

Baeza-Yates, Ricardo. Bias on the Web. Communications of ACM, vol. 61 (6), pp. 54-61, 6/2018. Presentación disponible en YouTube.

Baeza-Yates, Ricardo, Muñoz, Catherine. Sesgos codificados. Ciper Académico, 8/5/2021.

Baeza-Yates, Ricardo. Some thoughts on Responsible AI, sin publicar, 13/4/2022.

Baeza-Yates, Ricardo, Matthews, Jeanna, et al. Principles for Responsible Algorithmic Systems, ACM, 24/10/2022. En castellano en: https://www.acm.org/binaries/content/assets/public-policy/spanish-statement-ai.pdf.

Baeza-Yates, Ricardo. Introduction to Responsible AI. European Review 31 (4), 3/8/2023.

Helm, Rebecca. Let's talk about biological sex. Twitter, 19/12/2019.

Kolbert, Elizabeth. There is no scientific base for race – it's a made-up label. National Geographic, 4/2018.

Linardatos, Pantelis, Papastefanopoulos, Vasilis, Kotsiantis, Sotiris. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy 23, 18, 2021.

Lomas, Natasha. Italian court rules against 'discriminatory' Deliveroo rider-ranking algorithm. TechCrunch, 4/1/2021.

Rello, Luz, Baeza-Yates, Ricardo, Ali, Abdullah, Bigham, Jeffrey P., Serra, Miquel. Predicting risk of dyslexia with an online gamified test. PlosOne, 2/12/2020.

Shneiderman, Ben. Bridging the Gap between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems. ACM Transactions on Interactive Intelligent Systems 10, 4, 10/2020.