PROPUESTA A LA UE RESPECTO A LA GOBERNANZA EN LOS SISTEMAS DE IA

Tal como se ha descrito con anterioridad, la IA debe ser fiable, lo que significa que debe ser lícita, robusta y ética basándose en principios fundamentales tales como el respeto a la autonomía humana, la prevención del daño, la equidad y la explicabilidad.

Esta premisa, como se ha visto, requiere que los sistemas de IA cumplan una serie de requisitos (acción y supervisión humanas, solidez técnica y seguridad, gestión de la privacidad y de los datos, transparencia, diversidad, no discriminación, equidad, gestión social y medioambiental y rendición de cuentas) que son una lista abierta la cual debe ser revisada continuamente debido al crecimiento exponencial que experimenta la tecnología y, en concreto, la de la IA, que puede hacer que estos requisitos vayan mutando para adaptarse a la misma.

En consecuencia, analizando los principios fundamentales y los requisitos expuestos, una de las propuestas que se pueden elevar a la Comisión Europea es la de crear una certificación o evaluación de conformidad de los sistemas de IA, entendida como el proceso que se lleva a cabo para verificar si se cumplen o no los requisitos específicos relativos a un proceso, servicio o producto, antes de que éste llegue al usuario final o se inserte en un bien o servicio ya existente puesto que su aplicación englobaría el cumplimiento de todos los principios y requisitos descritos anteriormente para que el sistema de IA certificado fuera fiable:

- El **principio de la autonomía humana** puesto que se verificaría que el sistema de IA está diseñado para beneficiar las aptitudes cognitivas, sociales y/o culturales de la personas así como la existencia de un control humano sobre los procesos creados, siguiendo diseños antropocentristas y garantizando la acción y supervisión humana, tanto como requisito del sistema para su funcionamiento, como por la propia certificación al ser realizada por personas.

- El **principio de prevención del daño** ya que se comprobaría el estado de la técnica utilizada en el sistema específico validando su solidez y seguridad técnicas, asegurándose que no puedan ser utilizados para fines malintencionados y teniendo un plan de actuación ante posibles fallos del sistema.

Asimismo, se revisaría la protección a la dignidad humana que el sistema de IA ofrece, incluyendo en la evaluación, medidas que exijan la información necesaria prestando atención a los posibles efectos adversos que puedan afectar a personas vulnerables.

Igualmente, se verificaría la adecuada gestión de los datos a lo largo de todo el ciclo de vida del sistema de IA (protocolos de acceso, consentimiento de los interesados, en su caso, tipo de datos tratados, datos generados antes y después de sus tratamiento en el sistema de IA...) la cual, en todo momento, debe cumplir del RGPD y las legislaciones nacionales al efecto.

- El **principio de equidad** se constataría al garantizar, con las comprobaciones oportunas, que no se producen sesgos discriminatorios en el conjunto de datos utilizados (sobre todo, en la parte de recopilación de información) ni sesgos injustos en los algoritmos incluyendo la supervisión y la corrección humanas en las distintas fases del sistema.

Con la certificación propuesta se comprobará si el sistema de IA fomenta la igualdad en el acceso a la educación, a los bienes, a los servicios y/o a la tecnología.

Asimismo, revisará que los resultados obtenidos de la utilización del sistema de IA sean proporcionados y que los usuarios tengan la capacidad de oponerse a la mismos si no lo son.

Igualmente, se podría verificar si cumplen los requisitos de sostenibilidad y responsabilidad ecológica así como si presenta soluciones ante preocupaciones globales.

- El **principio de explicabilidad** ya que verificará si se informa de las capacidades y finalidades para las que se ha creado el sistema de IA y si es fácilmente identificable ante las personas como tal.

Igualmente, se examinará si es posible explicar como el sistema ha obtenido esos resultados y si documenta rigurosamente los algoritmos utilizados asegurando la trazabilidad que es lo que hará que se eviten las llamadas "cajas negras" y fomente la transparencia.

- Por último, el uso de certificaciones cumpliría con el requisito de **la rendición de cuentas** que garantizaría la responsabilidad proactiva (accountability) de los sistemas de IA y sus resultados tanto antes como después de la implantación del mismo al acreditar la diligencia de los desarrolladores y/o implantadores del sistema al haber realizado un análisis de riesgos y previendo las posibles contingencias.

Asimismo, al permitir la auditabilidad se evaluarían los algoritmos, los datos y los procesos de diseño del sistema que fomentarían la fiabilidad de la tecnología creada e incluso el respeto a los derechos fundamentales afectados.

Además, el uso de las certificaciones garantizaría la capacidad de informar sobre las acciones que concluyen los resultados y sus consecuencias así como las respuestas pudiendo minimizar los efectos negativos que puedan surgir al tener un plan de contingencia que reaccione ante esos adversos.

