Integración de los derechos humanos en el ciclo de vida de la IA

Un marco para el desarrollo de la Inteligencia Artificial

Autoras: Emma Kallina, Sofia Kypraiou, & Caitlin Kraft-Buchman
<AI & Equality> IA basada en Derechos Humanos, Diciembre 2025..

Abstract

Las investigaciones actuales ponen de relieve la posibilidad de que los sistemas de IA afecten negativamente a los Derechos Humanos (DD.HH) individuales y colectivos si se desarrollan sin una consideración cuidadosa. Mediante la incorporación de un análisis crítico y de puntos de reflexión sobre el impacto en los derechos humanos durante el desarrollo de la IA, estos daños pueden mitigarse o evitarse por completo. Este libro blanco esboza nuestro marco <IA & Equidad> que permite un enfoque de este tipo, yendo aún más lejos y promoviendo un desarrollo de la IA impulsado por el deseo de promover la dignidad humana. El marco consiste en las preguntas esenciales y los puntos de reflexión que son relevantes en cada una de las seis etapas del ciclo de vida de la IA, garantizando que los impactos sobre los DD.HH se tomen en cuenta a lo largo del proceso (en vez de después de que el sistema ya está terminado e implementado). Al integrar la evaluación del impacto en los derechos humanos del Instituto Alan Turing con nuestro enfoque práctico basado en los derechos humanos del ciclo de vida de la IA y el desarrollo de la IA, esta metodología facilita el cumplimiento de los próximos requisitos políticos, como la evaluación del impacto en los derechos humanos de la Ley de IA de la UE.

Sin embargo, nuestro objetivo es ir más allá del mero cumplimiento y **avanzar hacia un paradigma de desarrollo de la IA que promueva de forma proactiva los Derechos Humanos**, en lugar de mitigar los riesgos como un añadido o después de que se hayan producido los daños. Al **involucrar a las comunidades desde el principio** y con un poder de

decisión sustancial, promovemos y posibilitamos el desarrollo de sistemas que centran los Derechos Humanos, la igualdad y la inclusión en el núcleo del código, capaces de crear nuevas oportunidades y corregir de forma innovadora las desigualdades. Esperamos poner los programas sociales en consonancia con la investigación y los valores del siglo XXI, unidos en la búsqueda de formas de hacer que la IA sea más eficaz, no simplemente más «precisa» y «eficiente».

¿Cuál es el objetivo de un enfoque basado en los derechos humanos?

La IA está afectando a todas las partes de la sociedad e incluso cuando ha sido bienintencionada ha perjudicado o explotado repetidamente a comunidades, y especialmente a grupos vulnerables. Creemos que muchos de estos daños pueden evitarse mediante **puntos de reflexión críticos** desde la fase conceptual, a lo largo de todo el desarrollo de la IA y después de él. Estos puntos de reflexión promueven un **cambio de paradigma en la creación de la IA**, alejándose de los objetivos impulsados principalmente por la tecnología independiente y acercándose a la creación de sistemas sociotécnicos en colaboración **con las comunidades** con las que el sistema interactuará y a las que afectará.

Es probable que este enfoque dé lugar a sistemas más robustos, que permitan una asimilación, un uso y una evolución más eficaces de la tecnología, con el potencial de empoderar a las comunidades y a los ciudadanos para que alcancen y disfruten sus derechos humanos. También dará lugar a sistemas y soluciones con menos riesgo de afectar negativamente a los derechos humanos de las comunidades a las que la tecnología está destinada a servir.

¿Por qué un enfoque basado en los derechos humanos frente a una IA "ética" o responsable?

La ética, que tiene una importancia crucial, también es contextual o **situacional**[1]. Los principios éticos y responsables de la IA, elaborados por una amplia gama de organismos (por ejemplo, el mundo académico, las organizaciones de la sociedad civil, los institutos de investigación, los gobiernos y el sector privado) son la respuesta más común a las preocupaciones en torno a la ética de la IA[2]; sin embargo, son objeto de importantes críticas

[1] Sadek, M., Kallina, E., Bohné, T. et al. Challenges of responsible AI in practice: scoping review and recommended actions. AI & Society (2024). https://doi.org/10.1007/s00146-024-01880-9
[2] Jobin, A., Ienca, M. and Vayena, E. (2019) The Global Landscape of AI Ethics Guidelines. Nature Machine Intelligence, 1, 389-399. https://doi.org/10.1038/s42256-019-0088-2

por parte del mundo académico[3][4] y de la práctica de la IA[5][6]. Su carácter **abstracto** permite interpretaciones y aplicaciones divergentes, lo que impide o incluso socava la rendición de cuentas.

Evitamos esta ambigüedad centrándonos en los Derechos Humanos, un corpus consensuado de legislación internacional (y nacional) que refleja una **comprensión universal** de los aspectos necesarios para garantizar la dignidad humana centrándose en la igualdad y la no discriminación, la participación y la inclusión, la rendición de cuentas y el Estado de Derecho, que son principios indivisibles e interdependientes de los derechos humanos[7]. Así pues, los Derechos Humanos proporcionan un **punto de partida común y concreto** para alinear a diferentes actores, disciplinas y culturas.

