I Encuentro FIADYS

Impacto de la IA en el uso de algoritmos en justicia y seguridad

Editado por FIADYS en 2024

#1

Impacto de la IA: los algoritmos en justicia y seguridad a la luz del Reglamento Europeo de Inteligencia Artificial (IA)

A continuación, se presenta un resumen de las ponencias que dieron comienzo el encuentro [1] y las conclusiones alcanzadas tras un interesante debate en torno a un problema todavía cargado de interrogantes.

Ponentes

- Fernando Miró Llinares. Catedrático de Derecho Penal y Criminología de la Universidad Miguel Hernández de Elche, Director del centro CRIMINA y patrono de la Fundación FIADYS.
- Patricia Faraldo Cabana. Catedrática de Derecho Penal, Universidade da Coruña, Directora de ECRIM.

1. Resumen de las ponencias

1.1. Finalidad y ámbito de aplicación del Reglamento de IA

El reglamento europeo sobre IA busca regular los usos de la IA para limitar los riesgos que de ellos se derivan. Su finalidad es garantizar que los sistemas de IA sean seguros y respeten los derechos fundamentales, así como brindar apoyo a la innovación en materia de IA.

Esta normativa ha elegido un enfoque basado en riesgos y los daños potenciales que la IA puede generar en la sociedad. Crea un sistema de gobernanza y supervisión europeo, con órganos de la Unión, pero también dependiente de autoridades de supervisión nacionales.

El ámbito de aplicación del reglamento se circunscribe a proveedores de sistemas de IA que pongan en servicio o comercialicen dentro de la UE o

---

[1] Este encuentro ha sido posible gracias a la Fundación FIADYS y a la colaboración de dos proyectos de investigación, el proyecto de I+D+i TED2021-129356B-I00, financiado por MICIU/AEI/10.13039/501100011033 y por la "Unión Europea NextGenerationEU/ PRTR"; y el proyecto "La responsabilidad de la inteligencia artificial: un desafío para las ciencias penales" (PID2020-112637RB-I00).

cuya salida se utilice en la UE, aunque su origen proceda de países externos, y a usuarios de los mismos o personas que explotan estos sistemas y no a los afectados. Quedan fuera del ámbito del reglamento la utilización de sistemas de IA por las autoridades de terceros países y organizaciones internacionales cuando algunos utilicen sistemas de uso militar o en el contexto de seguridad nacional, ni los utilizados con un fin de investigación o desarrollo científico.

1.2. ¿Cuándo será el Reglamento plenamente aplicable?

Tras su adopción por el Parlamento Europeo y el Consejo, el Reglamento de IA entrará en vigor a los veinte días de su publicación en el Diario Oficial, pero será aplicable de forma progresiva, por norma general 24 meses después de su entrada en vigor.

1.3. ¿Qué sistemas entiende el Reglamento que son IA?

Sistemas considerados IA

Un sistema basado en una máquina diseñado para funcionar con distintos niveles de autonomía, que puede mostrar capacidad de adaptación tras el despliegue y que, para objetivos explícitos o implícitos, infiere, a partir de la información de entrada que recibe, la manera de generar información de salida, como predicciones, contenidos, recomendaciones o decisiones, que puede influir en entornos físicos o virtuales (art. 3. 1).

Una característica principal de los sistemas de IA es su capacidad de inferencia (…). Las técnicas que permiten la inferencia al construir un sistema de IA incluyen estrategias de aprendizaje automático. Estas técnicas aprenden de los datos cómo alcanzar determinados objetivos y estrategias basadas en la lógica y el conocimiento que infieren a partir de información codificada o de una representación simbólica de la tarea que debe resolverse. La capacidad de inferencia de un sistema de IA trasciende el tratamiento o análisis básico de datos y permite el aprendizaje automático, el razonamiento o la modelización de los datos (considerando 12).

Sistemas no considerados IA

Según el Reglamento de IA la definición adoptada se ideó basándose "en las principales características de los sistemas de IA que los distinguen de los sistemas de software o los planteamientos de programación tradicionales y más sencillos" excluyéndose "los sistemas basados en las normas definidas únicamente por personas físicas para ejecutar automáticamente operaciones" (considerando 12).

