ESTUDIOS

LA PREDICCIÓN DEL RIESGO Y EL RIESGO DE LA PREDICCIÓN EN LOS SISTEMAS PENITENCIARIOS

JUAN ALDAZ ARREGUI
AITOR MARTÍNEZ GONZÁLEZ
DIRECTORES

ARANZADI

LA PREDICCIÓN DEL RIESGO Y EL RIESGO DE LA PREDICCIÓN EN LOS SISTEMAS PENITENCIARIOS

JUAN ALDAZ ARREGUI
AITOR MARTÍNEZ GONZÁLEZ
Directores

La predicción del riesgo y el riesgo de la predicción en los sistemas penitenciarios

Prólogo
José Luis de la Cuesta Arzamendi

Autores

Juan Aldaz Arregui
Lorena Alemán Aróstegui
Antonio Andrés-Pueyo
José Luis de la Cuesta Arzamendi
Izaskun Ibabe Erostarbe
Iulen Ibáñez Baños
Lucía Martínez Garay
Aitor Martínez González
Fernando Miró Llinares
Marian Moreno Mata
Sandra Pérez Domínguez
Ana I. Pérez Machío
Alberto Saiz Garitaonandia
Ane Viana Salinas y Leyre Padilla López
Gema Varona Martínez

ARANZADI

© **Juan Aldaz Arregui y Aitor Martínez González (Dirs.) y autores, 2026**
© **ARANZADI LA LEY, S.A.U.**

ARANZADI LA LEY, S.A.U.
C/ Collado Mediano, 9
28231 Las Rozas (Madrid)
www.aranzadilaley.es
Atención al cliente: https://areacliente.aranzadilaley.es/publicaciones

Primera edición: marzo 2026

Depósito Legal: M-3525-2026
ISBN versión impresa: 978-84-1085-691-2
ISBN versión electrónica: 978-84-1085-692-9

Diseño, Preimpresión e Impresión: ARANZADI LA LEY, S.A.U.
Printed in Spain

© ARANZADI LA LEY, S.A.U. Todos los derechos reservados. A los efectos del art. 32 del Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba la Ley de Propiedad Intelectual, ARANZADI LA LEY, S.A.U., se opone expresamente a cualquier utilización del contenido de esta publicación sin su expresa autorización, lo cual incluye especialmente cualquier reproducción, modificación, registro, copia, explotación, distribución, comunicación, transmisión, envío, reutilización, publicación, tratamiento o cualquier otra utilización total o parcial en cualquier modo, medio o formato de esta publicación.
Cualquier forma de reproducción, distribución, comunicación pública o transformación de esta obra solo puede ser realizada con la autorización de sus titulares, salvo excepción prevista por la Ley. Diríjase a **Cedro** (Centro Español de Derechos Reprográficos, **www.cedro.org**) si necesita fotocopiar o escanear algún fragmento de esta obra.
El editor y los autores no asumirán ningún tipo de responsabilidad que pueda derivarse frente a terceros como consecuencia de la utilización total o parcial de cualquier modo y en cualquier medio o formato de esta publicación (reproducción, modificación, registro, copia, explotación, distribución, comunicación pública, transformación, publicación, reutilización, etc.) que no haya sido expresa y previamente autorizada.
El editor y los autores no aceptarán responsabilidades por las posibles consecuencias ocasionadas a las personas naturales o jurídicas que actúen o dejen de actuar como resultado de alguna información contenida en esta publicación.
ARANZADI LA LEY no será responsable de las opiniones vertidas por los autores de los contenidos, así como en foros, chats, o cualesquiera otras herramientas de participación. Igualmente, ARANZADI LA LEY se exime de las posibles vulneraciones de derechos de propiedad intelectual y que sean imputables a dichos autores.
ARANZADI LA LEY queda eximida de cualquier responsabilidad por los daños y perjuicios de toda naturaleza que puedan deberse a la falta de veracidad, exactitud, exhaustividad y/o actualidad de los contenidos transmitidos, difundidos, almacenados, puestos a disposición o recibidos, obtenidos o a los que se haya accedido a través de sus PRODUCTOS. Ni tampoco por los Contenidos prestados u ofertados por terceras personas o entidades.
ARANZADI LA LEY se reserva el derecho de eliminación de aquellos contenidos que resulten inveraces, inexactos y contrarios a la ley, la moral, el orden público y las buenas costumbres.
Nota de la Editorial: El texto de las resoluciones judiciales contenido en las publicaciones y productos de **ARANZADI LA LEY, S.A.U.**, es suministrado por el Centro de Documentación Judicial del Consejo General del Poder Judicial (Cendoj), excepto aquellas que puntualmente nos han sido proporcionadas por parte de los gabinetes de comunicación de los órganos judiciales colegiados. El Cendoj es el único organismo legalmente facultado para la recopilación de dichas resoluciones. El tratamiento de los datos de carácter personal contenidos en dichas resoluciones es realizado directamente por el citado organismo, desde julio de 2003, con sus propios criterios en cumplimiento de la normativa vigente sobre el particular, siendo por tanto de su exclusiva responsabilidad cualquier error o incidencia en esta materia.

Consejo Asesor Aranzadi LA LEY

D. Luis María Cazorla (Presidente)
Presidente de la Real Academia de Jurisprudencia y Legislación de España

D. Alberto Palomar (Secretario)
Magistrado de lo contencioso-administrativo (EV)

D. Ricardo Alonso
Catedrático de Derecho Administrativo y de la UE

D. Moisés Barrio
Letrado del Consejo de Estado

D. Jacobo Manuel Barja de Quiroga López
Presidente de la Sala de lo Militar del Tribunal Supremo

D. Alfredo Berges
Presidente de UNE

D.ª Sonia Calaza
Decana de la Facultad de Derecho de la UNED
Catedrática de Derecho procesal de la UNED

D.ª Concepción Campos
Experta en Gestión Pública y Presidenta de la Asociación de Mujeres en el Sector Público

D.ª Ana Belén Campuzano
Catedrática de Derecho mercantil de la Universidad CEU San Pablo

D. Adolfo Díaz Ambrona
Secretario General de la Cámara de Comercio de España

D. Antonio Fernández de Buján y Fernández
Académico de Número de las Reales Academias de Jurisprudencia y Legislación de España y de Galicia

D.ª Isabel Fernández Torres
Catedrática (Ac.) de Derecho Mercantil

D.ª Ana Fernández-Tresguerres
Notaria de Madrid

D. José Luis García Delgado
Catedrático de Economía Aplicada. Univ. Nebrija

D.ª Piedad García-Escudero
Letrada de las Cortes Generales

D. Rafael García Meiro
Consejero Delegado/CEO de AENOR

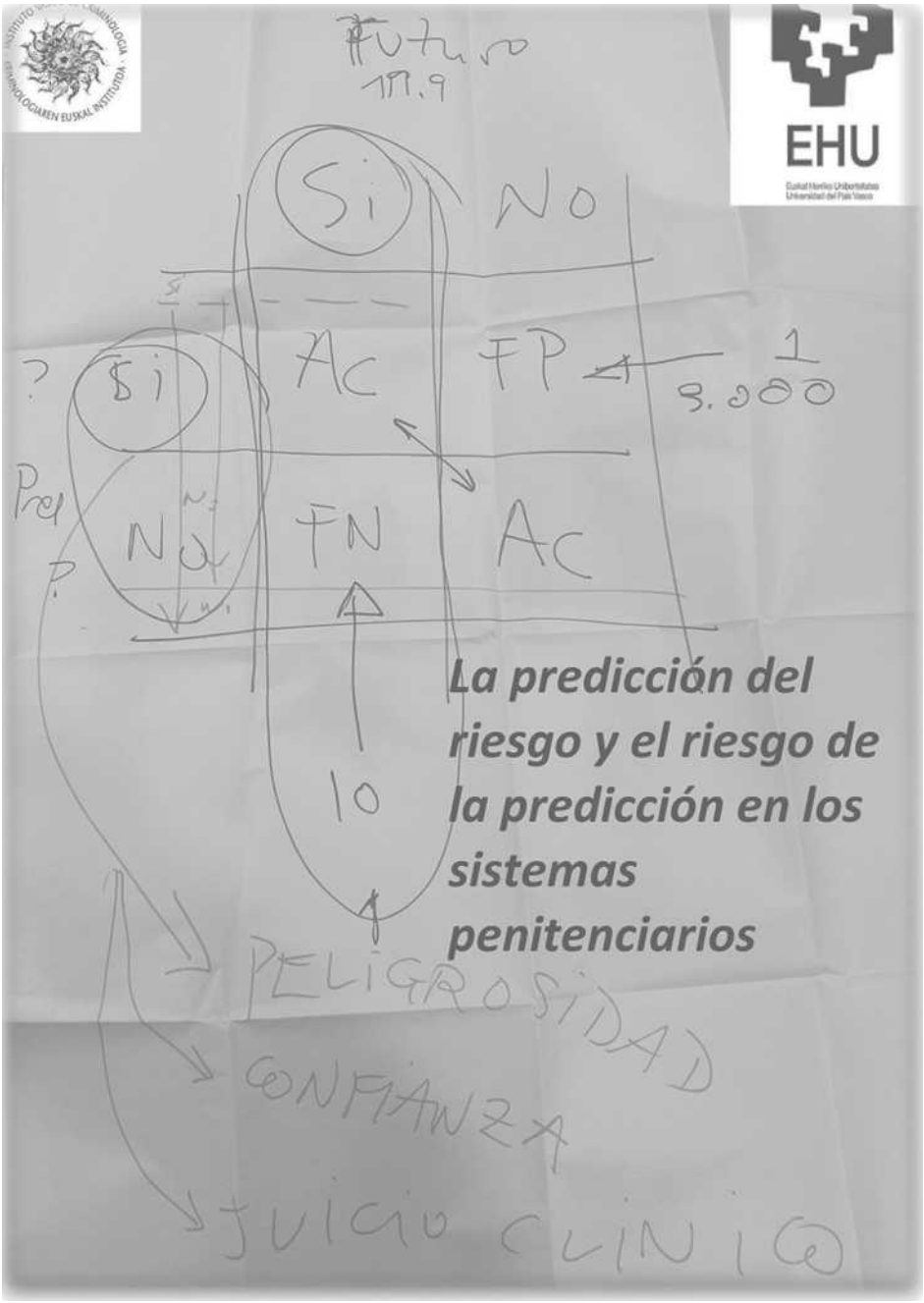
D. Isaac Merino Jara
Magistrado del Tribunal Supremo (Sala de lo Contencioso-Administrativo)

D.ª Encarnación Roca
Catedrática de Derecho Civil

D. Antonio V. Sempere Navarro
Magistrado del Tribunal Supremo

D.ª Rosario Silva
Abogada del Estado

D. Eduardo Torres-Dulce
Fiscal



La predicción del riesgo y el riesgo de la predicción en los sistemas penitenciarios

Juan Aldaz Arregui y Aitor Martínez González (Dirs.)

