Automatizando (In)Justicias:

Una auditoría adversarial a RisCanvi

La justicia penal ha experimentado una profunda transformación a raíz de la integración de algoritmos predictivos, como lo demuestran sistemas como COMPAS o PredPol, ambos utilizados en Estados Unidos. Este cambio de paradigma ha suscitado debates en torno a la **transparencia, los sesgos y la fiabilidad de la toma de decisiones algorítmicas.**

En Eticas hemos querido contribuir al debate realizando **la primera auditoría adversarial de un sistema de justicia penal basado en IA en Europa**: la herramienta RisCanvi. Utilizada desde 2009 en Cataluña, España, y diseñada específicamente para evaluar la reincidencia y el riesgo de violencia entre los reclusos. RisCanvi ejerce influencia sobre las decisiones de libertad condicional y condena, y aunque algunos autores han expresado su preocupación por su parcialidad y falta de precisión y transparencia, la mayoría de las personas, reclusos, abogados y actores judiciales desconocen su existencia o su funcionamiento interno.

En nuestra misión por comprender RisCanvi, el equipo de Eticas llevó a cabo una Auditoría Adversarial utilizando un enfoque de doble métodología. Esto implicó una Auditoría Etnográfica, que incluyó entrevistas con reclusos y personal tanto dentro como fuera del sistema de justicia penal, y una Auditoría Comparativa de Resultados, basada en datos públicos sobre la población reclusa y la reincidencia, y comparándolos con los factores de riesgo y los comportamientos de riesgo de RisCanvi.

Lo que hemos encontrado en **RisCanvi es un sistema que no es conocido por aquellos a los que más impacta, los reclusos; que no es de confianza para muchos de los que trabajan con él, que tampoco están formados en su funcionamiento y ponderaciones; que es opaco y que no se ha adherido a la normativa vigente sobre el uso de sistemas automatizados de toma de decisiones en España**, donde las auditorías de IA son obligatorias desde 2016. Pero, sobre todo, nuestros datos demuestran que RisCanvi puede no ser justo ni fiable, y que no ha conseguido hacer lo que la IA debería hacer mejor: estandarizar los resultados y limitar la capacidad discrecional. En consonancia con estudios anteriores, no consideramos que RisCanvi sea fiable, ya que esto requeriría una relación clara entre los factores de riesgo, los comportamientos de riesgo y las puntuaciones de riesgo.

Basándonos en los datos disponibles, llegamos a la conclusión de que **RisCanvi no cumple su función, y actualmente no es capaz de ofrecer las garantías necesarias a reclusos, abogados, jueces y autoridades de justicia penal.**

Como ocurre con cualquier auditoría adversarial, nuestras conclusiones no son definitivas. No pudimos acceder a los datos del sistema, por lo que no pudimos corroborar nuestras observaciones. Pero existen al parecer suficientes datos sobre la mesa para **justificar un nuevo análisis del sistema**. En la situación actual, cuando un recluso de bajo riesgo reincide,

es imposible saber si el hecho de que no se le clasifique correctamente es el resultado de un porcentaje de error inevitable o una característica de un sistema poco fiable. Del mismo modo, cuando a un recluso se le deniega el acceso a mayores niveles de libertad debido a su alto riesgo, actualmente no está claro si se trata de una decisión justa.

Tal y como establece el AI Act recientemente aprobado, el uso de la IA en entornos especialmente vulnerables, como el sistema de justicia penal, requiere un mayor nivel de transparencia y supervisión, tanto interna como externa, así como esfuerzos constantes para inspeccionar y supervisar el rendimiento y el impacto del sistema. No encontramos que este sea el caso en el desarrollo y uso de RisCanvi, añadiendo el cumplimiento legal a los muchos retos de esta herramienta de riesgo de IA.

***

Este informe es el cuarto del Programa de Auditoría Adversarial de Eticas. Le animamos a acceder a informes anteriores sobre sistemas de IA que cubren temas como la evaluación del riesgo de sesgo de género, las desigualdades inducidas por la IA y las vulneraciones laborales en las aplicaciones de transporte VTC, y el impacto del reconocimiento facial en las personas con discapacidad. Esté atento a los futuros trabajos de Eticas sobre ceguera y reconocimiento de emociones, redes sociales y salud mental y el uso de la IA en la vivienda, la banca y los servicios sociales.

Conclusiones

RisCanvi es un algoritmo predictivo utilizado para evaluar la reincidencia y el riesgo de violencia y determinar el acceso a la libertad condicional en el sistema penitenciario catalán (España). Creado en 2009, en 2022 afectaba ya a 7.713 personas en el sistema penitenciario español, según los últimos datos disponibles del Ministerio de Justicia. **El sistema nunca ha sido auditado**, a pesar de que la normativa española exige que todos los sistemas automatizados que afectan a los derechos individuales sean auditados desde 2016.

El grado de conocimiento y comprensión del sistema RisCanvi **varía considerablemente entre los distintos actores**. Mientras que algunos profesionales están muy familiarizados con él, otros admiten conocerlo poco o nada. Esto es especialmente preocupante, ya que la mayoría de los que desconocen la existencia o el funcionamiento de RisCanvi son reclusos.

El funcionamiento interno de RisCanvi sigue siendo opaco, incluso para el personal de primera línea, como los psicólogos. Los usuarios del sistema desconocen las variables y ponderaciones en que se basan las calificaciones de riesgo.

Los profesionales pueden influir en las puntuaciones de RisCanvi, **pero rara vez lo hacen**, produciéndose alteraciones en menos del 5% de los casos. Esto significa que, de hecho, RisCanvi es un sistema de IA totalmente automatizado sin intervención humana significativa («human in the loop»).

Los reclusos carecen de apoyo jurídico y de conocimiento del sistema y de sus propias clasificaciones de riesgo, lo que impide su participación efectiva o la impugnación de los resultados de RisCanvi.

No existe una relación sólida entre los factores de riesgo de RisCanvi, los comportamientos de riesgo y las puntuaciones de riesgo. Los resultados de nuestros esfuerzos de ingeniería inversa muestran **correlaciones aleatorias entre los factores relevantes**. Esto apunta a una **falta de estandarización** en la asignación del riesgo a los reclusos.

Los factores estáticos inmutables, como los problemas de la infancia, desempeñan un papel crucial en las evaluaciones de la conducta, lo que provoca **sesgos contra determinados grupos demográficos** y reclusos cuya infancia transcurrió en entornos difíciles.

La estructura actual de **RisCanvi no ofrece las garantías necesarias de imparcialidad**, claridad y solidez a reclusos, abogados, jueces y autoridades de justicia penal.

etic as FOUNDATION
Irídia_
DIVERSIFAIR Diversify with Intersectionally Fairer AI
Funded by the European Union