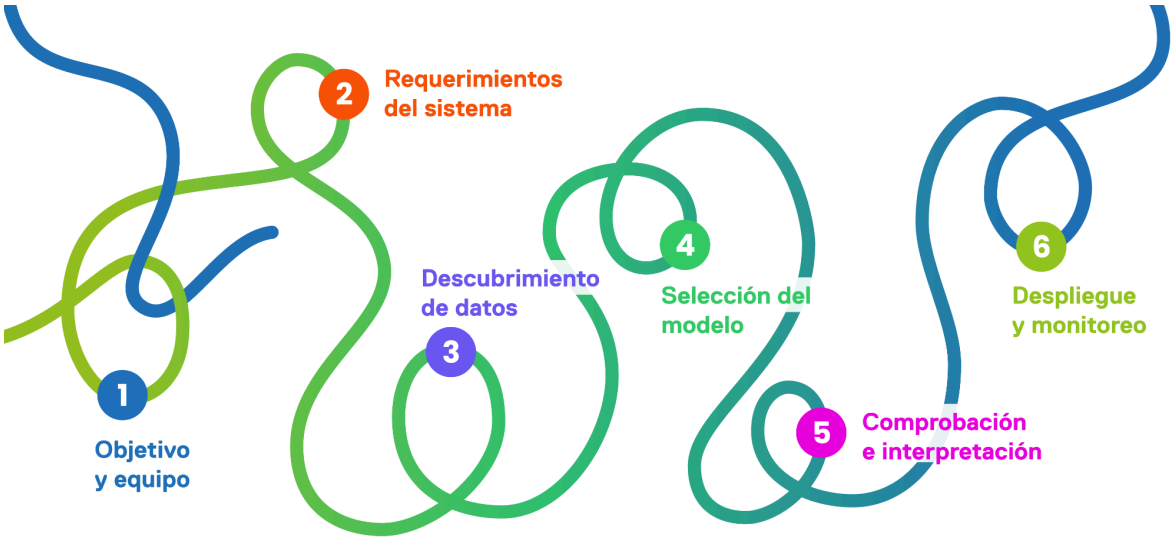
Además, nuevas políticas como la Ley de IA de la UE exigen **evaluaciones del impacto sobre los derechos humanos (EIDH)** por parte de quienes despliegan o adquieren tecnologías de alto riesgo, como la IA utilizada en recursos humanos, educación, decisiones financieras o asistencia sanitaria. Dado que actualmente no existe ninguna EIDH oficial como parte de la Ley de IA de la UE ni en ninguna otra parte, varios organismos e institutos de investigación están desarrollando sus versiones de EIDH. Después de revisar varias, decidimos integrar **la muy completa EIDH del Instituto Alan Turing en nuestro marco**, es decir, plantear las preguntas y reflexiones cubiertas por la HRIA en las etapas del ciclo de vida en las que se vuelven relevantes. De este modo, permitimos un enfoque del desarrollo de la IA que tiene en cuenta los aspectos pertinentes a lo largo de **todo el proceso de desarrollo**, en lugar de hacerlo como un añadido una vez desarrollado el sistema, es decir, en el momento de la adquisición. De este modo, los responsables del despliegue o la adquisición pueden revisar todas las medidas adoptadas, lo que facilita enormemente la rendición de cuentas y la transparencia, así como el proceso de realización de EIDH antes del despliegue. Por consiguiente, orientar nuestro marco en función de los derechos humanos tiene la ventaja adicional de que **facilita el cumplimiento de la próxima normativa sobre IA.**

[3] McNamara, A., SmithJ., and Murphy-Hill, E (2018). Does ACM's Code of Ethics Change Ethical Decision Making in Software Development?, https://doi.org/10.1145/3236024.3264833
[4] Munn, L. The uselessness of AI ethics. AI Ethics 3, 869–877 (2023). https://doi.org/10.1007/s43681-022-00209-w
[5] Ibáñez, J., Olmeda, Mónica (2022). Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study. AI and Society 37 (4):1663-1687
[6] Rakova, Bogdana, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. "Where Responsible AI Meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices." Proceedings of the ACM on Human-Computer Interaction 5 (April 13, 2021): 1–23. https://doi.org/10.1145/3449081
[7] United Nations. 1948. Universal Declaration of Human Rights.

El ciclo de vida de la IA

Para garantizar que nuestras recomendaciones sean **accionables para los profesionales de la IA**, hemos anclado **nuestras preguntas de reflexión sobre <IA & Equidad>** en el ciclo de vida de la IA, combinándolas con la EIDH del Instituto Alan Turing. El ciclo de vida no es estrictamente lineal, sino que está **entrelazado y es cíclico**, como un hilo que se repite una y otra vez. Esto subraya la importancia de **reflexionar, revisar y perfeccionar** a medida que aprendemos más sobre el contexto sociotécnico, los datos, el modelo y la **integración de las consideraciones basadas en los derechos humanos a lo largo del ciclo de vida de la IA**, en lugar de como un añadido después de que el sistema se haya desarrollado o incluso contemplado o programado para su uso.