A este respecto y, tras nuestra justificación para determinar que la certificación sobre un sistema de IA es uno de los métodos más completos para asegurar que éste es fiable y, en el entendido de que la solidez técnica y la supervisión de la misma en los referidos sistemas también deben ser certificados, vamos a concretar nuestra propuesta en una certificación sobre los riesgos legales que conlleva el uso de un sistema de IA, en el que se verificaría que el objeto del sistema de IA, sus algoritmos, los datos y sus resultados tienen riesgos legales nulos. Entendemos, igualmente, que ayudaría a que los usuarios finales tuvieran confianza en los sistemas de IA certificados mejorando su comercialización haciéndola más segura y, en cierta manera, más multitudinaria.

En cuanto al objetivo perseguido, consideramos que los sistemas de IA se ven afectados por una serie de riesgos legales que es preciso mitigar y, que dependen del objetivo y finalidad para el que se va a diseñar. Así, los riesgos legales a los que nos enfrentamos y los que es preciso mitigar son los riesgos por responsabilidad civil, es decir, los referidos a responsabilidad civil por productos defectuosos, responsabilidad civil profesional, responsabilidad por privacidad, por vulneración de derechos fundamentales y sus posibles indemnizaciones; los riesgos legales por responsabilidad penal refiriéndonos a los posibles delitos por uso fraudulento del sistema, por daños y lesiones personales, por responsabilidad profesional y; los riesgos legales por lagunas, cambios o inexistencia de normativa, ya que, como se ha determinado, los sistemas de IA y su tecnología recen de manera exponencial y muchas veces van por delante de la regulación.

En base a estos riesgos legales y a los principios que debe cumplir todo sistema de IA (Respecto a la autonomía humana, Prevención de daño, Equidad y Explicabilidad) se ha elaborado un

cuestionario y recopilación de documentación, a modo de estándar fijado y personalizado que cumpla con las exigencias éticas y legales cuyo resultado es la emisión de una certificación que garantice que el sistema de IA es fiable en lo que respecta al ámbito legal.

El procedimiento propuesto para llevar a cabo la certificación se basaría en una auditoria tras la que el informe final sería la certificación del sistema de IA libre de riesgos legales. Para ello, nos basaríamos en un proceso de 5 pasos:

1.- Planificar: Realizar reuniones con los responsables del proyecto para determinar el objetivo del sistema de IA y concretar la auditoria a aplicar, informando a los responsables de la firma de actuar en la auditoria, la duración de la misma y la futura certificación.

2.- Hacer: Se llevará a cabo un cuestionario adaptado a cada uno de los sistemas que se vayan a verificar en cuanto a los riesgos legales en los que pueden incurrir dependiendo del objeto y finalidad para los que han sido creados.

3.- Verificar: Realizar el cuestionario personalizado y recopilación de la documentación necesaria que acredite lo respondido en el cuestionario.

4.- Revisar: Análisis del cuestionario y documentación recopilada. En el caso de que alguno de los puntos del cuestionario no sean los correctos de acuerdo a los estándares legales, se revisaran de nuevo para intentar solventarla incidencia. En este caso, se volvería al paso 3, tras haber realizado las gestiones y modificaciones oportunas y haber planteado, en su caso, medidas de seguridad adecuadas, se revisaría de nuevo el estándar fijado.

5.- Certificar: Una vez revisado, se emitiría un informe apto que conllevaría la emisión de un certificado o, un informe no apto, en el que se expondrían las razones y riesgos legales que se habrían mantenido a pesar de las modificaciones y por las que no se ha emitido el correspondiente certificado, proponiendo en su caso, las acciones correspondientes para conseguir el certificado..

Dicha certificación deberá ser realizada por empresas certificadoras distintas al desarrollador y al cliente y que deberán ser autorizadas y legitimadas por un organismo público.

Así expuesto, y centrándonos en la certificación de los sistemas de IA que creen las empresas, se propone que las certificaciones se realicen tanto en la fase de diseño del sistema, para que los mismos sean fiables cumpliendo con todos los principios expuestos desde esa fase, como en la de implementación (en el caso de que la tengan) ya que es posible que insertando el sistema de IA creado en un producto o servicio existente, pudiera resultar, en conjunto, un producto o servicio no deseable ni fiable.

En el mismo sentido, también deberá evaluarse cuando el sistema de IA se actualice, ya que pueden modificar aspectos del mismo que lo hagan no fiable de acuerdo a los estándares establecidos.