Índice General

Página

PRÓLOGO

JOSÉ LUIS DE LA CUESTA ARZAMENDI 23

PARTE I: ANÁLISIS DESCRITIVO

CAPÍTULO 1

VALORACIÓN DEL RIESGO E IA EN EL ÁMBITO PENITENCIARIO EN LA ERA DE LAS *SMART PRISONS*: ANDAR ANTES DE CORRER

PROF. DR. FERNANDO MIRÓ LLINARES Y SANDRA PÉREZ DOMÍNGUEZ 27

1. Lo inteligente (y lo no tanto) del término "smart prisons" .. 27

2. El contexto: Digitalización, algoritmización e IA en prisiones 30

3. La valoración del riesgo en el ámbito penitenciario en tiempos de IA. 38

4. Andar (en la dirección adecuada) antes que correr o quedarse inmóvil 43

5. Referencias 46

CAPÍTULO 2

ALGORITMOS DE REINCIDENCIA: ¿HERRAMIENTA ÚTIL PARA EUSKADI? REVISIÓN SISTEMÁTICA SOBRE VARIABLES, MODELOS E INTERPRETACIÓN EN EL ÁMBITO PENITENCIARIO

ANE VIANA SALINAS Y LEYRE PADILLA LÓPEZ 51

1. Introducción: contexto y objetivos 51

Página

2. Análisis de la literatura sobre algoritmos de reincidencia ... 53

3. Evaluación crítica de modelos, variables y resultados 58

4. Discusión: Estado de la cuestión actual e implicaciones para Euskadi 63

5. Conclusiones 67

6. Referencias 70

CAPÍTULO 3

LA VALORACIÓN PREDICTIVA DEL RIESGO EN LA EJECUCIÓN PENAL: LÍMITES, GARANTÍAS Y PROPUESTAS PARA EL ÁMBITO PENITENCIARIO VASCO

IZASKUN IBABE 77

1. Introducción 77

2. Modelos de herramientas predictivas de valoración del riesgo 79

2.1. El modelo Riesgo-Necesidad-Responsividad como marco teórico 79

2.2. Tipologías de instrumentos de valoración del riesgo 80

2.2.1. Juicio clínico no estructurado 80

2.2.2. Juicio clínico estructurado 81

2.2.3. Métodos actuariales 82

2.2.4. Modelos mixtos básicos y avanzados 82

2.3. Límites predictivos, validez externa y garantías jurídicas ... 84

3. Uso de herramientas predictivas en el ámbito español 88

4. Multi-escala RisCanvi 90

5. Propuesta de marco prospectivo para la eventual incorporación de una herramienta en el sistema penitenciario vasco . 93

6. Conclusiones: bases para un modelo avanzado y garantista 97

7. Referencias 99

Página

CAPÍTULO 4

COMPARATIVA EUROPEA DE HERRAMIENTAS ALGORÍTMICAS E INTELIGENCIA ARTIFICIAL PARA PREDECIR LA REINCIDENCIA DE LA POBLACIÓN PENITENCIARIA Y RECOMENDACIONES PARA SU POTENCIAL IMPLEMENTACIÓN EN LA CAPV

AITOR MARTÍNEZ GONZÁLEZ 103

1. Introducción 103

2. Herramientas algorítmicas predictoras de reincidencia en la UE 105

2.1. Finlandia 106

2.2. Países Bajos 107

2.3. Estonia 110

2.4. Grecia, Chipre, Portugal y Bulgaria 111

2.5. Suiza 113

3. Consideraciones éticas y legales a tener en cuenta para su implementación en la CAPV 114

4. Conclusiones 117

5. Referencias 119

PARTE II: RisCanvi

CAPÍTULO 5

EL PROTOCOLO RISCANVI PARA LA GESTIÓN DEL RIESGO EN EL SISTEMA PENITENCIARIO DE CATALUÑA

DR. A. ANDRES-PUEYO 125

1. Introducción 125

2. RisCanvi en el modelo rehabilitador de las prisiones en Cataluña 128

3. Aplicaciones del RisCanvi en la gestión penitenciaria 133

4. Conclusiones 138

5. Referencias 140

Página

CAPÍTULO 6

EVALUACIÓN TÉCNICA Y ÉTICA DEL SISTEMA RISCANVI: POTENCIALIDADES Y RIESGOS EN LA PREDICCIÓN ALGORÍTMICA PENITENCIARIA

IULEN IBÁÑEZ BAÑOS 145

1. Introducción 145
2. Presupuestos técnicos y de datos 146
3. Potencialidades de evolución 147
4. Discusión 147
5. Minimización de riesgos 148
6. Conclusiones 148
7. Referencias 149

CAPÍTULO 7

LA ARQUITECTURA DE LA TOMA DE DECISIONES EN EQUIPO EN EL ÁMBITO PENITENCIARIO: RETOS ÉTICO-JURÍDICOS DE LAS HERRAMIENTAS PREDICTIVAS DE VIOLENCIA Y DE QUEBRANTAMIENTO DE CONDENA

GEMA VARONA MARTÍNEZ 151

1. Introducción: un acercamiento ético-jurídico y criminológico a la adaptación de la herramienta RisCanvi 153
1.1. Datos básicos sobre RisCanvi 154
1.2. Algunas ventajas y desventajas de las herramientas de valoración y gestión del riesgo, en particular en el funcionamiento de RisCanvi 155
1.3. Requisitos tecnológicos y jurídicos básicos para una posible adaptación de RisCanvi 161
1.4. Objetivo y estructura del capítulo 161
2. Marco normativo del Consejo de Europa en el ámbito penal sobre el uso de la inteligencia artificial y tecnologías similares 162

Página

3. Marco jurídico sobre la inteligencia artificial en la Unión Europea 168

4. Marco jurídico básico, comunitario y español, sobre protección de datos, en particular en relación con el uso de perfilaciones y predicciones en el ámbito penitenciario: El Reglamento General de Protección de Datos (RGPD) de 2016 de la UE, la Ley Orgánica 3/2018 de Protección de Datos Personales y Garantía de los Derechos Digitales, la Directiva 2016/680 y la Ley Orgánica 7/2021 sobre protección de datos personales en el ámbito penal, y su conexión con la Ley Orgánica General Penitenciaria y el Reglamento Penitenciario 173

4.1. Normativa europea y española general sobre protección de datos 173

4.2. Normativa europea y española de protección de datos en el ámbito penal-penitenciario: la Directiva 2016/680, la Ley Orgánica 7/2021 sobre protección de datos personales en el ámbito penal y el Reglamento penitenciario español 176

5. Contraste del marco normativo con un análisis de resoluciones judiciales que mencionan la Tabla de Valoración de Riesgo y la herramienta RisCanvi 180

6. Contraste a través de un trabajo de campo realizado en Euskadi: Observación no participante de la toma de decisiones técnicas en las tres prisiones vascas 196

6.1. Metodología del trabajo de campo de observación no participante 198

6.2. Consideraciones éticas 199

6.3. Limitaciones 200

6.4. Resultados del análisis temático con palabras y extractos ilustrativos 200

6.4.1. Complejidad en el manejo y limitaciones de la información disponible 201

6.4.2. Interdisciplinariedad en el trabajo en equipo 202

6.4.3. Retos y complementariedad a la hora de llegar a consensos desde la diversidad de perspectivas

Página

técnicas y humanas, donde se reconocen posibles sesgos 202

6.4.4. Manejo diverso de tablas en papel de factores de valoración como criterios de orientación y justificación 205

6.4.5. Desconocimiento, suspicacia, interés y ofrecimiento de colaboración ante la experiencia del RisCanvi en Cataluña y su posible adaptación en el País Vasco 209

6.4.6. Necesidad de más recursos humanos, materiales y tecnológicos 210

6.5. Consideración final 210

7. Contraste a través de un cuestionario y un grupo de opinión desarrollado con estudiantes universitarios vascos 211

7.1. Resultados: número de respuestas y perfiles sociodemográficos 212

7.2. Mayor identificación con alguna de las personas que participan en el pódcast sobre RisCanvi (abogados/as, funcionarios/as, técnicos/as que la elaboraron, perfeccionan y/o auditan, condenados/as...) y mayor preocupación en el uso de RisCanvi 213

7.3. Mayor beneficio que aportaría la utilización de este tipo de herramienta en las prisiones vascas 217

7.4. Adecuación de la inversión de fondos públicos en este tipo de herramienta y, en su caso, de sistemas de control 220

7.5. Otros comentarios adicionales 223

7.6. Grupo de discusión 224

7.7. Consideraciones generales 224

8. Recapitulación: Navegando entre un juicio clínico, un juicio probabilístico y un sistema mixto 224

9. Referencias 233

Página

CAPÍTULO 8

LAS DOS CARAS DE LOS ALGORITMOS DE ESTIMACIÓN DE RIESGO EN EL ÁMBITO PENITENCIARIO: RECOMENDACIONES PARA CONJUGAR PREVENCIÓN Y GARANTÍAS

LUCÍA MARTÍNEZ GARAY 245

1. Introducción 246

2. Las dos caras de RisCanvi 246

3. La primacía del aspecto tratamental en el diseño, implementación y supervisión de RisCanvi 248

4. Paradojas que produce la tensión entre la dimensión tratamental/preventiva, y la de garantías 251

4.1. ¿Puede la capacidad predictiva de RisCanvi ser buena y mala a la vez? ¿Puede afirmarse al mismo tiempo que RisCanvi minimiza los falsos positivos, y que produce demasiados? 251