Identificamos las siguientes seis etapas del ciclo de vida la IA:



IA & EQUIDAD
WOMEN AT THE TABLE | CENIA

Ciclo de vida de la Inteligencia Artificial

1. Objetivo + Composición del equipo
2. Definición de los requisitos del sistema
3. Descubrimiento de datos
4. Selección y desarrollo de un modelo
5. Pruebas e interpretación de resultados
6. Despliegue y supervisión posterior

Preguntas esenciales por etapa del ciclo de vida de la IA

En las siguientes secciones, ofreceremos una breve visión general sobre las seis etapas, los conceptos cruciales y las cuestiones esenciales sobre las que los creadores de IA deben reflexionar en cada etapa específica (cuadro morado).

¿Cómo abordar las preguntas de reflexión?

Es esencial responder las preguntas solo o solamente con tu equipo. Por el contrario, para muchas preguntas es esencial **debatir las preguntas y las posibles respuestas con representantes de las comunidades específicamente afectadas y, especialmente, con grupos históricamente marginalizados.** Además, sus respuestas pueden cambiar a medida que vaya aprendiendo cosas nuevas, así que **no dude en revisar y modificar sus respuestas.**

Evaluación de impacto en derechos humanos (EIDH) del Instituto de Alan Turing

El Instituto Alan Turing publicó una versión en proceso de su marco de garantía de los DD.HH, la democracia y el Estado de Derecho para los sistemas de IA. **Ubicamos las áreas cubiertas en su plantilla EIDH (véanse las páginas 251 a 276[8]) a lo largo del ciclo de vida de la IA** para permitir un desarrollo de la IA que las tenga en cuenta antes de su despliegue, y también en la fase del ciclo de vida en la que **se vuelven relevantes**. De esta manera, ayudamos a construir sistemas con los DD.HH. en su núcleo, lo que **no sólo implica el cumplimiento de la EIDH, sino que hace que el proceso de llevar a cabo EIDH previas al despliegue sea más fácil, eficiente y eficaz.**

[8] The Alan Turing Institute, Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal (2022). P.251-276, https://doi.org/10.5281/zenodo.5981675

Etapa 1. Definición del objetivo y composición del equipo

A. Definición del objetivo

Es **esencial empezar por el objetivo y la finalidad de un sistema:** Siempre debe quedar claro por qué se necesita un sistema en particular, qué problema resuelve y para quién. Frecuentemente esta visión sólo refleja las necesidades de las personas que desarrollan el sistema de forma aislada, que tienen una gran influencia en este contexto, en contraposición a las necesidades de las comunidades a las que el sistema está destinado a servir y afectar.

Por lo tanto, es esencial implicar a las comunidades afectadas desde el principio mediante prácticas de **desarrollo participativ**o (véase el recuadro). Para empezar, se debe consultar a la comunidad afectada y acordar que **un sistema de IA es la mejor manera de ayudar a resolver su problema,** ya que puede haber formas más sencillas, eficientes y rentables de abordar el problema central.

Desarrollo participativo:
En este contexto, se refiere al proceso de creación de tecnología en colaboración con las comunidades afectadas[9]. Esto incluye explorar sus necesidades, valores y preocupaciones en el contexto de la aplicación y abordarlas en el diseño del sistema.

Las comunidades afectadas pueden ser usuarios del sistema (por ejemplo, un hospital, un banco, la Administración), usuarios del sistema (por ejemplo, radiólogos, empleados de un banco, funcionarios), las personas sobre las que se utiliza el sistema (por ejemplo, un paciente, alguien que solicita un préstamo, un ciudadano), así como las comunidades más vulnerables.

En este caso, es esencial que todas las **comunidades afectadas** (y no sólo los grupos críticos para los ingresos) participen y tengan **poder real de decisión y agencia en el proceso**. De este modo se evita una forma extractiva de desarrollo participativo, en la que se recogen las necesidades de la comunidad pero se desatiende su aplicación por intereses comerciales o agendas internas.

---

[9] Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2023, October). The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*

B. Composición del equipo

En la creación y el funcionamiento de un sistema de IA intervienen muchas personas, ¡muchas más que las que escriben el código! El **objetivo** de un sistema debe **informar fundamentalmente la composición de su equipo de creadores**, es decir, qué tipos de conocimientos y experiencia vivida son necesarios para hacer realidad plenamente el objetivo previsto. Esto incluiría no sólo los conocimientos y aptitudes técnicas necesarios, sino también las **diversas procedencias, perspectivas y experiencias** con el entorno para el que se desarrolla el sistema. Queremos destacar dos papeles que a menudo se olvidan: las comunidades afectadas y los científicos sociales.