Asimismo, en el caso de que el sistema de IA no se actualizase antes de un año, habría que evaluarlo de manera proactiva , al menos en lo que se refiere al uso de la técnica y la seguridad de la misma, con controles internos (existiendo la posibilidad de que sea mediante autoevaluaciones) que deberían ser acreditadas mediante informes que la propia empresa deberá remitir al organismo público correspondiente.

Creemos que la utilización de certificaciones, a pesar de que suponga un gasto adicional a las inversiones realizadas por cada una de las empresas desarrolladoras y/o implementadoras, podría dar lugar a incentivos. Por un lado, de tipo monetario, accediendo o teniendo la posibilidad de acceder a nuevas ayudas a la inversión para el mantenimiento del sistema creado y certificado y/o para la creación de nuevos sistemas y, de tipo fiscal, obteniendo deducciones fiscales para la propia empresa (como las ya existentes en la legislación española, por ejemplo, el previsto en el artículo 35 de la Ley de Impuesto de Sociedades, para empresas cuya actividad es el I+D+i).

Otro de los posibles incentivos a las certificaciones, sería la de que, en caso de que todas las validaciones y revisiones hubiesen sido positivas y, en base al principio de responsabilidad pro activa (accountability), la empresa que la tuviera y hubiera tenido un incidente con reclamación de indemnizaciones y daños, pudiera ser exonerada o, al menos reducida su responsabilidad, al haber tenido más diligencia que la que hubiera tenido un buen padre de familia habiendo analizado su sistema desde todos los puntos de vista y tomando las precauciones debidas ante los incidentes que se hubiesen detectado en las evaluaciones, estando reconocido por terceros de reconocido prestigio.

La creación de esta certificación así como los organismos necesarios para ello, podrían ser los mismos, aumentando sus competencias y adaptando sus estructuras, que los creados en la directiva de ciberseguridad (Reglamento 2019/881 del Parlamento Europeo y de Consejo de 17 de abril) que crea el organismo de ENISA (la Agencia de la Unión Europea para la Ciberseguridad que en España, a nivel local, se fija el CNN-CERT, de acuerdo a los requisitos establecidos en el anexo del mismo reglamento) así como el certificado de ciberseguridad. De esta manera, se minimizarían los riesgos económicos que suponen la creación de un órgano nuevo de estas características.

Tras nuestra breve exposición acerca de la creación de una certificación para los sistemas de IA fiable, procedemos a explicar el proceso que se llevaría a cabo en la fase de verificación de nuestra auditoria y que sería la parte que, una vez revisada y aplicada en el sistema de IA podría determinar la emisión de su certificación.

En primer lugar, se debería describir de manera breve el objetivo y la finalidad del sistema de la IA, indicando el nombre del sistema y la población a la que va dirigido.

Tras la exposición de la descripción del sistema de IA se personalizaría el cuestionario (basado en el preparado por la Comisión Europea en su versión piloto, que se adjunta como Anexo I) y la recopilación de documentación en base a los siguientes puntos:

1.- Vulneración de PRINCIPIOS.

En este punto se analizaría y verificaría la posible vulneración y afectación de principios y derechos fundamentales mediante una evaluación de impacto.

2.- DATOS.

Basándonos en la privacidad desde el diseño (RGPD) se verificaría la forma de recoger los datos (fuentes privadas o públicas, consentimiento del titular, base legitimadora), si existe o no seudonimización, minimización y granularidad, la transparencia en las finalidades y tratamiento de los datos (información al titular), control de los datos por parte del titular, medidas de seguridad adecuadas por parte del responsable, tiempo por el que se almacenarán los datos, si

los datos están sesgados porque las muestras no son adecuadas para el objetivo del sistema de IA auditado y la capacidad de resiliencia a la hora de adaptarse a la nueva regulación si la hubiera.

3- ALGORITMOS.

Se verificaría la transparencia y capacidad de explicar el algoritmo para que no haya caja negra, si se utilizan datos sesgados o si son adecuados para el objetivo. Asimismo, se analizaría si el propio algoritmo es adecuado para la finalidad y si se ha elaborado bajo la premisa de que las personas son el centro, cumpliendo con las pautas matemáticas adecuadas y si es correcto, es decir, sin ser defectuoso desde el inicio. Se observaría si tiene las medidas de seguridad adecuadas para que no sea atacado cibernéticamente, ni para que sea usado de manera delictiva y, si el mismo tiene capacidad para adaptarse a los cambios normativos o a una nueva regulación.

4.- RESULTADOS.

En este punto, se verificaría si el control de los mismos lo tiene una persona, si puede ser explicado (eliminando la posibilidad de la opacidad), si la usabilidad de la misma puede ser decidida por una persona y qué medidas de seguridad tiene para garantizar que el resultado no es delictivo, no causa daños y cumple con la normativa existente a pesar de que el resto de elementos hayan sido los adecuados.