4.2. La capacidad predictiva es mejor o peor según cuál sea el objetivo que perseguimos. ¿Cuál es el objetivo y quién debe establecerlo? 258

4.3. ¿Produce RisCanvi efectos discriminatorios prohibidos, o no los produce? 260

4.4. ¿Es RisCanvi una herramienta que mejora la transparencia de las decisiones, o un algoritmo opaco que vulnera el derecho de acceso a los datos personales y el derecho de defensa? 267

5. Conclusiones generales 271

6. Recomendaciones concretas si se quiere implantar alguna herramienta parecida a RisCanvi en el sistema penitenciario vasco 276

6.1. Mantener la propiedad y el control sobre los modelos y sobre los datos 276

6.2. Necesidad de evaluación previa, continua, y por resultados agregados 277

6.3. Transparencia 278

Página

6.4. Conflicto con protección de derechos 279

7. Referencias 284

PARTE III: ANÁLISIS CRÍTICO

CAPÍTULO 9

ALGORITMO, ALGORITMO… ¿ME PODRÍAS DECIR QUÉ PUEDES APORTAR MEDIANTE LA PREDICCIÓN DE RIESGO DE REINCIDENCIA EN LA MISIÓN DE FACILITAR QUE TODAS LAS PERSONAS QUE DEBAN CUMPLIR UNA PENA Y MEDIDA PENAL EN EUSKADI LO HAGAN EN CONDICIONES DE DIGNIDAD, RESPETANDO SUS DERECHOS Y AYUDÁNDOLES A PREPARARSE BIEN PARA SU RÁPIDA REINTEGRACIÓN A LA SOCIEDAD?

JUAN ALDAZ ARREGUI 291

1. Introducción 292

2. Riesgos de la predicción de riesgos 294

2.1. Una mirada general al marco de la sociedad del riesgo 295

2.2. El paradigma del riesgo en el ámbito penitenciario 296

3. La dificultad de cuantificar y predecir el futuro 299

4. Algunos de los dilemas centrales de la potencial aplicación de algoritmos predictivos al ámbito penitenciario 303

4.1. Datificación de la realidad. Entre la calidad y la calidez de los datos 304

4.2. El procesamiento de los datos. Entre la búsqueda de la agilización y el riesgo de la precipitación 307

4.3. Decisiones informadas. Entre sesgos y aprendizajes 309

5. Retos organizativos previos a la potencial implementación de un algoritmo predictivo 310

6. Conclusiones 313

7. Referencias 314

Página

CAPÍTULO 10

APROXIMACIÓN CRÍTICA AL USO DE INSTRUMENTOS ALGORÍTMICOS EN LA EJECUCIÓN PENITENCIARIA DESDE LA PERSPECTIVA DEL DERECHO PENAL

LORENA ALEMÁN ARÓSTEGUI 317

1. Introducción: el uso de algoritmos y la determinación del castigo penal 317

2. El Derecho penal y la incidencia de la peligrosidad en la determinación del castigo 319

2.1. Derecho penal de la culpabilidad por el hecho 320

2.2. El ámbito de valoración de la peligrosidad: su extensión a las penas y a los sujetos imputables 322

3. Peligrosidad, riesgo y uso de algoritmos en la ejecución penitenciaria 324

3.1. Peligrosidad versus valoración del riesgo 325

3.2. Aspectos positivos y negativos del uso de instrumentos algorítmicos frente al juicio clínico en la ejecución penitenciaria . 330

4. Conclusiones 337

5. Referencias 343

CAPÍTULO 11

INSTRUMENTOS DE PREDICCIÓN DE LA PELIGROSIDAD Y EJECUCIÓN PENITENCIARIA: EL AVANCE DEL DERECHO PENITENCIARIO DEL ENEMIGO

ANA I. PÉREZ MACHÍO 349

1. Referencias 354

CAPÍTULO 12

PERSPECTIVAS DE UNA PSICÓLOGA EN PRISIONES VASCAS SOBRE LA EVALUACIÓN DEL RIESGO DE REINCIDENCIA

MARIAN MORENO MATA 357

Página

CAPÍTULO 13

LA REVOLUCIÓN DE LA IA EN EL MUNDO JURÍDICO-JUDICIAL. ALGUNOS PROBLEMAS ACTUALES Y UNA MIRADA HACIA EL FUTURO

ALBERTO SAIZ GARITAONANDIA 361

1. Introducción. la IA en su momento "Manhattan" 361

2. La IA y su uso en el ámbito jurídico-judicial: de las primeras experiencias predictivas a la IA generativa 363

2.1. Experiencias de uso de IA en la predicción de resoluciones judiciales 363

2.2. El salto cualitativo de la IA generativa y su capacidad de aportar textos complejos 365

3. Algunas notas sobre el funcionamiento y los problemas del uso de la IA y su proyección en el ámbito jurídico 368

3.1. La aproximación probabilística como base del funcionamiento de la IA 368

3.2. La trascendencia y limitaciones de los datos pasados como fuente del modelo de IA 370

3.3. Los bias o sesgos algorítmicos 371

3.4. Los sesgos humanos ocultos y su proyección algorítmica 372

3.5. ¿Y qué hacemos con los datos sesgados? 374

4. La ironía de la automatización como riesgo en el uso de IA 375

5. A modo de conclusión: herramientas del presente y observaciones hacia el futuro de una tecnología en plena efervescencia 376

6. Referencias 378

Prólogo

La orientación resocializadora de la ejecución de la pena privativa de libertad —que, no por compartida, deja de recibir fuertes críticas, tanto desde las perspectivas ideológicas más conservadoras como por sus propias dificultades teóricas y prácticas y a la vista de los excesos propiciados en el plano terapéutico, contrarios a los derechos fundamentales de los presos— conlleva importantes consecuencias en la vida penitenciaria. Esta, en primer lugar, debería esforzarse no ya por no agravar, sino, en la medida de lo posible, atenuar el efecto desocializador inherente a toda decisión de internamiento en prisión, estructurando una convivencia basada, de verdad, en el reiterado principio de que el interno sigue formando parte de la sociedad y, por tanto, asimilada a la vida en libertad, plenamente respetuosa de los derechos no afectados por la condena y abierta a la comunicación con la vida del entorno, cuya participación en la esfera prisional es igualmente de fomentar.

Pieza central de la vida penitenciaria ha de ser, igualmente, el ofrecimiento de alternativas al comportamiento criminal, lo que, lejos de cualquier socialización coactiva y sin perjuicio, en su caso, de las apropiadas terapias psicológicas y similares, ha de buscarse principalmente a través de la oferta de programas de educación para la convivencia, propiciando en los ciudadanos presos el respeto de los principios democráticos y la legalidad formal.

Caracteriza, en todo caso, a los sistemas penitenciarios el atender en no pocas decisiones (en particular, las relativas a clasificación y grado) al progreso en la reeducación y el riesgo de reincidencia. La disponibilidad de las mejores herramientas que permitan una adecuada medición de estas circunstancias constituye, por ello, una preocupación muy extendida en las administraciones competentes, que buscan inspirarse en los mejores modelos para incorporar a sus prácticas los sistemas más apropiados de evaluación y gestión de riesgos penitenciarios. Destacan entre ellos los sistemas predictivos basados en algoritmos, como el sistema "RisCanvi", aplicado desde hace más de una década por la administración penitenciaria catalana.

No son pocas las cuestiones problemáticas que suscitan estos modelos y sistemas, cuya aplicación específica ha sido recientemente estudiada en el marco del Consejo de Europa, por parte de la *Recomendación CM/Rec(2024)5 del Comité de Ministros a los Estados miembros sobre los aspectos éticos y organizativos del uso de la inteligencia artificial y las tecnologías digitales conexas por los servicios penitenciarios y de libertad condicional.* Tras reclamar la garantía de los principios de respeto de la dignidad humana y derechos fundamentales, legalidad, seguridad jurídica y responsabilidad, igualdad de trato y no discriminación, proporcionalidad, eficacia y necesidad de la IA, buena gobernanza, transparencia, trazabilidad y explicabilidad, derecho al control humano de las decisiones tomadas, calidad, fiabilidad y seguridad, uso centrado en el ser humano, así como educación digital y en IA (1-9), dedica la Recomendación su apartado V B) a la regulación de su aplicación "con fines de gestión de los autores de delitos, evaluación de riesgos, rehabilitación y reinserción". Se advierte ahí, entre otras, de "las dificultades que plantean los sesgos algorítmicos y la calidad y representatividad de los datos", afirmando expresamente que "la sensibilidad hacia todos los tipos de diversidad, incluida la perspectiva de género y el multiculturalismo, debería guiar el diseño y el uso de las herramientas de evaluación de riesgos para evitar cualquier discriminación" (19). Añade, asimismo, la Recomendación que "los resultados de la evaluación de riesgos solo deberían utilizarse para la gestión de riesgos", sin permitir la automatización de las decisiones, que han de "ser tomadas por profesionales debidamente designados" (20).

A la vista de lo anterior y del interés por parte de la Administración penitenciaria de Euskadi por analizar estos sistemas de cara a su posible incorporación en el sistema vasco, el día 13 de octubre de 2025 tuvo lugar en el Campus de Gipuzkoa, de la Universidad del País Vasco, un Círculo de diálogo, entre personas expertas y profesionales de diferentes ramas del saber, sobre: "*Retos criminológicos ante los instrumentos de valoración y gestión de riesgos en el ámbito penitenciario. Una exploración de su adaptación al ámbito vasco*". Este Círculo de diálogo se desarrolló en el marco de las actividades del Convenio de Colaboración en materia penitenciaria entre el Departamento de Justicia del Gobierno Vasco y el Instituto Vasco de Criminología (EHU) (2025-2028), en concreto, dentro de una investigación liderada por Gema Varona y Juan Aldaz.