Comunidades afectadas:

Las comunidades afectadas son las **expertas en el contexto en el que se desplegará el sistema** (es decir, en su experiencia vivida) y cargarán con las consecuencias del despliegue del sistema. Debe prestarse especial atención a las comunidades ya marginadas, ya que los sistemas de IA pueden tener efectos particularmente adversos en la capacidad de estas comunidades para participar plena y significativamente en los nuevos sistemas que se creen[10]. Las aportaciones de las comunidades afectadas contribuyen a crear sistemas más adecuados, garantizan una mayor aceptación y ayudan a prever riesgos y daños.

Preguntas esenciales en esta etapa:

Objetivo y contexto del sistema

- ¿Qué problema intenta resolver el sistema?
  - ¿Existe un historial de discriminación en el ámbito en cuestión?
  - ¿Existe el riesgo de que su sistema pueda reforzar o perpetuar resultados históricamente desiguales?
  - ¿Cómo puede contrarrestar dicha discriminación histórica?
- ¿El sistema tendrá una función esencial o de alto riesgo, o se implementará en

[10] Buolamwini & Gebru (2018) http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf; Angwin et al. (2016) 'Machine Bias'. ProPublica, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

un sector de alto impacto o crítico para la seguridad (véase, por ejemplo, la Ley de IA de la UE)?
    - ¿Cómo se garantiza un funcionamiento seguro, tanto en el diseño como en caso de fallo del sistema?
- ¿Se ha entablado un diálogo sobre el sistema con las comunidades afectadas por él?
    - ¿Es un sistema de IA la mejor manera de abordar el problema?
    - ¿Responde a las necesidades más acuciantes de la comunidad?
    - ¿Hay comunidades vulnerables, por ejemplo, debido a características protegidas?
- ¿Se supone que el sistema se implementará a gran escala? ¿Es esto prudente?
- ¿El uso del sistema o el hecho de que el sistema se utilice en alguien es voluntario (uso directo e indirecto)?

Efectos del sistema

- ¿Quién se beneficia del sistema y quién puede verse perjudicado?
    - ¿Refleja o nivela las estructuras de poder actuales?
    - ¿Cómo podemos involucrar a las comunidades y, en especial, a los grupos históricamente marginados?
- ¿Contribuye activamente el sistema a los derechos humanos?
    - ¿Ha realizado una primera evaluación del impacto sobre los derechos humanos para identificar los riesgos antes de invertir recursos (págs. 21 a 47)? Entre los riesgos potenciales se incluyen la manipulación, la discriminación o la protección de las estructuras de poder actuales.
        - ¿Qué ocurre si el sistema se utiliza de forma no prevista?
    - ¿Contribuye el sistema a promover los principios y prioridades de los derechos humanos?
    - ¿A quién se debe incluir o consultar durante esta evaluación?
    - ¿Cómo se garantiza que los riesgos identificados se eliminen o mitiguen?
- ¿Quién es responsable de las inexactitudes y los daños resultantes?
    - ¿Cómo se documentan las decisiones de diseño del sistema, las responsabilidades y las obligaciones generales para que se puedan rastrear?
    - ¿Ha tenido en cuenta todas las preguntas anteriores (especialmente las repercusiones en los derechos humanos) para toda la cadena de valor de su sistema, por ejemplo, para proveedores, subcontratistas, auditores, etc.?

        - ¿Cómo se garantiza el escrutinio continuo y exhaustivo de la cadena de valor?

Empoderamiento de las comunidades afectadas

- ¿Cómo pueden las comunidades afectadas estar representadas en el equipo para que este pueda beneficiarse de sus conocimientos y experiencia en el mundo real?
- Además de la pertenencia al equipo, ¿cómo involucra el equipo a las comunidades afectadas?
    - ¿Estas comunidades reciben la capacidad necesaria para influir en las decisiones?
    - ¿Tiene el equipo de desarrollo la mentalidad y las habilidades necesarias para lograrlo?

Composición del equipo

- ¿Qué conocimientos especializados necesita en su equipo?
- ¿Tienes diversidad en cuanto a cultura, demografía, experiencias vividas, disciplinas y habilidades (sociotécnicas, jurídicas, antropológicas, UX, técnicas, entre otras)?
- ¿Cómo garantizas jerarquías planas y la comunicación entre las disciplinas?
- ¿El equipo tiene:
    - Conciencia de los riesgos que los sistemas de IA suponen para los derechos humanos y las razones subyacentes?
    - Conocimientos y experiencia sobre el problema que están tratando de resolver?
    - Conocimientos y experiencia sobre posibles soluciones para este problema?