ANEXO I

En la siguiente lista de evaluación se adjunta la versión piloto que preparó la Comisión Europea a la que se han hecho algunas sugerencias(en cursiva) que nos parece apropiadas para que cumplan con los principios y requisitos de los sistemas de IA fiables.

LISTA DE EVALUACIÓN PARA UNA IA FIABLE (VERSIÓN PILOTO)

1. Acción y supervisión humanas

Derechos fundamentales:

- En aquellos casos de usos en los que puedan producirse efectos potencialmente negativos para los derechos fundamentales, ¿ha llevado usted a cabo una evaluación del impacto sobre los derechos fundamentales? ¿Ha identificado y documentado los posibles equilibrios entre los diferentes principios y derechos? *(¿Que derechos fundamentales se podrían ver afectados? Se debería presentar documentación.)*

- ¿Interactúa el sistema de IA con el proceso de adopción de decisiones por parte de usuarios finales humanos (por ejemplo, con las acciones recomendadas, las decisiones que es preciso adoptar o la presentación de opciones)

- ¿Existe en esos casos el riesgo de que el sistema de IA afecte a la autonomía humana al interferir con el proceso de adopción de decisiones del usuario final de forma imprevista?

- ¿Ha considerado usted si el sistema de IA debería informar a los usuarios de que una decisión, contenido, recomendación o resultado es fruto de una decisión algorítmica? *(Presentar la documentación de la trazabilidad y explicar como lo hace)*

- En el caso de que el sistema de IA cuente con un bot de charla o un sistema conversacional, ¿son los usuarios finales humanos conocedores de que están interactuando con un agente no humano?

Acción humana:

- En el caso de que el sistema de IA se implante en el proceso de trabajo, ¿ha tenido usted en cuenta la asignación de tareas entre el sistema de IA y los trabajadores humanos para garantizar interacciones adecuadas y una supervisión y control humanas apropiadas?

- ¿El sistema de IA mejora o aumenta las capacidades humanas?

- ¿Se han adoptado medidas para evitar que los procesos de trabajo confíen o dependan en exceso del sistema de IA? *(¿Cuales?)*

Supervisión humana:

- ¿Ha analizado cuál sería el nivel adecuado de control humano sobre el sistema de IA específico y para el caso de uso concreto de que se trate?

- ¿Puede describir el nivel de control o implicación humana, si procede? ¿Quién es la persona que ostenta el control del sistema y cuáles son los momentos o herramientas para la intervención humana?

- ¿Ha establecido mecanismos y adoptado medidas para garantizar la posibilidad de dicho control o supervisión humanos o para asegurar que las decisiones se tomen bajo la responsabilidad exclusiva de seres humanos?*(Cuáles?)*

- ¿Ha adoptado alguna medida para posibilitar la realización de auditorías y para solucionar cualquier problema relacionado con la gestión de la autonomía de la IA?*(Cuáles?)*

- En el caso de que exista un sistema de IA (o un caso de uso) autónomo o con capacidad de autoaprendizaje, ¿ha establecido mecanismos de control y supervisión más concretos?

- ¿Qué tipo de mecanismos de detección y respuesta ha establecido para evaluar si algo puede salir mal?

- ¿Se ha asegurado de disponer de un botón de desconexión o un procedimiento que permita abortar una operación en condiciones de seguridad en caso necesario? ¿Implica ese procedimiento que se aborta el proceso en su totalidad, en parte o la delegación del control a un ser humano?

2. Solidez técnica y seguridad

<u>Resistencia a los ataques y seguridad</u>:

- ¿Ha evaluado las posibles formas de ataque a las que puede ser vulnerable el sistema de IA?

- En particular, ¿ha analizado los diferentes tipos y naturalezas de las vulnerabilidades, como la contaminación de los datos, la infraestructura física o los ciberataques?

- ¿Ha adoptado medidas o sistemas para garantizar la integridad del sistema de IA y su capacidad para resistir posibles ataques? *(Relacionar cada una de las vulnerabilidades a dichas medidas*)

- ¿Ha evaluado el comportamiento de su sistema en situaciones o entornos imprevistos?*(En que escenarios y como se ha comportado? Tiene plan de contingencia en dichos casos?)*

- ¿Ha analizado si su sistema se puede utilizar (y, en caso afirmativo, en qué medida) para diferentes fines? Si es así, ¿ha adoptado medidas adecuadas para prevenir su uso con fines no deseados (como, por ejemplo, la no divulgación de la investigación o despliegue del sistema)? *(Cuáles son dichos fines y que medidas son las fijadas en ese caso? Dispone de plan de contingencia en dichos casos? Cuál es?)*

Plan de repliegue y seguridad general *(aportar el plan de contingencia que deberá contener todos estos puntos)*:

- ¿Se ha asegurado de que su sistema cuente con un plan de repliegue suficiente en el caso de que se enfrente a algún ataque malintencionado o a otro tipo de situación inesperada (por ejemplo, procedimientos técnicos de conmutación o formulación de preguntas a un ser humano antes de continuar)?