1.3. ¿Qué sistemas entiende el Reglamento que son IA?

a) Sistemas predictivos o de policía predictiva

Son técnicas analíticas (preferentemente cuantitativas) que mediante predicciones estadísticas tratan de identificar posibles objetivos de intervención policial o prevenir el crimen o resolver delitos pasados. Por ejemplo, permiten predecir qué, cómo y quién será la próxima víctima o el autor de un delito.

Estos sistemas pueden ser de dos tipos: sistemas basados en el lugar ("Place Based Predictive Policing") o en las personas o sujetos ("Person Based Predictive Policing"). El Reglamento regula los sistemas de policía predictiva de forma amplia, incluyéndose aquellos sistemas utilizados para la ejecución de la ley penal, ya sea en el ámbito policial, judicial o penitenciario ("Law enforcement").

Respecto a los sistemas de policía predictiva ("predictive policing"), el reglamento incorpora prácticas prohibidas y sistemas considerados de alto riesgo.

Prácticas prohibidas

El uso de sistemas de IA para realizar evaluaciones de riesgo de personas físicas únicamente con el fin de evaluar o predecir la probabilidad de que una persona física cometa una infracción penal, basándose solamente en la elaboración del perfil de una persona física o en la evaluación de rasgos o características de su personalidad. La prohibición no se aplica a sistemas de IA para el apoyo de la evaluación humana de la implicación de una persona en un hecho delictivo que se base en datos objetivos y verificables directamente relacionados con la actividad delictiva.

Es decir, están prohibidos los sistemas de IA:

- Cuyo fin sea la valoración del riesgo de la comisión de una infracción penal por parte de una persona física.
- Que esa valoración se base únicamente en el "profiling".

¿Qué se entiende por "profiling"?
Toda forma de tratamiento automatizado de datos personales consistente en "utilizar datos personales para evaluar determinados aspectos personales de una persona física, en particular, para analizar o predecir aspectos relativos al rendimiento profesional, situación económica, salud, preferencias personales, intereses, fiabilidad, comportamiento, ubicación o movimientos de dicha persona física" (art. 4. 4 Reglamento General de Protección de Datos).

¿Qué se entiende por "basado únicamente en la elaboración de perfiles"?
Es un concepto aún no explorado en detalle, pero utilizando conceptos de la normativa de protección de datos podemos sugerir lo siguiente:

- Que la adscripción de una persona a un perfil o grupo sea la única circunstancia que se tenga en cuenta en la valoración del riesgo.
- Por tanto, la decisión final no ha sido tomada por un humano y es la consecuencia automática de la integración de una persona en un perfil. Consiguientemente, el responsable debe garantizar una supervisión significativa de la decisión por una persona autorizada y competente que pueda modificar esa decisión (similitud a la normativa de protección de datos).

Sistemas de IA de alto riesgo

No se encuentran prohibidos, pero su uso se supedita a introducir una serie de garantías. En el ámbito del "predictive policing" son sistemas de alto riesgo los sistemas utilizados para evaluar el riesgo de la comisión de una infracción penal de una persona física que estén basados exclusivamente en el "profiling".

Sistemas de IA excluidos

Quedan fuera las herramientas que no sean consideradas IA bajo la definición general del Reglamento y específicamente los sistemas predictivos basados en el lugar pues no realizan una evaluación del riesgo de personas físicas individualizadas.

b) Identificación biométrica

Prácticas prohibidas o de riesgo inaceptable:

- Sistemas de IA que creen o amplíen bases de datos de reconocimiento facial mediante la extracción no selectiva de imágenes faciales de internet o de circuitos cerrados de televisión.
- Sistemas de identificación biométrica remota "en tiempo real" en espacios de acceso público con fines de aplicación de la ley.