Etapa 2: Definición de los requisitos del sistema

En la segunda etapa, el objetivo del sistema se formaliza en una lista de requisitos, d**esarrollada nuevamente en diálogo entre diversas funciones y comunidades**. Esto incluye gestionar los compromisos entre las diferentes necesidades y los requisitos deseados, ya que los sistemas existen en un **ecosistema de valores.**

Ecosistema de valores

Diferentes aspectos de un sistema lo vuelve un sistema responsable. Algunos ejemplos son que sus decisiones sean justas (**equidad**[11]), que sus decisiones sean fáciles de entender (**explicabilidad**), que su proceso de desarrollo y sus motivaciones subyacentes sean claros (**transparencia**) o que funcione con pocos errores (**precisión**). Es **imposible optimizar todos estos aspectos simultáneamente en la misma métrica**, por lo que es necesario hacer **concesiones** (aunque estas concesiones no reducen necesariamente la precisión de manera fundamental). Por ejemplo, los modelos altamente explicables suelen tener menos precisión que las formas más opacas de modelos de IA.
En algunos contextos, la explicabilidad puede ser tan importante (o incluso más) que la minimización de errores (precisión): sólo si la persona que supervisa el sistema puede comprender y cuestionar los resultados, podrá detectar y corregir los errores, lo que en última instancia conducirá a menos errores que la alta precisión por sí sola. Por lo tanto, es esencial no centrarse únicamente en una métrica (como se suele hacer con la precisión), sino **tomar una decisión consciente sobre la jerarquía y la importancia de las métricas en el contexto específico.**

Es importante destacar que **la precisión nunca debe considerarse sin tener en cuenta la equidad,** ya que puede ocultar una distribución desigual de la precisión, por ejemplo, que el sistema sea muy preciso para la mayoría de los casos, pero muy impreciso para un grupo minoritario. Esto puede tener repercusiones negativas en los derechos humanos, en la atención médica, el reconocimiento facial, las finanzas, las subvenciones y otros sectores importantes.

El proceso de definición de los requisitos del sistema debe ser iterativo y fluido; es muy probable que **la lista de requisitos cambie a medida que se conozcan más detalles sobre el contexto social y las necesidades de las comunidades afectadas.** Por lo tanto, es importante proporcionar una plataforma en la que los operadores y las comunidades afectadas puedan comunicar al equipo cualquier información nueva que pueda influir en los requisitos.

[11] Traducido del inglés *fairness*.

Preguntas esenciales en esta etapa:

Participación de las comunidades afectadas

- ¿**Quiénes** deben participar en la definición de los requisitos del sistema? ¡No se limite a pensar en los operadores, los usuarios o las partes con ingresos críticos!
- ¿Existen **tensiones** entre los objetivos del sistema y las necesidades de las comunidades afectadas? ¿Cómo se pueden abordar estas tensiones, dando siempre prioridad a los derechos humanos?
- ¿Ha **revisado** su evaluación inicial del impacto sobre los derechos humanos, ahora que se prevén más capacidades?
- ¿Ha solicitado la **opinión de expertos**, por ejemplo, de las comunidades afectadas con experiencia vivida, un departamento gubernamental (o un departamento gubernamental aliado), el mundo académico o un organismo público?

Consideraciones sobre la explicabilidad

- ¿Cuál es el **objetivo** de las explicaciones?
    - ¿Quién es el **público** de las explicaciones y por qué?
    - ¿Las explicaciones estarán disponibles para todas las comunidades afectadas, lo que facilitará el escrutinio público?
    - ¿Las explicaciones proporcionadas son fáciles de procesar para todos los públicos a los que van dirigidas?
- ¿Ha considerado qué **aspectos** de la explicabilidad son los más relevantes?
    - Por ejemplo, cómo se toman las decisiones en general, cómo se tomó una decisión concreta, etc.
- ¿Cómo puede utilizar las explicaciones para aumentar la **capacidad de acción** de las comunidades afectadas, por ejemplo, detallando qué tendría que cambiar para obtener un resultado diferente (explicación contrafactual)?
    - ¿Cómo se asegura de que sus explicaciones ayuden a las comunidades afectadas a comprender los límites y los impactos del sistema?

Ecosistema de valores

- ¿Existen **tensiones** entre la precisión y otras métricas más necesarias en este contexto?
- **Equidad**: ¿Qué métricas de equidad considera que serían útiles en este contexto? ¡Explore varias!
- **Privacidad**: ¿Se respeta la privacidad de todas las comunidades afectadas y los interesados?
    - ¿Cómo se puede minimizar la recopilación de datos en esferas privadas, por ejemplo, en los hogares?
    - ¿Vale la pena la intrusión restante?
- **Transparencia**: ¿Cómo permitirá a las comunidades afectadas acceder a la información sobre su metodología, por ejemplo, los datos de entrenamiento, el proceso analítico, cómo se entrenó el modelo, los metadatos de diversas métricas?
    - ¿Cómo puede garantizar que las comunidades afectadas sean conscientes de que están utilizando un sistema de IA o de que se utiliza sobre ellas?
- **Responsabilidad**: ¿Cuál es la estructura de responsabilidad?
    - ¿Qué supervisión humana se debe buscar?
    - ¿Qué conocimientos y formación necesitarán las personas que participen en el proceso?
    - ¿Cómo puede permitir que las comunidades afectadas impugnen un resultado?
- **Usabilidad**: ¿Cómo podemos garantizar que la interfaz sea intuitiva y accesible para todos?