- ¿Ha analizado el nivel de riesgo que plantea el sistema de IA en el caso de uso concreto previsto?

- ¿Ha introducido algún proceso para medir y evaluar los riesgos y la seguridad?

- ¿Ha proporcionado la información necesaria en caso de que exista algún riesgo para la integridad física de las personas?

- ¿Ha estudiado la posibilidad de contratar una póliza de seguro para hacer frente a los posibles daños que provoque el sistema de IA?

- ¿Ha identificado los riesgos potenciales para la seguridad asociados a (otros) usos previsibles de la tecnología, incluidos los usos accidentales o malintencionados? ¿Existe algún plan para mitigar o gestionar esos riesgos?

- ¿Ha evaluado si es probable que el sistema de IA cause daños a los usuarios o a terceros? En caso afirmativo, ¿ha evaluado la probabilidad, el daño potencial, el público afectado y la gravedad de tales daños?

- Si existe el riesgo de que el sistema de IA ocasione daños, ¿ha tenido en cuenta las leyes de responsabilidad civil y de protección de los consumidores? ¿Cómo?

- ¿Ha analizado los efectos potenciales o el riesgo para la seguridad del medio ambiente o de la fauna?

- ¿Ha tenido en cuenta en su análisis de riesgos si los problemas de seguridad o de la red (por ejemplo, los peligros para la ciberseguridad) plantean riesgos para la seguridad o pueden causar daños debido a un comportamiento imprevisto del sistema de IA?

- ¿Ha estimado el efecto probable de un fallo de su sistema de IA que provoque que el sistema ofrezca resultados erróneos, quede fuera de servicio o proporcione resultados socialmente inaceptables (como, por ejemplo, prácticas discriminatorias)?

- ¿Ha definido umbrales y mecanismos de gestión para los escenarios anteriores a fin de activar planes alternativos o de repliegue?

- ¿Ha definido y ensayado planes de repliegue?

<u>Precisión</u>

-¿Ha evaluado qué nivel y definición de precisión se requerirá en el contexto del sistema de IA y para el caso de uso previsto?

- ¿Ha evaluado cómo se mide y garantiza la precisión?

- ¿Ha adoptado medidas para garantizar que los datos utilizados sean exhaustivos y estén

- actualizados?

- ¿Ha adoptado medidas para evaluar si es necesario disponer de datos adicionales, por ejemplo para mejorar la precisión o eliminar sesgos?

- ¿Ha evaluado los daños que se ocasionarían si el sistema de IA realizara predicciones incorrectas?*(Cómo y cuáles son?)*

- ¿Ha establecido algún mecanismo para medir si el sistema está realizando una cantidad inaceptable de predicciones erróneas?

- Si el sistema está realizando predicciones erróneas, ¿ha establecido una serie de pasos que permitan subsanar el problema?

Fiabilidad y reproducibilidad:

- ¿Ha diseñado una estrategia para supervisar y verificar que el sistema cumple los objetivos, el propósito y las aplicaciones previstas? *(Cómo y cuál es?)*

- ¿Ha comprobado si es necesario tener en cuenta algún contexto o condición particular para garantizar la reproducibilidad?

- ¿Ha introducido procesos o métodos de verificación para medir y garantizar los diferentes aspectos de la fiabilidad y la reproducibilidad?*(Cuáles?)*

- ¿Ha establecido algún proceso para describir las situaciones en las que un sistema de IA falla en determinados tipos de entornos?

- ¿Ha documentado y detallado claramente esos procesos para la verificación de la fiabilidad de los sistemas de IA? *(Aportarlo?)*

- ¿Ha establecido algún mecanismo o comunicación para garantizar a los usuarios (finales) que el sistema de IA es fiable?

3. Gestión de la privacidad y de los datos

Respeto de la privacidad y de la protección de datos:

- Dependiendo del caso de uso, ¿ha establecido un mecanismo que permita notificar los problemas relacionados con la privacidad o la protección de datos en los procesos de recopilación de datos de los sistemas de IA (tanto con fines de formación como de funcionamiento) y su tratamiento?

- ¿Ha evaluado el tipo y alcance de los datos incluidos en sus bases de datos (por ejemplo, si estas contienen datos de carácter personal)?

- ¿Ha analizado formas de desarrollar el sistema de IA o de formar el modelo en las que no sea necesario utilizar datos personales o potencialmente sensibles (o que utilicen la mínima cantidad posible de este tipo de datos)?