El presente volumen recoge los materiales más relevantes del Círculo de Diálogo, junto con la investigación propia realizada por miembros y personal investigador del IVAC-KREI. A lo largo de estas páginas se resalta la preocupación por el modo de asegurar tanto el respeto de la normativa de

protección de datos a la hora del manejo de estas herramientas internacionalmente reconocidas "de alto riesgo" como, asimismo, la fiabilidad y calidad de unos datos que, basados en modelos estadísticos construidos sobre variables predefinidas de personas con perfiles similares a las evaluadas, deberían quedar libres de todo sesgo o riesgo de generar estigmatización, y ayudar en la toma de decisiones individualizadas, sin sustituir a los profesionales responsables cuya capacitación en el uso de estas herramientas debe quedar igualmente bien garantizada. Se subraya igualmente la necesidad de establecimiento de una arquitectura clara de responsabilidades y dación de cuentas, así como la transparencia del funcionamiento interno del algoritmo —que debería contribuir a la reinserción social y no dar lugar a "automatismos punitivos", cuestiones especialmente subrayadas en el Círculo de Diálogo y matizadas en el intercambio de ideas, también con profesionales de prisión y estudiantes del grado de Criminología y otras disciplinas.

Este volumen se presenta como una invitación a seguir profundizando y reflexionando sobre todo ello desde una perspectiva interdisciplinar.

José Luis de la Cuesta Arzamendi

Presidente del Comité de Honor. IVAC-KREI

PARTE I: ANÁLISIS DESCRITIVO

¿Qué son las herramientas algorítmicas?
¿Qué es la Inteligencia Artificial?
¿Dónde se aplican en los sistemas penales
y penitenciarios?, y ¿para qué?

Capítulo 1

Valoración del riesgo e IA en el ámbito penitenciario en la era de las *Smart Prisons*: andar antes de correr

Prof. Dr. Fernando Miró Llinares
Crímina. Universidad Miguel Hernández de Elche

Sandra Pérez Domínguez
Crímina. Universidad Miguel Hernández de Elche

SUMARIO: 1. LO INTELIGENTE (Y LO NO TANTO) DEL TÉRMINO "SMART PRISONS". 2. EL CONTEXTO: DIGITALIZACIÓN, ALGORITMIZACIÓN E IA EN PRISIONES. 3. LA VALORACIÓN DEL RIESGO EN EL ÁMBITO PENITENCIARIO EN TIEMPOS DE IA. 4. ANDAR (EN LA DIRECCIÓN ADECUADA) ANTES QUE CORRER O QUEDARSE INMÓVIL. 5. REFERENCIAS.

1. LO INTELIGENTE (Y LO NO TANTO) DEL TÉRMINO "SMART PRISONS"

La prisión como espacio de experimentación tecnológica, especialmente en el ámbito de la seguridad, la vigilancia y el control, no es algo nuevo. En los últimos años, sin embargo, las dinámicas de innovación tecnológica en la prisión y en lo penitenciario se han intensificado y, podríamos decir, complejizado, en un nuevo ecosistema de datos, infraestructuras digitales y tecnologías emergentes (Jewkes & Reisdorf, 2016) que ha dado lugar a lo que ha venido denominándose *Smart Prisons* (McKay, 2022). Es esta una etiqueta que, a modo de eslogan, viene a amalgamar procesos heterogéneos, hijos de la revolución digital, unidos por el ámbito en que se aplican, lo penitenciario. Entre ellos está la digitalización de expedientes de internos

y la gestión de procesos administrativos en las prisiones; la implantación de sensores biométricos y sistemas de gestión de movimientos entre otras tecnologías para la seguridad de los centros; el uso de plataformas de e-learning, sistemas de acceso a internet "blanco" o controlado y la utilización de videoconferencias para el día a día en las prisiones; o la proliferación de herramientas algorítmicas, desde las más clásicas de valoración del riesgo hasta las que usan IA, para la toma de decisiones relacionadas con la clasificación, el tratamiento o la progresión de grado (Miró Llinares, 2023). Pero, al integrarse todas estas tecnologías y técnicas de diversa naturaleza, y con muy diferentes implicaciones cada una de ellas, dentro del paraguas de la adjetivación "smart", lo que se expresa es algo más que encontrarnos ante un proceso natural de adopción puntual de tecnologías digitales en el ámbito penitenciario. La *smart prison* se caracterizaría por la creación de un entorno penitenciario altamente datificado y conectado, en el que la información generada por las personas privadas de libertad, el personal penitenciario y los dispositivos tecnológicos se recopila, centraliza y procesa de forma continua para convertirse en un recurso estratégico destinado a optimizar procesos, anticipar riesgos y orientar la toma de decisiones "inteligentes" (McKay, 2022). Con ello, implícitamente, el término "prisión inteligente" transmite que estaríamos ante una auténtica transformación contemporánea de la prisión en la que las tecnologías de la información y la comunicación se integran de manera estructural en su infraestructura y en el propio sentido del sistema penitenciario para, básicamente, convertirla en algo "mejor".

Ahora bien, y sin entrar en un estudio detallado de si tras los procesos reales de incorporación de variadas tecnologías digitales en las prisiones hay una racionalidad común o, como ha ocurrido en tantos otros ámbitos de la vida social, nos encontramos ante una mera tendencia a la acumulación sucesiva de capas de "innovación" tecnológica bajo la creencia de que todo cambio conlleva progreso, lo que sí es evidente es que el uso acrítico de la etiqueta *Smart prison* podría oscurecer lo que, a nuestro parecer, debería ser lo esencial siempre que una tecnología viene a sustituir a otra en un proceso. Nos referimos al análisis de qué mejoras y qué problemas concretos introduce la tecnología; cómo redistribuye capacidades y vulnerabilidades entre los actores afectados e implicados en el castigo y su gestión; y, hasta qué punto su uso es compatible con los principios que legitiman el ejercicio del poder punitivo. Se trata, en definitiva, de pensar en cuál es el objetivo del uso de una determinada tecnología, y de hacerlo antes de su adopción. Comenzar por ahí, y hacerlo no mezclando tecnologías y procesos que son distintos entre sí, sino diferenciando el análisis para cada una de las concretas herramientas y para cada uno de los posibles usos de las mismas, no solo sería lo más "inteligente",

sino que nos ayudaría a escapar de la tradicional dicotomía de actitudes ante el uso de nuevas tecnologías (Miró Llinares, 2020) que también ha alcanzado a las *Smart prisons*. Por un lado, la visión tecnoutópica, donde la prisión inteligente se presenta como una institución más eficiente, segura y orientada a la rehabilitación, capaz de mejorar las condiciones laborales del personal penitenciario, optimizar el uso de los recursos disponibles y ofrecer respuestas más individualizadas a las necesidades de los internos y que se correspondería con casos como el proyecto "Smart Prison Project" de la Agencia Finlandesa de sanciones penales, en la que la digitalización y la IA se definen como un instrumento al servicio de los fines rehabilitadores del sistema mediante la promoción de la autonomía, la agencia y las competencias para la vida de las personas privadas de libertad (Puolakka, 2022). Por otro, la visión distópica, para la que todas estas tecnologías acaban siendo instrumentos que refuerzan y amplifican las lógicas tradicionales de control, vigilancia e inocuización, profundizando la deshumanización del encierro y extendiendo la capacidad de supervisión e incapacitación más allá de los muros de la prisión y que se vería plasmada en los modelos penitenciarios altamente tecnificados de países como China, Singapur, Australia, Estados Unidos o Turquía en los que el énfasis se sitúa fundamentalmente en la automatización de la vigilancia y el control de la población reclusa (McKay, 2022).

Ambas visiones tienen serias limitaciones. La de la defensa de la "Smart prison" obvia la cuestión fundamental de plantearse qué podría llegar a significar que los muros se llenen de sensores, que los expedientes se conviertan en bases de datos interoperables a diversos efectos, o que la trayectoria de una persona presa se traduzca en puntuaciones de riesgo, perfiles de tratamiento o indicadores de "progresión" calculados de forma automatizada; la de la crítica genérica a las "Smart prisons" acierta en los peligros pero en ocasiones niega la posibilidad de que se use la tecnología para mejorar los modos defectuosos o limitados de hoy para tomar decisiones que afectan a los derechos de los internos, las enormes posibilidades de la digitalización para la reintegración futura de los presos o el mero hecho de que la IA pueda utilizarse para identificar sesgos, malos usos y necesidades específicas en el ámbito penitenciario (Miró Llinares, 2020). Conviene, por ello, dejar generalizaciones y etiquetas a un lado y desagregar tecnologías y usos para comprender qué añaden realmente las tecnologías digitales, los algoritmos y la IA sobre las formas previas del gobierno y el día a día penitenciario; qué continuidades y qué rupturas introducen y de qué manera reconfiguran la toma de decisiones, la distribución de responsabilidades y los márgenes de control y discrecionalidad. Esto está en línea con la reciente Recomendación CM/Rec(2024)5 del Comité de Ministros del Consejo de Europa, que esta-

blece un marco ético y organizativo para el uso de la inteligencia artificial y de las tecnologías digitales en los servicios penitenciarios y de probation. Lejos de avalar una adopción indiscriminada de soluciones "inteligentes", la Recomendación insiste en la necesidad de un uso legítimo, proporcionado y orientado a la rehabilitación, subrayando principios como el control humano efectivo, la prohibición de decisiones automatizadas que afecten a derechos fundamentales, la transparencia, la rendición de cuentas y la centralidad de la dignidad humana. Este marco refuerza la idea de que el debate sobre las *smart prisons* no debe formularse en términos abstractos o globales, sino a partir del análisis concreto de tecnologías, finalidades y contextos de uso.

En este contexto y desde ese enfoque, el objetivo de este trabajo es doble: por un lado, y en el marco del proyecto en el que se ha realizado el mismo, pretendemos contribuir a la cuestión de si es recomendable el uso de herramientas de valoración del riesgo en el ámbito penitenciario y de cuál debería ser el propósito y, desde él, el diseño de estas. Por otro lado, pretendemos llevar a cabo la reflexión en el marco de un contexto de discusión académica más profunda y de una profunda tendencia de transformación de las prisiones protagonizada por la utilización de múltiples tecnologías digitales no siempre llevada a cabo con la reflexión necesaria. Por eso comenzaremos por tratar de diferenciar conceptualmente los procesos de digitalización, algoritmización e incorporación de sistemas de inteligencia artificial en el ámbito penitenciario, con el objetivo de clarificar sus lógicas específicas. Luego veremos los distintos usos y ámbitos de aplicación que se plantea que puede ofrecer la inteligencia artificial en las prisiones. Y finalmente entraremos en la discusión sobre el uso de herramientas de valoración del riesgo en el ámbito penitenciario, analizando las condiciones bajo las cuales puede resultar recomendable el uso de estas e, incluso, planteando si debemos o no avanzar hacia modelos basados en inteligencia artificial para este tipo de decisiones.