Etapa 3: Descubrimiento de datos

Un objetivo del sistema relevante y sus requisitos pueden verse comprometidos si el conjunto de datos utilizado para entrenar el sistema de IA **no es representativo de su caso de uso y contexto**. Un buen ajuste sociocultural del conjunto de datos incluye diversos aspectos, como la demografía de las personas que lo componen, su cultura o factores ambientales. Será imprescindible consultar a expertos en la materia para garantizar que se recojan adecuadamente los aspectos relevantes.

Si no se encuentra un conjunto de datos que se ajuste bien, es posible que el equipo tenga que generar uno nuevo, ya sea recopilando datos nuevos o mejorando o ampliando los conjuntos de datos existentes mediante pasos de **preprocesamiento** (es decir, matemáticos).

> El **preprocesamiento** se refiere a la **manipulación y transformación de los datos brutos antes de introducirlos en un modelo**. Implica diversas técnicas para mejorar la calidad, la relevancia y la imparcialidad de los datos, por ejemplo, equilibrando la frecuencia de una clase específica (como el género o la raza) en el conjunto de datos, de modo que el modelo se entrene por igual en ellas.

Preguntas esenciales en esta etapa:

Origen de los datos

- ¿**Quién** recopiló los datos y con qué finalidad?
- ¿Los interesados dieron su consentimiento para el uso de sus datos?
    - ¿Se respetó su **privacidad**?
- ¿Qué grado de **confidencialidad** tiene la información? Por ejemplo, ¿los datos revelan atributos sensibles como el origen racial o étnico, la orientación sexual, el estado de salud o las creencias religiosas?
    - ¿Existe alguna forma de anonimizar los datos personales para que se respete la privacidady se pueda obtener información sobre la edad, el género y la ubicación geográfica?

Sesgos en los datos

- ¿Quiénes están **representados** en los datos? ¿Quiénes están **excluidos**? ¿Por qué podría ser así?
    - ¿Qué **regiones geográficas y cultura**s están incluidas y cuáles no?
    - ¿Qué **consecuencias** tiene esto para el funcionamiento de su sistema?
- ¿Qué **sesgos históricos o actuales** podrían existir en los datos, con el riesgo de comprometer los derechos humanos?
- ¿Qué tipo de **preprocesamiento** de datos son necesarios para crear un modelo que sea justo en este contexto?
- En su caso de uso específico, ¿es más beneficioso ignorar (mostrar la posible injusticia en los datos), "borrar" (eliminar la posible injusticia en los datos) o incluso contrarrestar (contrarrestar este sesgo de manera que el grupo desfavorecido pase a estar favorecido) este sesgo?

Documentación

- ¿Ha **documentado** qué conjuntos de datos está utilizando y por qué los ha elegido, de modo que los posibles implementadores puedan evaluar si sus datos de entrenamiento se ajustan a su contexto?
- ¿Ha documentado todos los **pasos de preprocesamiento** que ha seguido (información esencial para futuros usos de su sistema o código)?
- ¿Ha guardado sus datos "sin procesar", además de los datos preprocesados, para poder utilizarlos en el futuro?

Etapa 4: Selección y desarrollo de un modelo

Es hora de considerar **qué tipo de modelo de IA es el mejor para satisfacer los requisitos del sistema.** ¡No siempre es el algoritmo de aprendizaje profundo más complicado!

En cambio, se trata de **elegir el modelo más adecuado para el alcance requerido,** al tiempo que se gestionan los compromisos ya identificados. Por ejemplo, los modelos menos complejos suelen ser más explicables, pero pueden alcanzar una precisión ligeramente inferior. Dado que la explicabilidad es un requisito previo para una buena detección de errores y sesgos, estos modelos parecen especialmente importantes en escenarios de alto riesgo. Por ejemplo, el Banco Central Europeo exige un alto nivel de explicabilidad para las decisiones de calificación crediticia y, por lo tanto, excluye las redes neuronales y otros

tipos de algoritmos menos explicables que impiden el descubrimiento de resultados discriminatorios y el escrutinio.

El desarrollo del modelo en sí mismo es un **proceso iterativo** en el que se ajustan diferentes aspectos del modelo para cumplir con los distintos requisitos del sistema (por ejemplo, mediante métodos de procesamiento previo o posterior, o ajustando las ponderaciones o los parámetros de un modelo). En este sentido, es importante reflexionar sobre las etapas anteriores para garantizar que el objetivo, los requisitos, los datos y el modelo estén alineados.

> Los métodos de **procesamiento interno (in-processing)** están diseñados para mitigar el sesgo y aumentar la imparcialidad mientras se entrena el modelo, mientras que los métodos de **posprocesamiento** incluyen la modificación de los resultados del modelo una vez completado el entrenamiento.

Preguntas esenciales en esta etapa:

Requisitos del tipo de modelo y explicabilidad

- El modelo desarrollado:
  - ¿Logra una explicabilidad adecuada, teniendo en cuenta lo que está en juego en la situación?...
  - ¿**Minimiza** la complejidad?
  - ¿Alerta al usuario si no está seguro de una decisión y/o cuando se enfrenta a un caso que no se refleja suficientemente en sus datos de entrenamiento (por ejemplo, un modelo entrenado solo con piel clara con poco pigmento se enfrenta a un caso de piel oscura con más pigmento, lo que alerta al usuario de que no sabe cómo clasificar este caso)?