- ¿Ha introducido mecanismos de aviso y control sobre los datos personales en función del caso de uso (como, por ejemplo, el consentimiento válido y la posibilidad de revocar el uso de dichos datos, cuando proceda)?

- ¿Ha tomado medidas para mejorar la privacidad, por ejemplo a través de procesos como el encriptado, la anonimización y la agregación?

- En los casos en que exista una persona responsable de la privacidad de los datos, ¿la ha implicado desde una fase inicial del proceso?

Calidad e integridad de los datos:

- ¿Ha alineado su sistema con las normas potencialmente pertinentes (por ejemplo, ISO, IEEE) o ha adoptado protocolos generales para la gestión y gobernanza cotidianas de sus datos? *(Cuáles son? Aportación del certificado o en su caso del protocolo en cuestión)*

- ¿Ha establecido mecanismos de supervisión para la recopilación, almacenamiento, tratamiento y utilización de los datos?

- ¿Ha evaluado su grado de control sobre la calidad de las fuentes de datos externas utilizadas?

- ¿Ha instaurado procesos para garantizar la calidad y la integridad de sus datos? ¿Ha estudiado la posibilidad de introducir otros procesos? ¿Cómo está verificando que sus conjuntos de datos no son vulnerados ni objeto de ataques?

Acceso a los datos:

- ¿Qué protocolos, procesos y procedimientos se han seguido para gestionar y garantizar una adecuada gobernanza de los datos

- ¿Ha evaluado quién puede acceder a los datos de los usuarios y en qué circunstancias?

- ¿Se ha asegurado de que esas personas poseen la cualificación para acceder a los datos, que se les exige acceder a ellos y que cuentan con las competencias necesarias para comprender los detalles de la política de protección de datos?

- ¿Ha asegurado la existencia de un mecanismo de supervisión que permita registrar cuándo, dónde, cómo y quién accede a los datos, y con qué propósito?

4. Transparencia

Trazabilidad:

- ¿Ha adoptado medidas que puedan garantizar la trazabilidad? Esto puede conllevar la presentación de la documentación de:

- Los métodos utilizados para diseñar y desarrollar el sistema algorítmico:

* en el caso de un sistema de IA basado en reglas, se debería documentar el método de programación o la forma en que se creó el modelo;

* o en el caso de un sistema de IA basado en el aprendizaje, se debería documentar el método de formación del algoritmo, incluidos los datos de entrada que se recopilaron y seleccionaron y la forma en que se hizo;

- Los métodos empleados para ensayar y validar el sistema algorítmico:

* en el caso de un sistema de IA basado en reglas, se deberían documentar los escenarios o casos de uso utilizados para los ensayos y la validación;

* en el caso de un modelo basado en el aprendizaje, se debería documentar la información sobre los datos utilizados para los ensayos y la validación;

- Los resultados del sistema algorítmico:

* se deberían documentar los resultados del algoritmo o las decisiones adoptadas por este, así como otras posibles decisiones que se producirían en casos diferentes (por ejemplo, para otros subgrupos de usuarios).

Explicabilidad:

- ¿Ha evaluado en qué medida son comprensibles las decisiones y, por tanto, el resultado producido por el sistema de IA? *(Aportar la manera en la que se explican los resultados obtenidos (puede ser confidencial)*

- ¿Se ha asegurado de que se pueda elaborar una explicación comprensible para todos los usuarios que puedan desearla sobre las razones por las que un sistema adoptó una decisión determinada que diera lugar a un resultado específico?

- ¿Ha evaluado en qué medida la decisión del sistema influye en los procesos de adopción de decisiones de la organización?

- ¿Ha evaluado por qué se desplegó ese sistema en particular en esa área concreta?

- ¿Ha evaluado el modelo de negocio del sistema (por ejemplo, de qué modo crea valor para la organización)?

- ¿Ha diseñado el sistema de IA teniendo en mente desde el principio la interpretabilidad?

- ¿Ha investigado y tratado de utilizar el modelo más sencillo e interpretable posible para la aplicación en cuestión?

- ¿Ha evaluado si puede analizar sus datos relativos a la formación y los ensayos realizados? ¿Puede modificar y actualizar estos datos a lo largo del tiempo?

- ¿Ha evaluado si, tras la formación y el desarrollo del modelo, tiene alguna posibilidad de examinar su interpretabilidad o si dispone de acceso al flujo de trabajo interno del modelo?

Comunicación:

- ¿Ha informado a los usuarios (finales) —mediante cláusulas de exención de responsabilidad u otros medios— de que están interactuando con un sistema de IA y no con otro ser humano? ¿Ha etiquetado su sistema de IA como tal?