2. EL CONTEXTO: DIGITALIZACIÓN, ALGORITMIZACIÓN E IA EN PRISIONES

Como se ha dicho, la incorporación de tecnologías digitales en el sistema penitenciario se inscribe dentro de una dinámica general de modernización del sistema de justicia penal, impulsada por la denominada Cuarta Revolución Industrial (Schwab, 2016). En este contexto, la digitalización, la algoritmización y la inteligencia artificial (IA) se presentan con frecuencia como elementos de una misma dinámica de modernización institucional. Sin embargo, el uso indistinto de estos conceptos ha generado una notable

confusión analítica, dificultando la delimitación de sus efectos específicos y planteando la cuestión de si nos encontramos ante procesos progresivos y convergentes o ante un *totum revolutum* caracterizado por la superposición acrítica de tecnologías heterogéneas.

La digitalización constituye el nivel más básico de esta transformación tecnológica y se refiere, fundamentalmente, a la conversión de información, procedimientos y soportes analógicos en formatos digitales, sin que ello implique necesariamente un cambio en la lógica sustantiva de funcionamiento del sistema (Brennen & Kreiss, 2016). Desde una perspectiva conceptual, la digitalización actúa como un proceso de traducción tecnológica, orientado a trasladar prácticas preexistentes a entornos digitales, conservando intactos los marcos normativos, los criterios de decisión y las estructuras organizativas subyacentes. En el ámbito penitenciario, este proceso se ha materializado en la implantación de sistemas de información penitenciaria, la digitalización de expedientes administrativos y judiciales, la gestión electrónica de datos personales y penitenciarios, así como en el uso de tecnologías de videovigilancia y control biométrico (Güerri Ferrández et al., 2021; López Lorca, 2023). Estas herramientas han permitido optimizar la gestión administrativa, mejorar la coordinación interinstitucional y aumentar la trazabilidad, la estandarización y la transparencia formal de las actuaciones penitenciarias. No obstante, la digitalización no implica por sí misma una transformación sustantiva de la toma de decisiones, en la medida en que no altera los criterios normativos ni desplaza el papel central del juicio humano, limitándose a reforzar la eficiencia, rapidez y capacidad de registro de procesos ya existentes (Hildebrandt, 2015).

Un segundo nivel de complejidad viene dado por la algoritmización, entendida como la introducción de procedimientos formalizados de cálculo y modelización estadística capaces de procesar grandes volúmenes de datos y generar resultados estructurados, como puntuaciones, clasificaciones o predicciones, que influyen de manera directa o indirecta en la toma de decisiones. A diferencia de la digitalización, la algoritmización sí introduce una modificación relevante en la forma en que la información es tratada y jerarquizada, ya que incorpora una lógica cuantitativa y probabilística que orienta la valoración de los casos. En el ámbito penitenciario, este proceso se ha desarrollado principalmente a través de las denominadas herramientas de valoración del riesgo, diseñadas para estimar la probabilidad de reincidencia, de comisión de nuevos delitos o de comportamientos problemáticos, así como para apoyar decisiones relacionadas con la clasificación penitenciaria, la progresión de grado o la concesión de permisos (Brandariz García, 2016;

Rizer & Watney, 2018). Estos sistemas, de carácter actuarial, se presentan como mecanismos orientados a reducir la discrecionalidad individual y aumentar la consistencia y previsibilidad de las decisiones, incorporando una lógica de *evidence-based sentencing* y de gestión penitenciaria basada en datos empíricos (Hanson & Bussière, 1998; Gendreau et al., 1996). Sin embargo, aunque formalmente se configuran como herramientas de apoyo, la algoritmización introduce una reconfiguración del peso relativo de la información en el proceso decisorio, al priorizar determinados factores y correlaciones en detrimento de otros, lo que plantea interrogantes sobre su impacto real en la autonomía profesional y en la racionalidad jurídica.

La inteligencia artificial representa el estadio más avanzado de este proceso de transformación tecnológica. Desde un punto de vista conceptual, la IA se distingue de los algoritmos tradicionales por la incorporación de técnicas de aprendizaje automático (*machine learning*) que permiten a los sistemas identificar patrones complejos, ajustar sus modelos en función de nuevos datos y mejorar su rendimiento sin intervención humana directa constante. Esta capacidad de aprendizaje y adaptación incrementa tanto su potencial predictivo como su grado de autonomía operativa (Alarie et al., 2018), diluyendo progresivamente la trazabilidad entre los datos de entrada y los resultados generados. Y es que la inteligencia artificial adquiere relevancia como elemento distintivo del paradigma de las *smart prisons*. Si bien muchas de las tecnologías actualmente implantadas en las prisiones, como los sistemas de gestión penitenciaria, la videovigilancia o los dispositivos de control de accesos pueden funcionar sin recurrir a técnicas avanzadas de IA, la incorporación de sistemas algorítmicos capaces de analizar grandes volúmenes de datos, identificar patrones de comportamiento y formular predicciones marca un salto cualitativo en el proceso de transformación digital. La IA permite pasar de una prisión digitalizada a una prisión "inteligente" en sentido estricto, en la medida en que introduce capacidades de aprendizaje, anticipación y toma de decisiones automatizadas.

Ahora bien, una vez delimitados conceptualmente los distintos niveles de transformación tecnológica, resulta necesario preguntarse en qué ámbitos concretos podría desplegarse el uso de la inteligencia artificial en el contexto penitenciario. Como ha señalado uno de los autores de este trabajo en otro lugar (Miró Llinares, 2023), pueden distinguirse tres grandes ámbitos en los que el desarrollo y la aplicación de sistemas de inteligencia artificial presentan un potencial particularmente significativo: la gestión cotidiana de las prisiones; la seguridad, el control y la vigilancia de los internos; y el tratamiento de los condenados conforme a los fines rehabilitadores del

sistema penitenciario. Es cierto que los dos últimos son especificaciones del primero, pero a efectos didácticos parece más adecuado diferenciarlos.

A) En primer lugar el sentido del desarrollo de las *Smart prisons* y del uso de sistemas de IA dentro de ellas responde a la voluntad de mejorar la eficiencia de la gestión penitenciaria. Desde una perspectiva tecnoutópica la gestión de los servicios penitenciarios y de libertad condicional, al igual que el de otras instituciones públicas y privadas, podría mejorarse, con la optimización de los procesos humanos y de gestión por medio de sistemas de IA (Miró Llinares, 2023). Si la administración penitenciaria busca racionalizar la gestión del personal penitenciario, mejorar la seguridad y dar más autonomía a los detenidos en ciertas acciones diarias, la digitalización de los procesos estaría encaminada a ello para, en última instancia, mejorar las relaciones entre el personal y lograr los efectos pretendidos. Son ya muchos los estados que, con ese propósito de eficiencia, ya han procedido a la recopilación y gestión de información sobre los detenidos en relación con su encarcelamiento (ingreso inicial, registros de cargos, sentencia, liberación temporal), gestión de la vida diaria en prisión (médico, planificación de comidas, programación de actividades y visitas), gestión del tiempo del personal (Krikorian & Coye, 2019). El siguiente paso, en la utopía de la eficiencia por medio de IA, consistiría en la utilización de esos datos para la automatización de algunas decisiones, la identificación de las fallas y de los aciertos y el análisis inteligente del actuar en el ámbito penitenciario para eliminar la subjetividad y lograr la objetividad (Miró Llinares, 2023). Y a la luz de esos objetivos de lo que suelen ser instituciones públicas, las empresas privadas han comenzado a desarrollar y ofrecer *software* de gestión penitenciaria con cada vez más funcionalidades y módulos novedosos, destinados a la gestión diaria de la prisión (Krikorian & Coye, 2019).

Como se ha dicho, la IA es la herramienta esencial que lograría el siguiente paso: la información en tiempo real proporcionada por este tipo de herramientas serviría para optimizar el uso de los recursos, comprender cómo están funcionando la organización y el personal y, de ese modo, mejorar la toma de decisiones a nivel de gestión de las organizaciones (procesos de selección y contratación de personal, para su formación, para la determinación del presupuesto de la empresa, etc.). Eso sí, lo que la herramienta defina como uso óptimo determinará todo: un uso rentable de los recursos debiera favorecer el bienestar del personal y ayudar a equilibrar su vida profesional y familiar, pero quizás para lo que sirva es para beneficiar únicamente a los fines de la organización en términos de eficiencia. Del mismo modo la tecnología de IA también puede utilizarse para reducir el sesgo

humano inconsciente en los procesos de RRHH, sustituyendo la toma de decisiones humana en determinados pasos, pero dado que su diseño también está hecho por humanos también es posible que genere sesgos y que estos se vayan consolidando como algo "objetivo" y neutro una vez se han datificado (Miró Llinares, 2023).