Dentro de esta prohibición, hay excepciones:

Casos permitidos de identificación biométrica remota en tiempo real en espacios públicos con fines de aplicación de la ley (que exigen autorización judicial, evaluación previa de las repercusiones y notificación a la Agencia Española de Protección de Datos (AEPD):

- Localización o identificación de una persona sospechosa de haber cometido alguno de los delitos mencionados en el anexo II[2] que en el Estado miembro de que se trate se castigue con una pena o una medida de seguridad privativas de libertad cuya duración máxima sea de al menos cuatro años.
- Búsqueda de víctimas concretas de secuestro, trata de seres humanos o explotación sexual de seres humanos, así como de personas desaparecidas.
- Prevención de una amenaza específica, importante e inminente para la vida o la seguridad física de las personas físicas o de una amenaza real y actual o real y previsible de un atentado terrorista.

---

[2] Terrorismo, trata de seres humanos, explotación sexual de menores y pornografía infantil, tráfico ilícito de estupefacientes o sustancias psicotrópicas, tráfico ilícito de armas, municiones y explosivos, homicidio voluntario, agresión con lesiones graves, tráfico ilícito de órganos o tejidos humanos, tráfico ilícito de materiales nucleares o radiactivos, secuestro, detención ilegal o toma de rehenes, delitos que son competencia de la Corte Penal Internacional, secuestro de aeronaves o buques, violación, delitos contra el medio ambiente, robo organizado o a mano armada, sabotaje, participación en una organización delictiva implicada en uno o varios de los delitos enumerados en esta lista.

Prácticas de alto riesgo

Biometría, en la medida en que su uso esté permitido por el Derecho de la Unión o nacional aplicable:

a) Sistemas de identificación biométrica remota

Quedan excluidos de la consideración de sistemas de alto riesgo los sistemas de IA destinados a ser utilizados con fines de verificación biométrica. A diferencia de la identificación biométrica, que compara el patrón biométrico de una persona con todos los que están recogidos en una base de datos (búsqueda de correspondencia uno-a-muchos), para saber quién es una persona, la verificación compara el patrón biométrico de una persona con otro ya registrado, para verificar que es quién dice ser (proceso de búsqueda de correspondencia uno-a-uno).

La posibilidad de utilizar estos sistemas de alto riesgo que cuenten con una autorización legal está sometida al cumplimiento de múltiples requisitos: a) sistema de gestión de riesgos, b) gobernanza y gestión de los datos de entrenamiento y prueba, c) documentación técnica actualizada, d) registros de actividad del sistema, e) información a los usuarios sobre las capacidades del sistema, f) supervisión humana, g) nivel adecuado de precisión, robustez y ciberseguridad. Además, los organismos de Derecho público o agentes privados que presten servicios públicos y operadores que suministren sistemas de alto riesgo deben llevar a cabo una evaluación de impacto.

2. Conclusiones del debate

¿Son las herramientas como el VioGén o RisCanvi sistemas de IA?

Es un debate abierto pero la opinión mayoritaria se mostró contraria a su reconocimiento como IA por los siguientes motivos:

- La decisión final sobre la valoración del riesgo es supervisada o adoptada por un profesional en base a la información analizada siguiendo el protocolo.
- No hay aprendizaje automático ni una adaptación del sistema en base a los resultados de ese aprendizaje.
- El RisCanvi fue elaborado a partir de la recolección inicial de casos reales en prisión y mediante unas técnicas regresión múltiple se identificaron una serie de factores de riesgo que permitían clasificar a los sujetos en función del riesgo a partir de un punto de corte que fue establecido por parte de los profesionales que generaron la herramienta. Estos puntos de corte no se modifican con el funcionamiento del sistema.
- Este proceso es similar a herramientas como el VioGén y PREVI-A.
- No extraen la información de forma automática, sino que son los profesionales, mediante su actividad, los que alimentan el sistema.

Aun así, se consideró que en el caso de que se entendiese que son sistemas de IA, estas herramientas podrían ser consideradas de alto riesgo y, por consiguiente, será necesario atenerse a una serie de requisitos y controles tal y como establece el Reglamento.

¿Pueden mejorarse estas herramientas con la incorporación de la IA?

- Los sistemas de valoración de riesgo existentes no son sensibles a fenómenos infrecuentes de la conducta humana como algunos comportamientos violentos excepcionales. La IA podría ayudar a mejorar la capacidad predictiva de estos sistemas ya que se alimentan de todos los casos nuevos y se calibra de forma automática en base a la información introducida.