Aspectos relacionados con la equidad

- ¿Cuál es la **métrica de equidad** más adecuada y por qué?
- ¿Han probado diferentes métricas y resultados?
- ¿Qué **aspectos de la equidad** se tienen en cuenta, por ejemplo, en función del género, el origen étnico, la educación...?
  - ¿Ha tenido en cuenta las **interseccionalidades** pertinentes?

- ¿Se ha asegurado de que el modelo no se base en variables o indicadores que puedan ser injustamente discriminatorios? Por ejemplo, el código postal de una persona podría permitirle inferir su origen étnico.
- ¿Por qué se han elegido determinados **pasos de procesamiento interno (modelo) y posterior (evaluación)**?

Otros

- ¿El modelo es **transparente** para las comunidades afectadas? Es decir, quién lo financió, cuál es su objetivo, quiénes participaron, datos de entrenamiento, rendimiento, etc.
- ¿Cuál es el **impacto ambiental** del modelo? ¿Vale la pena el costo?
- ¿Se han realizado esfuerzos para minimizar o compensar el impacto ambiental?

Etapa 5: Testeo e interpretación de resultados

Una vez desarrollado el modelo, debemos **comprobar si cumple los requisitos del sistema definidos por el equipo en la fase 2**. Para algunas métricas, esto se puede hacer mediante **pruebas técnicas,** mientras que **otras requieren la opinión de las comunidades afectadas**[12], por ejemplo, si se ha alcanzado el nivel de explicabilidad previsto.

Para las pruebas técnicas, es importante que e**l conjunto de datos de prueba sea tan representativo del contexto como el conjunto de datos de entrenamiento.** La inclusión de ejemplos o casos extremos puede ayudar a descubrir posibles problemas que pueden no ser evidentes durante las pruebas rutinarias, revelando así cualquier limitación o debilidad en el rendimiento del modelo.

Los insights adquiridos deben plasmarse en un “manual de funcionamiento” (u otro tipo de documentación adecuada) que se entregará a los **futuros usuarios u operadores del sistema.** Al indicar los contextos para los que se ha entrenado el sistema (en los que se espera que funcione bien) y aquellos en los que no (en los que es probable que haya imprecisiones), los operadores pueden calibrar su confianza y adherencia en consecuencia.

[12] Mireia Yurrita, Dave Murray-Rust, Agathe Balayn, and Alessandro Bozzon. 2022. Towards a multi-stakeholder value-based assessment framework for algorithmic systems. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 535–563. https://doi.org/10.1145/3531146.3533118

Además, el manual debe incluir recomendaciones sobre **el nivel de supervisión humana necesario,** lo que permitirá una formación adecuada de los operadores.

Preguntas esenciales en esta etapa:

Contexto y resultados de las pruebas

- ¿El sistema **cumple con el objetivo y los requisitos del sistema**?
    - ¿Qué **medidas de rendimiento** del modelo se incluyen y por qué se seleccionaron otras en lugar de otras (incluidos aspectos cuantitativos Y cualitativos)?
    - ¿Sigue siendo válida esta selección después de haber obtenido más información sobre el contexto de la aplicación? ¿Deberíamos añadir algo?
    - ¿Qué **opiniones** se han tenido en cuenta en estas pruebas?
- ¿Se puede dar a conocer el modelo entrenado al **público o a expertos externos** para que lo prueben y examinen con el fin de detectar posibles problemas?
- ¿Se ha **probado** el modelo en un contexto lo más parecido posible a su aplicación real (incluidos sus usuarios reales) para identificar posibles daños?
- ¿Se han **incluido** los aprendizajes y comentarios resultantes?

Manual de funcionamiento

- ¿Existe un manual **fácil de entender** para los operadores del sistema?
- ¿Qué podemos recomendar como **mejores prácticas** en torno al funcionamiento, por ejemplo, cuánta supervisión humana se requiere y con qué experiencia?
- ¿Para qué **contextos** se ha entrenado el sistema?
    - ¿En qué casos podría resultar injusto o inexacto?
- ¿Cómo se capacitará a los operadores sobre cómo utilizar e interpretar el sistema, incluyendo cómo calibrar su confianza y capacidad para cuestionar el funcionamiento del sistema?
- ¿Cómo se registrarán los cambios futuros en el sistema?

Etapa 6: Implementación y postimplementación, auditoría y monitoreo

Implementación: La etapa de implementación es la **última verificación de integridad del sistema,** es decir, si se han considerado, comunicado y tenido en cuenta todos los daños, impactos discriminatorios y consecuencias. Revise su **Evaluación de Impacto sobre los Derechos Humanos** inicial, revisandola de manera más exhaustiva ahora que conoce el sistema completo, para asegurarse de que se haya evaluado el impacto negativo del sistema sobre los derechos humanos en su forma final.