Gran parte de la gestión del día a día tiene que ver con el trato con los internos, y la IA también acabaría teniendo un rol en este ámbito. Como ha señalado McKay (2022), hoy las tecnologías contemporáneas también están llegando a las prisiones y la crisis del Covid-19 ha acelerado estos procesos también: las videoconferencias, las tabletas digitales, los iKiosks y las computadoras en las celdas, no han sustituido aún pero lo harán, al teléfono fijo compartido tradicional de las prisiones y a las cartas postales. Los dispositivos digitales personales facilitan nuevas formas de mantener conexiones familiares y oportunidades para acceder a la justicia y permiten a los presos emprender programas, estudiar y desarrollar su alfabetización digital (McKay, 2022)[1]. En principio la IA podría usarse para la automatización del uso de determinados dispositivos, para la selección personalizada de los programas a usar por cada interno, entre otras funcionalidades. Pero todos estos dispositivos también plantean la cuestión de la gestión de los datos personales de quienes los usan, así como de la posibilidad de obtener información (aún anonimizada) del uso masivo de este tipo de dispositivos por parte de muchos presos y que ello tenga un valor económico y sea utilizado como objeto de comercio por las propias prisiones para crear herramientas de IA a partir de estos datos (Miró Llinares, 2023). De hecho, una de las potenciales funciones de la IA puede ser la mejora de la gestión de la controvertida cuestión de la conectividad digital de los reclusos, que está comenzando a reconocerse gradualmente en muchos países[2]. En EEUU hasta nueve prisiones estatales han adoptado tabletas para ser utilizadas por personas privadas de libertad desde 2017 (Finkel & Bertram, 2019) y, en un esfuerzo por mejorar la dimensión rehabilitadora del castigo, varios estados de la UE se comprometieron recientemente a brindar a las personas detenidas acceso a aprendizaje electrónico, búsqueda de empleo en línea o ciertos sitios web de servicios públicos (Krikorian & Coye, 2019). Aun así, esos servicios a menudo se reservan para las instalaciones de seguridad mínima, no están disponibles para los detenidos en instalaciones cerradas, y existe una discusión sobre si las personas encarceladas deben poderse

1. Véase también (Krikorian, J., & Coye, S., 2019). Véase desde las percepciones de las personas en prisión, Thaler, Barkworth, & Howard, 2022.
2. Véase, sobre esta cuestión, el interesante trabajo de (Jewkes & Reisdorf, 2016).

comunicar con personas fuera de la prisión por correo electrónico, tarjetas electrónicas, videogramas o videollamadas o realizar otras actividades de ocio, o de gestión de su vida penitenciaria o para su capacitación laboral y de rehabilitación (Mufarreh, Waitkus & Booker, 2022). Sin entrar en la cuestión de la bondad de estas medidas[3], lo cierto es que las herramientas de IA podrían servir para la selección de contenidos en el caso de que se quiera limitar el acceso a listas blancas, o incluso para controlar el tipo de acceso a internet y evitar que se puedan acceder a contenidos nocivos o realizar conductas no permitidas que podrían facilitar la perpetración de delitos a distancia (Miró Llinares, 2023).

B) Dentro de la gestión de las prisiones en el día a día el ámbito en el que se percibe mayor potencialidad al uso de sistemas de IA en "lo penitenciario" es el de la seguridad y la vigilancia en las propias prisiones. Conforme a la visión tecnofílica, que ya se basa en algunas experiencias reales especialmente en prisiones asiáticas altamente tecnificadas, la digitalización permite un significativo incremento de la seguridad y el control sobre las poblaciones carcelarias, y la IA podría constituir el paso definitivo para la automatización de procesos y decisiones que irían en la línea de mejorar tal control y vigilancia (Miró Llinares, 2023). Como ha señalado McKay (2022), las prisiones pueden ser entornos volátiles para los funcionarios de prisiones y los reclusos, con ataques de recluso a recluso y de recluso al personal, que ocurren particularmente en instalaciones superpobladas, y para contrarrestar tales peligros, especialmente para el personal penitenciario, las prisiones inteligentes con formas de monitoreo y vigilancia automatizadas, IA y/o robóticas pueden reducir la necesidad de que los funcionarios de prisiones interactúen físicamente con los reclusos, proporcionando un entorno de trabajo más seguro.

El paradigma más futurista, pero ya casi real, constituiría el celador robot que vendría a sustituir al trabajador humano que se ve en situaciones de riesgo para garantizar la seguridad en las prisiones (Miró Llinares, 2023). Ya se han reportado en Asia pruebas de guardias robot que, por medio de cámaras tridimensionales, sensores y algoritmos de reconocimiento de patrones para identificar comportamientos irregulares o inapropiados, patrullan la prisión de forma autónoma bajo la supervisión de un funcionario penitenciario remoto (Kim, 2012)[4]. También en Singapur se ha planteado un modelo de

3. Véase la opinión de Jewkes & Reisdorf (2016) a favor de la provisión de dispositivos a los reclusos para abordar el analfabetismo digital y acceder a programas educativos y de rehabilitación.

4. Cita aquí McKay (2022) los ejemplos de Corea del Sur y Hong Kong. También se están utilizando robots para detectar drogas en la materia fecal de los reclusos. Véanse las referencias que cita en su trabajo.

"prisión sin guardias"[5], y en China la prisión de Yancheng proporciona una vigilancia general continua de los reclusos mediante una red de sensores y cámaras conectadas al «cerebro» de IA que puede rastrear y monitorear continuamente a cada recluso (Chen, 2019). Más allá de esos ejemplos en Asia en los que la IA desempeña ya un rol significativo en las denominadas "Smart prisons", en Europa la IA lo que podría permitir es incrementar las potencialidades de sistemas ya existentes, como las pulseras de identificación que automatizan las puertas de las prisiones, el reconocimiento facial y el análisis de imágenes de circuito cerrado de televisión (CCTV) por medio de IA para evitar peleas o intentos de fuga (Kaun & Stiernstedt, 2019), entre otras que, de nuevo desde la tecnoutopía, permitirían a las autoridades penitenciarias mejorar la automatización y la eficacia en la gestión de tales medidas. Lo mismo puede ocurrir con la gestión de los datos de las cámaras corporales usadas por funcionarios en algunas prisiones en situaciones en las que puede haber problemas de seguridad (Dodd, Antrobus & Sydes, 2020). Más allá de que estas cámaras pueden aumentar la transparencia operativa, y de hecho los oficiales penitenciarios creen que la presencia de cámaras corporales reduce la amenaza de acusaciones falsas y, por lo tanto, mejora su seguridad "profesional" (Adams, 2022; Sydes, Dodd & Antrobus, 2022), la incorporación de sistemas de IA podría automatizar tanto su uso como la gestión de algunos de los datos que la misma puede permitir. Y la vigilancia con la IA no solo permitiría evitar delitos en el exterior. En varias prisiones de EE. UU., se ha reportado que el uso de algoritmos de IA para el análisis automatizado de llamadas telefónicas de presos habilitado por IA evita la organización de delitos desde la prisión (Kaun & Stiernstedt, 2019).

Pero, quizás, donde la IA puede mostrarse especialmente eficaz en el ámbito de la seguridad en las prisiones es en la prevención del contrabando. Como es sabido el contrabando, bien sea de drogas, armas, dinero, tabaco y más recientemente, teléfonos móviles (Peterson, Kizzort, Kim & Shukla) entra en las cárceles de distintas formas (el propio personal de prisiones, los visitantes, los vehículos, el correo) y representa un problema de primer orden para las prisiones (Miró Llinares, 2023). La prevención del contra-

5. Señala McKay (2022) que el Servicio penitenciario de Singapur ha desarrollado un sistema de combinación de cinco tecnologías que incluyen el análisis de video avanzado para detectar agresiones, cámaras de reconocimiento facial en celdas, etiquetas de muñeca escaneables con chips para el día a día de los reclusos, tabletas digitales compartidas para la digitalización de la rehabilitación y la corrección de los reclusos y una herramienta denominada iKiosk, que permite a los reclusos realizar una variedad de tareas de autoservicio, como solicitudes administrativas y privilegios de redención.

bando siempre ha hecho uso de la tecnología, desde los rayos X hasta los detectores de metales, pero al igual que también hay nuevas formas de lograr el contrabando relaciona con las tecnologías más modernas (como el uso de drones), también hay múltiples herramientas digitales, desde las tecnologías de detección y deflector/bloqueo de drones, hasta nuevos sistemas como los "olfateadores" digitales similares a los diseñados en aeropuertos para detectar explosivos en pasajeros civiles, pasando por nuevos sistemas de escáneres de imágenes de cuerpo completo (Leighton, 2014) que ya usan IA o que con la IA podrían aumentar significativamente su eficacia.

Y las posibilidades de la IA en materia de seguridad y control no abarcan solo la prisión, sino que también pueden extenderse al monitoreo electrónico. Australia está probando diferentes soluciones de IA que combinan el análisis de CCTV con sensores inteligentes en el contexto de la libertad condicional y la detención domiciliaria.

C) El último ámbito es el tratamiento de los delincuentes. A dos niveles: por un lado, la IA puede servir para mejorar los sistemas actuariales de valoración del riesgo y, en general, para mejorar la toma de decisiones respecto a la valoración del régimen aplicable a los penados; y por otro, puede servir para ayudar a los internos a mejorar sus habilidades de cara a una potencial reinserción social (Miró Llinares, 2023). Lo primero será objeto de análisis en el capítulo posterior, en el que se examina el papel que desempeñan las herramientas de valoración del riesgo en el ámbito penitenciario, partiendo del debate público suscitado en torno a instrumentos como RisCanvi y aclarando en qué medida estos sistemas pueden (o no) considerarse formas de IA. Hay una segunda vertiente de las posibilidades de uso de la IA en relación con el tratamiento de delincuentes: la de que la digitalización pueda conllevar beneficios de personalización y conectividad que faciliten su rehabilitación y que la automatización de esos sistemas aumente la efectividad de tales acciones. Como han señalado algunos autores (Järveläinen & Rantanen, 2021), la rehabilitación en la sociedad digital debe poder aprovecharse de las posibilidades de conectividad que las TIC ofrecen, como proporcionar experiencias online, usar sistemas de realidad virtual, entre otras tecnologías que pueden facilitar la formación y reeducación de un modo como nunca antes (Miró Llinares, 2023). Aunque son muchos los sistemas y programas que se han implantado, destaca el de las prisiones finlandesas, que busca proporcionar herramientas tecnológicas para promover la rehabilitación y reducir la reincidencia mediante el desarrollo de la autonomía, la agencia, la autoestima y las habilidades de gestión de la vida de los reclusos (Lindström & Puolakka, 2020).