- Esto no quiere decir que la introducción de la información y el resultado que se obtiene de estas herramientas no tenga que ser supervisado por los profesionales.
- Si son considerados IA podrían ser consideradas de alto riesgo y van a ser necesarios una serie de requisitos y controles. Sin embargo, se considera que es necesario aprovechar esta situación para mejorar estos instrumentos con la introducción de IA.
- Uno de los problemas de este tipo de sistemas es que existe un sesgo tecnológico en favor de los resultados automáticos de estos programas considerándolos de mayor valor que los de los profesionales. Ante ello, hay que generar precauciones teniendo en cuenta el juicio profesional.
- La IA no va a sustituir la comprensión de los fenómenos porque no permite rastrear e identificar la información en la que se basa para llegar a estas predicciones. Por lo tanto, por el momento, será necesario seguir contando con investigación alternativa para la comprensión profunda de los resultados.

Respecto a los sistemas de reconocimiento biométrico, se abordó la siguiente cuestión:

¿Qué tasa de error de los reconocimientos biométricos priorizamos? ¿Qué pasa con los falsos reconocimientos?

El reconocimiento facial automático es una tecnología inherentemente probabilística y, por tanto, intrínsecamente falible. Una vez adquiridos, los datos biométricos en bruto de un individuo se evalúan y, cuando es necesario, se someten a algoritmos de mejora de la señal para aumentar su calidad. La muestra biométrica inicial se transforma en una plantilla digital que contiene solo la información necesaria para ejecutar el algoritmo de reconocimiento de patrones. En la fase de comparación, la plantilla se compara con otra plantilla registrada en el sistema para producir una ratio de probabilidad o ratio de coincidencia basada en la puntuación, según la cual se valida o rechaza la identificación de una persona o la verificación de su identidad.

En esta fase, la tasa de falsas no coincidencias (FNMR) y la tasa de falsas coincidencias (FMR) son funciones del umbral del sistema: si el diseñador disminuye el umbral de aceptación para que el sistema sea más tolerante a las variaciones de entrada, la FMR aumenta, mientras que, si el umbral de aceptación se eleva para que el sistema sea más seguro, la FNMR aumenta en consecuencia. En resumen, los diseñadores pueden fijar el valor del umbral de aceptación a voluntad. Para ello, sin embargo, es importante tener en cuenta la finalidad y el contexto de uso del reconocimiento facial automático: cuando se aplica la tecnología en lugares visitados por millones de personas -como aeropuertos o estaciones de tren-, una proporción relativamente pequeña de errores sigue suponiendo que cientos de individuos sean marcados erróneamente, es decir, que sean identificados o rechazados incorrectamente como coincidentes. Las consecuencias de estos dos errores son diferentes según la situación. Por ejemplo, si la policía utiliza un algoritmo de reconocimiento facial en sus esfuerzos por localizar a un fugitivo, un falso positivo puede llevar a la detención errónea de una persona inocente, mientras que un falso negativo puede ayudar a que el sospechoso escape. Cada caso requiere una determinación del coste de los diferentes tipos de errores y una decisión sobre qué tipo de errores priorizar, teniendo en cuenta que la naturaleza probabilística de los resultados y la incorporación de determinados valores en la herramienta plantean dudas. sobre la justificación de considerar los resultados obtenidos como indicios "objetivos" que apunten a una sospecha razonable. Es exigible que los ajustes del diseño, como el mencionado umbral de aceptación, reflejen la arquitectura institucional del proceso penal, incluidos sus objetivos primordiales, es decir, absolver al inocente, condenar al culpable y la tasa aceptable de errores entre estos objetivos. La famosa relación de Blackstone (1769 [1893], p. 358) que subraya el "valor fundamental [...] de nuestra sociedad de que es mucho peor condenar a un inocente que dejar libre a un culpable", ilustra perfectamente este punto.

Finalmente se consideró que la finalidad del uso de este tipo de herramientas debería ser su utilización como guía de actuación, no como una solución independiente al criterio humano y experto.

FIADYS

secretaria@fiadys.org
fiadys.org