La decisión sobre si el sistema está listo para ser implementado es muy importante. Recomendamos **empoderar verdaderamente a las comunidades afectadas**; al fin y al cabo, son ellas las que tienen que soportar las consecuencias de un funcionamiento defectuoso. Además, es fundamental **establecer vías que permitan a los operadores y a las comunidades más afectadas alertar sobre los problemas** que experimentan en relación con el sistema.

Después de la implementación: El sistema debe ser auditado y monitoreado regularmente en después de la implementación, incluyendo oportunidades para que las **comunidades afectadas proporcionen comentarios.** Esto es especialmente relevante inmediatamente después de la implementación, ya que el sistema recién implementado podría exponer desafíos o problemas previamente desconocidos.

Incluso si el sistema funciona según lo previsto, **es probable que el contexto de aplicación del modelo cambie con el tiempo.** Esto no solo puede alterar los datos de entrada o los resultados que se consideran justos, sino que incluso puede afectar al objetivo, por ejemplo, hacer que el objetivo quede obsoleto, de modo que el sistema deba retirarse. Por lo tanto, es **esencial auditar continuamente el sistema,** incluyendo tanto auditorías cuantitativas como cualitativas en colaboración con las comunidades afectadas (véase, por ejemplo, un marco para poner en práctica dichas auditorías[13]). Aquí puede consultarse una descripción detallada de los diferentes tipos de auditorías, incluidas las realizadas por terceros externos[14].

[13] Yurrita, M., Murray-Rust, D., Balayn, A., & Bozzon, A. (2022, June). Towards a multi-stakeholder value-based assessment framework for algorithmic systems. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 535-563). https://dl.acm.org/doi/abs/10.1145/3531146.3533118
[14] Birhane, A., Steed, R., Ojewale, V., Vecchione, B., & Raji, I. D. (2024, April). AI auditing: The broken bus on the road to AI accountability. In 2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) (pp. 612-643). IEEE. https://ieeexplore.ieee.org/abstract/document/10516659

Preguntas esenciales en esta etapa:

Implementación

- ¿**Quién decide** que el modelo está listo para ser implementado?
    - ¿Los reguladores, los expertos en la materia y las comunidades afectadas han dado su consentimiento para la implementación?
    - ¿Las comunidades más afectadas tienen la capacidad de retrasar o detener la implementación?
- ¿Ha revisado su evaluación inicial del impacto sobre los DD.HH[15] y ha realizado una más exhaustiva, ahora que se conocen todas las capacidades del modelo?
- Antes de la implementación: ¿Existen procesos para **detectar posibles fallos del sistema o daños inesperados?**
    - ¿Son los tomadores de decisiones responsables de los daños que puedan causarse?
- ¿Qué **mecanismos** existen para cuando se identifica un problema?
    - ¿Quién es responsable de abordar los daños que puedan producirse? ¿Cuál es el plazo?

Monitoreo

- ¿Existen procesos o funciones que permitan a los operadores y a las comunidades afectadas **alertar sobre posibles inexactitudes o fallos del sistema**?
- ¿Cómo se puede garantizar que las comunidades afectadas puedan **optar por no utilizar el sistema**?
- ¿Cómo se supervisan los **cambios en el contexto**?
    - ¿Cuál es el proceso para conocer los nuevos riesgos o daños?
    - ¿Cuál es el mecanismo para conocer las nuevas necesidades de los usuarios sobre el terreno?
    - ¿Cómo podemos incluirlas en los requisitos y tenerlas en cuenta?
    - ¿En qué casos es mejor desconectar el sistema hasta que se hayan tenido en cuenta los riesgos?
    - ¿Cómo comprobará que el modelo sigue cumpliendo su objetivo?
    - ¿Cómo sabría que es el momento de retirar el sistema?

[15] The Alan Turing Institute, Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal (2022). P.251-276, https://doi.org/10.5281/zenodo.5981675

Resumen

Hemos destacado una serie de preguntas esenciales a lo largo de las seis etapas del ciclo de vida de la IA para que los creadores de este tipo de tecnología puedan reflexionar sobre **los objetivos que buscan, el impacto que pueden tener en los derechos humanos y los efectos sociales más amplios** de los sistemas que crean en colaboración con las comunidades afectadas por ellos.

Queremos destacar que estas preguntas, como mínimo, **facilitan la creación de tecnología que cumple con los principios de derechos humanos** de igualdad y no discriminación, participación e inclusión, rendición de cuentas y estado de derecho. Sin embargo, estas preguntas pueden ayudar a ir más allá del mero cumplimiento y permitir la creación de tecnologías que:

- se **guíen** por los principios de los derechos humanos,
- **contribuyan** a su acceso y cumplimiento, y
- aspiren a **empoderar a los seres humanos y a los responsables** de cumplir con sus obligaciones para que alcancen y disfruten de sus derechos humanos.

Únete a la comunidad <AI & Equality>

Comunidad interdisciplinaria global que reúne a personas apasionadas por la IA inclusiva y basada en los derechos humanos.

AI & EQUALITY