3. LA VALORACIÓN DEL RIESGO EN EL ÁMBITO PENITENCIARIO EN TIEMPOS DE IA.

En los últimos dos años en los medios de comunicación de España han surgido diversas noticias en las que de un modo u otro se hacía mención al uso de algoritmos de IA en el ámbito penitenciario catalán. Con mensajes como "el algoritmo que determina el futuro de los presos en Cataluña" o "algoritmos que ponen en juego las posibilidades de salir o no de prisión", se referían a RisCanvi (Andrés-Pueyo & Echeburúa, 2010), la herramienta utilizada por la administración penitenciaria catalana como apoyo en la adopción de decisiones relativas a la clasificación penitenciaria, la concesión de permisos ordinarios y la progresión a regímenes de semilibertad o libertad condicional. Este tipo de titulares, han contribuido a construir una narrativa pública en la que la toma de decisiones penitenciarias aparece crecientemente automatizada y sin control humano de las decisiones.

Sin embargo, esta representación mediática simplifica en exceso una realidad institucional considerablemente más compleja. Es cierto que la IA puede usarse en el ámbito penitenciario para decisiones como las propuestas por RisCanvi y que, en otros ámbitos, como el penitenciario español, se toman por parte de juntas clínicas sin el apoyo de sistemas de valoración del riesgo. Sin embargo, de ello no se sigue ni que RisCanvi sea, en sentido estricto, un sistema de inteligencia artificial, ni que la toma de decisiones penitenciarias en Cataluña se encuentre automatizada, como sugieren algunas de las narrativas mediáticas. Por tanto, antes de plantear si la IA debería o no incorporarse a este tipo de decisiones, resulta necesario aclarar en qué consistiría exactamente su uso en este ámbito. Esta exigencia de clarificación previa y de cautela no responde únicamente a consideraciones académicas o empíricas, sino que se encuentra expresamente recogida en los estándares europeos más recientes. En particular, el artículo 20 de la Recomendación CM/Rec(2024)5 establece que los resultados de la valoración del riesgo deben utilizarse exclusivamente para la gestión del riesgo y que las decisiones derivadas de dicha valoración no deben ser automatizadas, debiendo ser adoptadas por profesionales debidamente designados, garantizando así la intervención humana cuando están en juego derechos fundamentales.

En este sentido, el debate sobre la inteligencia artificial no puede abordarse como una cuestión meramente técnica ni como un problema instrumental centrado en la eficiencia predictiva, sino que exige una reflexión previa sobre los fines perseguidos, los riesgos asumidos y la capacidad real del sistema penitenciario para integrar estas herramientas sin alterar de forma problemática los equilibrios entre seguridad, tratamiento y garantías. En realidad,

en lo que consistiría la incorporación de la inteligencia artificial a este ámbito es en la introducción de una lógica correlacional avanzada dentro de una práctica que ya existe, la de los algoritmos de valoración del riesgo, con el objetivo de sumar capacidad predictiva y analítica a un mismo propósito: mejorar la eficacia en la anticipación de comportamientos futuros relevantes para la gestión penitenciaria. Ello no implica, sin embargo, una continuidad plena entre ambos enfoques, ya que la lógica subyacente a los sistemas de inteligencia artificial difiere en aspectos relevantes de la que caracteriza a las herramientas actuariales tradicionales (Miró Llinares & Castro Toledo, 2022; Hannah-Moffat, 2019) Mientras que los instrumentos clásicos de valoración del riesgo suelen construirse a partir de marcos teóricos criminológicos explícitos, identificando factores de riesgo y de protección validados empíricamente y organizados en modelos relativamente transparentes (Andrews & Bonta, 2010; Hanson & Morton-Bourgon, 2009), la inteligencia artificial opera de manera esencialmente correlacional, otorgando un peso menor a la explicación causal y a la fundamentación teórica de las variables, y priorizando la identificación de patrones estadísticos complejos en grandes volúmenes de datos (Alarie et al., 2018). En este sentido, la capacidad de la IA para "aprender" a partir de los datos no descansa tanto en hipótesis criminológicas previas cuanto en la detección automática de regularidades empíricas, lo que incrementa su potencial predictivo, pero también su opacidad y dificultad de interpretación. Este desplazamiento epistemológico no elimina las decisiones normativas inherentes a la valoración del riesgo, sino que las traslada a fases menos visibles del proceso, como el diseño del modelo, la selección de los datos de entrenamiento o la definición de los objetivos de optimización. Como advierte Hildebrandt (2015), esta opacidad dificulta la comprensión, la impugnación y el control efectivo de las decisiones, lo que resulta especialmente problemático en un ámbito en el que están en juego derechos fundamentales y expectativas legítimas de libertad. No obstante, esta divergencia lógica y epistemológica obliga, en primer término, a precisar el contenido y el funcionamiento de las herramientas tradicionales de valoración del riesgo, en la medida en que constituyen el punto de partida institucional y conceptual sobre el que se proyecta la incorporación de la inteligencia artificial. Por ello, resulta imprescindible situar la valoración del riesgo en el lugar que actualmente ocupa dentro de la ejecución penal, ya que la inteligencia artificial no irrumpe en un vacío normativo o institucional, sino que se inserta en una práctica ya consolidada que estructura de manera decisiva la toma de decisiones penitenciarias.

La valoración del riesgo se ha convertido, en las últimas décadas, en uno de los ejes estructurales de numerosos sistemas penitenciarios europeos

(Chelioudakis, 2020). Lejos de constituir una técnica auxiliar o marginal, orienta hoy decisiones clave que afectan de manera directa a la situación jurídica y vital de las personas privadas de libertad, como la clasificación inicial, la progresión de grado, la concesión de permisos o el acceso a programas de tratamiento. Esta centralidad se inscribe en un proceso más amplio de transformación de la política penal, caracterizado por el desplazamiento desde modelos centrados en la responsabilidad individual por hechos pasados hacia enfoques orientados a la gestión anticipatoria de comportamientos futuros. Como señalaron tempranamente Feeley y Simon (1992), la denominada *new penology* introduce una racionalidad actuarial que concibe a los sujetos del sistema penal como portadores de probabilidades de riesgo susceptibles de ser administradas mediante técnicas de clasificación y control, desplazando el foco desde la culpabilidad hacia la prevención.

Esta racionalidad prospectiva reconfigura progresivamente la ejecución penal en torno a la estimación de escenarios futuros, introduciendo tensiones relevantes con los modelos clásicos de individualización y rehabilitación, en la medida en que la atención se desplaza desde el análisis retrospectivo de la conducta pasada hacia la anticipación del comportamiento esperado. Conviene subrayar, además, que la centralidad de la valoración estructurada del riesgo y el grado de formalización o algoritmización de las decisiones penitenciarias no son homogéneos en todos los contextos institucionales. En el Estado español coexisten modelos diferenciados, en los que algunas administraciones penitenciarias han desarrollado herramientas específicas de valoración del riesgo como apoyo sistemático a la toma de decisiones, como es el caso de RisCanvi en el contexto catalán (Andrés-Pueyo, & Echeburúa, 2010; Andrés-Pueyo & Redondo, 2007), mientras que otras continúan basándose predominantemente en valoraciones clínicas no estructuradas o en criterios profesionales menos formalizados. Estas diferencias organizativas y normativas condicionan tanto el alcance práctico de los instrumentos existentes como el tipo de problemas que plantea una eventual incorporación de tecnologías más avanzadas, incluida la inteligencia artificial, y obligan a situar el análisis en configuraciones institucionales concretas.

Desde una perspectiva empírica, la expansión de las herramientas estructuradas de valoración del riesgo respondió, en buena medida, a las limitaciones del juicio profesional no estructurado. Una extensa literatura en psicología y criminología ha mostrado de manera consistente que las predicciones basadas exclusivamente en la intuición clínica presentan menores niveles de fiabilidad y validez, así como una elevada variabilidad interevaluador, en comparación con los instrumentos actuariales construidos a partir de evidencia empírica (Gendreau et al., 1996; Grove et al., 2000; Hanson & Morton-Bourgon, 2009).

En este contexto, las herramientas de valoración del riesgo se presentan como un intento de racionalizar la toma de decisiones, introduciendo criterios explícitos, procedimientos homogéneos y marcos de evaluación estandarizados que permitan aumentar la coherencia interna del sistema y reducir la arbitrariedad. Instrumentos como RisCanvi, se inscriben en esta tradición metodológica y no deben equipararse, sin más, a sistemas de inteligencia artificial basados en aprendizaje automático. Se trata de herramientas actuariales estructuradas que organizan factores de riesgo y de protección identificados por la literatura científica mediante ítems definidos, escalas fijas y procedimientos de aplicación homogéneos, manteniendo formalmente la intervención del juicio profesional como elemento central del proceso decisional.

Una de las principales mejoras que introducen estas herramientas reside precisamente en la estandarización del proceso de valoración. Al delimitar de forma explícita las variables consideradas relevantes y establecer criterios comunes para su ponderación, las herramientas de riesgo permiten reconstruir la lógica que subyace a las decisiones adoptadas y facilitan su eventual control externo (Dawes et al., 1989; Grove et al., 2000; Hart, 1998; Hannah-Moffat, 2019; Monahan & Skeem, 2016; Ægisdóttir et al., 2006). En ámbitos tradicionalmente caracterizados por una amplia discrecionalidad profesional y por decisiones difícilmente motivables, esta formalización puede contribuir a reforzar la transparencia y la rendición de cuentas. En la práctica penitenciaria, la valoración del riesgo no constituye un acto aislado ni un resultado autosuficiente, sino un proceso institucional que se despliega a lo largo de distintas fases y a través de diversos actores. La información relevante es recopilada por profesionales técnicos, como psicólogos, trabajadores sociales o educadores, y se integra en informes que son posteriormente valorados por órganos colegiados, como las juntas de tratamiento, en el marco de decisiones administrativas o propuestas elevadas a instancias judiciales. Las puntuaciones resultantes no operan, por lo general, como decisiones automáticas, sino como elementos que estructuran y orientan la deliberación institucional.