- ¿Ha establecido mecanismos para informar a los usuarios de las razones y criterios subyacentes a los resultados del sistema de IA?

- ¿Se han comunicado claramente estos a los usuarios previstos?

- ¿Ha establecido procesos que tengan en cuenta las opiniones de los usuarios y que utilicen dichas

- opiniones para adaptar el sistema?

- ¿Ha informado sobre los riesgos potenciales o percibidos, como la posible existencia de sesgos?

- ¿Ha tenido también en cuenta, según el caso de uso, la comunicación y la transparencia hacia otras audiencias, hacia terceros o hacia el público en general?

- ¿Ha dejado claro el propósito del sistema de IA y quién o qué podrá beneficiarse del producto o servicio que ofrezca este?

- ¿Se han especificado y se ha informado claramente sobre los escenarios de utilización del producto, estudiando posibles métodos de comunicación alternativos para garantizar que dicha información sea comprensible y adecuada para los usuarios a los que se dirige?

- Según el caso de uso, ¿ha tenido en cuenta la psicología humana y sus posibles limitaciones, como el riesgo de confusión, el sesgo de confirmación o la fatiga cognitiva?

- ¿Ha comunicado con claridad las características, limitaciones y posibles carencias del sistema de IA:

- en caso de desarrollo: a las personas encargadas de su despliegue en un producto o servicio?

- en caso de despliegue: a los usuarios finales o consumidores?

5. Diversidad, no discriminación y equidad

<u>Necesidad de evitar sesgos injustos:</u>

- ¿Se ha asegurado de que exista una estrategia o un conjunto de procedimientos para evitar crear o reforzar un sesgo injusto en el sistema de IA, tanto en relación con el uso de los datos de entrada como en lo referente al diseño del algoritmo?

- ¿Ha evaluado y reconocido las posibles limitaciones que emanan de la composición de los conjuntos de datos utilizados?

- ¿Ha tenido en cuenta la diversidad y representatividad de los usuarios en los datos? ¿Ha realizado ensayos para poblaciones específicas o casos de uso problemáticos?

- ¿Ha investigado y utilizado las herramientas técnicas disponibles para mejorar su comprensión de los datos, el modelo y su rendimiento?

- ¿Ha establecido procesos para verificar la existencia de posibles sesgos y llevar a cabo un seguimiento de estos durante las fases de desarrollo, despliegue y utilización del sistema?

- Dependiendo del caso de uso, ¿se ha asegurado de introducir un mecanismo que permita a otras personas informar sobre posibles problemas relacionados con la existencia de sesgos, discriminación o un rendimiento deficiente del sistema de IA? *(Cuál es y como funciona?)*

- ¿Ha estudiado vías y métodos de comunicación claros sobre cómo y a quién informar sobre este tipo de problemas?

- ¿Ha tenido en cuenta no solo a los usuarios (finales) sino también a otras personas que puedan verse indirectamente afectadas por el sistema de IA?*(cuáles son esas personas)*

- ¿Ha evaluado si existe la posibilidad de que las decisiones varíen aunque las condiciones no cambien?

- Si es así, ¿ha estudiado cuáles podrían ser las causas de ello?

- En caso de variabilidad, ¿ha establecido algún mecanismo de medición o evaluación del impacto potencial de dicha variabilidad sobre los derechos fundamentales ? *(Cuáles son?*

- ¿Se ha asegurado de utilizar una definición operativa adecuada de «equidad» para aplicarla en el diseño de sistemas de IA?

- ¿Se trata de una definición de uso común? ¿Estudió otras definiciones antes de optar por la seleccionada?

- ¿Ha instaurado análisis o parámetros cuantitativos para medir y poner a prueba la definición de equidad aplicada? *(Cuáles son?)*

- ¿Ha establecido mecanismos para garantizar la equidad en sus sistemas de IA? ¿Ha considerado otros posibles mecanismos?*(Cuáles y por qué?)*

<u>Accesibilidad y diseño universal:</u>

- ¿Se ha asegurado de que el sistema de IA se adapte a una amplia variedad de preferencias y capacidades individuales?

- ¿Ha evaluado si las personas con discapacidad, con necesidades especiales o en riesgo de exclusión pueden utilizar el sistema de IA? ¿Cómo se integró este aspecto en el sistema y cómo se verifica su funcionamiento?

- ¿Se ha asegurado de que la información sobre el sistema de IA también sea accesible para los usuarios de tecnologías asistenciales?*(Cómo lo hace?)*

- ¿Implicó o consultó a esta comunidad durante la fase de desarrollo del sistema de IA? *(Aportación de la comunicación)*

- ¿Ha tenido en cuenta el impacto de su sistema de IA en sus usuarios potenciales?

- ¿Es el equipo involucrado en el desarrollo del sistema de IA representativo de la audiencia a la que va dirigido? ¿Es representativo de la población en general y tiene también en cuenta a otros grupos que pudieran verse afectados de manera tangencial por el sistema?

- ¿Ha evaluado la posibilidad de que haya personas o grupos que puedan verse afectados de forma desproporcionada por las implicaciones negativas del sistema? *(Qué grupos son y por qué?)*

- ¿Ha recabado la opinión de otros equipos o grupos representativos de diferentes contextos y experiencias? *(Cuáles han sido y de qué manera las ha obtenido?)*

Participación de las partes interesadas:

- ¿Ha estudiado la posibilidad de introducir algún mecanismo para incorporar la participación de diferentes partes interesadas en el desarrollo y la utilización del sistema de IA?

- ¿Ha allanado el camino para la introducción del sistema de IA en su organización, informando e implicando previamente a los trabajadores afectados y sus representantes?

6. Bienestar social y ambiental

Una IA sostenible y respetuosa con el medio ambiente:

- ¿Ha establecido mecanismos para medir el impacto ambiental del desarrollo, despliegue y utilización del sistema de IA (por ejemplo, energía consumida por cada centro de datos, tipo de energía utilizada por los centros de datos, etc.)?

- ¿Se ha asegurado de introducir medidas para reducir el impacto ambiental de su sistema de IA a lo largo de todo su ciclo de vida?

Impacto social:

- En el caso de que el sistema de IA interactúe directamente con seres humanos:

- ¿Ha evaluado si el sistema de IA alienta a los humanos a establecer un vínculo y desarrollar la

- empatía con el sistema?

- ¿Se ha asegurado de que el sistema indique claramente que su interacción social es simulada y que no tiene capacidad para «entender» ni «sentir»?

- ¿Se ha asegurado de que se entiendan correctamente los efectos sociales del sistema de IA? Por ejemplo, ¿ha evaluado si existe un riesgo de pérdida de puestos de trabajo o de descualificación de la mano de obra? ¿Qué pasos se han dado para contrarrestar esos riesgos?

Sociedad y democracia:

- ¿Ha evaluado el impacto social global asociado al uso del sistema de IA más allá del que tenga sobre el usuario (final), como, por ejemplo, las partes interesadas que pueden verse indirectamente afectadas por dicho sistema?

7. Rendición de cuentas

Auditabilidad:

- ¿Ha establecido mecanismos para facilitar la auditabilidad del sistema por parte de agentes internos o independientes (garantizando, por ejemplo, la trazabilidad y registro de los procesos y resultados del sistema de IA)?

Minimización de efectos negativos y notificación de estos:

- ¿Ha llevado a cabo una evaluación de riesgos o de impacto del sistema de IA que tenga en cuenta a las diferentes partes interesadas que se vean afectadas por este de forma directa o indirecta? *(Aportarla)*

- ¿Ha establecido marcos de formación y educación para el desarrollo de prácticas de rendición de cuentas? *(Aportarlos)*

- ¿Qué trabajadores o partes del equipo están implicados en ello? ¿Trasciende la fase de desarrollo?

- ¿Se explica también en esa formación el posible marco jurídico aplicable al sistema de IA?

- ¿Ha considerado la posibilidad de crear una «junta de revisión ética de la IA» u otro mecanismo similar para debatir sobre las prácticas éticas y de rendición de cuentas en general, incluidas las posibles «zonas grises»?

- Además de las iniciativas o marcos internos para supervisar la ética y la rendición de cuentas, ¿se cuenta con algún tipo de orientación externa o se han establecido también procesos de auditoría? *(Cuáles son?)*

- ¿Existe algún proceso para que los trabajadores o agentes externos (por ejemplo, proveedores, consumidores, distribuidores/vendedores) informen sobre posibles vulnerabilidades, riesgos o sesgos en el sistema de IA o su aplicación? *(Cuáles son?)*

Documentación de los equilibrios alcanzados:

- ¿Se ha establecido algún mecanismo para identificar los intereses y valores que implica el sistema de IA y los posibles equilibrios entre ellos? *(Cuáles son?)*

- ¿Qué procesos ha seguido para decidir sobre los equilibrios necesarios? ¿Se ha asegurado de documentar la decisión sobre la búsqueda de dichos equilibrios?

Capacidad de obtener compensación:

- ¿Ha establecido un conjunto de mecanismos adecuado que permita obtener compensación en el caso de que se produzca cualquier daño o efecto adverso?*(Cuáles son?)*

- ¿Se han instaurado mecanismos para proporcionar información a usuarios (finales) y a terceros sobre las oportunidades de obtener compensación? *(Cuáles son?)*