Sin embargo, estas mejoras técnicas no están exentas de costes normativos y prácticos. La traducción de trayectorias vitales complejas en puntuaciones de riesgo implica necesariamente una simplificación de la realidad personal y social de las personas internas, privilegiando aquellos factores que resultan fácilmente medibles y comparables en detrimento de dimensiones cualitativas más difíciles de capturar mediante instrumentos estandarizados. Como ha señalado Harcourt (2007), esta lógica actuarial tiende a operar sobre probabilidades grupales, lo que limita su capacidad para anticipar comportamientos individuales y favorece la reificación del riesgo como un atributo estable del sujeto. En el ámbito penitenciario, esta

reificación puede consolidar trayectorias institucionales rígidas y dificultar el reconocimiento de los procesos de cambio personal y evolución a lo largo del tiempo, elementos centrales en sistemas que afirman orientarse a la rehabilitación (Zedner, 2020). Ante ello resulta esencial que la adopción de herramientas de valoración del riesgo no suponga una renuncia a la valoración individual de la persona, al juicio de valoración del riesgo sobre la persona en concreto, sino que su uso sea concebido más bien como una mejora del juicio individual al permitir incorporar conocimiento objetivo general (también aplicable al particular) que puede haber sido obviado.

Es indiscutible que la progresiva algoritmización de la valoración del riesgo, el aumento de los datos utilizados y, sobre todo, la aparición de la Inteligencia Artificial podría intensificar estas dinámicas negativas y olvidar el carácter de "apoyo" que debería tener la parte "algorítmica" de la valoración del riesgo. Al facilitar el procesamiento sistemático de grandes volúmenes de información y la combinación simultánea de múltiples variables, los sistemas algorítmicos refuerzan la centralidad de las puntuaciones de riesgo como descriptores aparentemente objetivos del individuo. Esta evolución no supone necesariamente la automatización de la decisión ni la eliminación del juicio profesional, pero sí una reconfiguración de los equilibrios del proceso decisional, al desplazar capacidad cognitiva y simbólica hacia el instrumento técnico, que pasa a estructurar el marco dentro del cual se ejerce la valoración profesional. De este modo, los profesionales penitenciarios podrían dejar de ser evaluadores principales para convertirse, en muchos casos, en intérpretes o validadores de resultados producidos por sistemas formalizados, lo que altera los márgenes efectivos de discrecionalidad y responsabilidad. Desde la perspectiva de las personas privadas de libertad, esta transformación suele traducirse en un aumento de la vulnerabilidad. Las herramientas de valoración del riesgo operan sobre datos históricos producidos por las propias instituciones penales y penitenciarias, datos que reflejan desigualdades estructurales preexistentes en términos sociales, económicos o culturales. La literatura ha mostrado que los algoritmos utilizados en el sistema de justicia penal tienden a reproducir los sesgos presentes en los datos con los que se construyen o validan, incluso cuando no incorporan explícitamente variables tradicionalmente consideradas sensibles (Angwin et al., 2016; Jefferson, 2020). En contextos penitenciarios caracterizados por la sobrerrepresentación de colectivos socialmente desfavorecidos, esta dinámica adquiere una relevancia normativa particular, al interactuar con el principio de igualdad y con la exigencia de individualización del tratamiento.

El reconocimiento de los límites estructurales de las herramientas de valoración del riesgo y de los riesgos asociados al uso de la inteligencia arti-

ficial en este ámbito nos debería obligar, por tanto, a ser enormemente cautos en su uso y, como sostendremos más adelante en las conclusiones, optar por andar antes de empezar a correr. Lo cual no significa que el uso de IA no pueda tener un rol en lo relativo a la valoración del riesgo en el ámbito penitenciario. Diversos autores han subrayado que el potencial normativamente más defendible de la IA en contextos sensibles reside en su utilización como herramienta de auditoría, supervisión y rendición de cuentas, más que como instrumento de automatización decisoria (Kroll et al., 2017; Binns, 2018; Miró Llinares, 2020). En este sentido, la IA puede desempeñar un papel relevante en la evaluación crítica de las propias herramientas actuariales, permitiendo precisamente identificar los errores de sesgos o discriminación y poner de relieve los límites epistemológicos de las herramientas de valoración del riesgo que ya estructuran la toma de decisiones penitenciarias (Angwin et al., 2016; Chouldechova & Lum, 2020). Y es que la formalización propia de estos instrumentos abre la posibilidad de identificar, auditar y analizar de manera sistemática dichas fuentes de sesgo, siempre que existan mecanismos adecuados de transparencia, validación periódica y control institucional.

4. ANDAR (EN LA DIRECCIÓN ADECUADA) ANTES QUE CORRER O QUEDARSE INMÓVIL

La discusión desarrollada a lo largo de este trabajo nos sitúa en un momento especialmente delicado en el que la presión por incorporar tecnologías digitales, algoritmos y sistemas de inteligencia artificial en el ámbito penitenciario se combina con una narrativa de modernización casi inevitable, impulsada por un cierto determinismo tecnológico que desplaza el foco desde la pregunta por qué *problemas queremos resolver* hacia la simple cuestión de qué *tecnología está disponible*. En un ecosistema institucional en el que la digitalización avanza con rapidez y donde se adoptan herramientas bajo la promesa de eficiencia, objetividad o neutralidad técnica, la tentación de "correr", de implementar sin evaluar, de automatizar sin comprender, es grande. Lo que aquí propugnamos es lo contrario: que antes de decidir si implementar una tecnología, si cambiar un procedimiento debido a la aparición de un avance tecnológico, valoremos de manera reflexiva, apoyándonos en lo empírico pero teniendo en cuenta lo normativo, lo que su incorporación supondría; qué transforma, qué desplaza, qué riesgos introduce y, sobre todo, en qué modelo de castigo y de gestión penitenciaria se inserta.

Esta exigencia de prudencia explica por qué somos especialmente cautos a la hora de recomendar el uso o no de tecnologías como las herramientas de valoración del riesgo o, aún más, de sistemas de IA aplicados a las mismas. Y

está también en línea con lo señalado por la Recomendación del Consejo de Europa sobre el uso de IA en prisiones CM/Rec(2024)5, que en sus artículos 4 y 8 subraya que el uso de la inteligencia artificial y de tecnologías digitales en los servicios penitenciarios y de probation solo resulta legítimo cuando contribuye efectivamente a la rehabilitación y reintegración, no sustituye al personal profesional y se emplea de manera necesaria y proporcionada, evitando impactos negativos sobre los derechos fundamentales y la dignidad de las personas privadas de libertad. Porque la cuestión no es únicamente si estas herramientas funcionan, o si mejoran ligeramente las predicciones; la cuestión es si su incorporación resulta coherente con los fines constitucionales de la ejecución penal, si fortalece o debilita los derechos fundamentales de quienes viven en prisión y si convierte el sistema penitenciario en un espacio más humano o más automatizado, más reflexivo o más opaco. Y lo primero que cabe afirmar es que carece de sentido pensar en IA si antes no hemos pensado, seriamente, en las HVR. La IA aplicada a la predicción del riesgo no es una ruptura absoluta, sino una extensión intensificada de la lógica actuarial: depende de datos estructurados, se apoya en correlaciones y pretende aumentar la eficiencia del mismo tipo de decisiones que ya estructuran las herramientas tradicionales. Pero mientras se sigue debatiendo el alcance y las implicaciones del uso de herramientas de valoración del riesgo, pretender avanzar hacia el uso de sistemas de IA para sustituir a las clásicas herramientas de valoración del riesgo, siendo como son mucho más opacos y, por tanto, más difíciles de auditar, no solo nos parece prematuro, sino potencialmente peligroso.

Ahora bien, ¿debemos "empezar a andar" con las herramientas de valoración del riesgo? Como hemos visto, las críticas a las HVR suelen confundirse con reproches sobre aquello que no son: no son sistemas de inteligencia artificial opacos que ocultan las razones de la decisión (o no tienen por qué serlo), tampoco son garantías contra los sesgos estructurales del sistema penal. Otras críticas apuntan a límites reales: su diseño puede reflejar opciones de política criminal discutibles, pueden influir en exceso en el juicio profesional o pueden reforzar concepciones estáticas del riesgo. Pero lo que aquí hemos defendido es que cualquier tecnología debe ser evaluada en relación con aquello a lo que pretende sustituir. Y lo cierto es que, conforme a las evidencias disponibles, las HVR cuando están bien construidas, bien aplicadas y bien revisadas presentan mayores niveles de consistencia y capacidad predictiva que el juicio clínico no estructurado y, además, aportan transparencia y trazabilidad a decisiones que antes podían descansar en intuiciones difícilmente reconstruibles. También generan riesgos propios: pueden desplazar de manera indebida la autonomía del profesional, favorecer la lectura mecanicista de sus puntuaciones o consolidar visiones

del castigo orientadas al control más que a la rehabilitación. Pero rechazar de plano las capacidades de conocimiento que aportan estas herramientas tampoco parece razonable, sobre todo si, como ocurre con el juicio clínico no estructurado, las alternativas presentan problemas similares o incluso más intensos de opacidad, sesgo y variabilidad.

La cuestión decisiva, por tanto, no es si usar herramientas de valoración del riesgo, sino cómo y *para qué* hacerlo. Las HVR solo adquieren sentido dentro de un marco en el que la rehabilitación sea un fin real, no retórico, del sistema penitenciario; dentro de un diseño que asegure muestras adecuadas, procesos de validación periódicos y mecanismos de transparencia que permitan a las personas internas comprender, impugnar y defenderse frente a las valoraciones que les afectan. En ese contexto, pueden constituir un instrumento útil, siempre subordinado al juicio profesional y siempre integrado en un proceso deliberativo que reconozca la evolución, la agency y la capacidad de cambio de las personas.

Con la IA el panorama cambia significativamente. La IA introduce niveles de opacidad, correlacionalidad extrema y dificultades de explicabilidad que resultan incompatibles, al menos por ahora, con decisiones que afectan de manera directa a la libertad personal o a las expectativas de vida de quienes están en prisión. En un ámbito donde los errores no son meros "fallos del sistema", sino afectaciones reales a derechos fundamentales, delegar decisiones relevantes en modelos no explicables es una apuesta que un Estado de derecho no puede permitirse. De ahí que, a corto plazo, no sea recomendable su uso para decisiones clasificatorias, de progresión o de acceso a beneficios. Pero ello no implica renunciar a la IA por completo. Si algo muestra el análisis realizado es que la IA puede ser extraordinariamente útil *en sentido inverso*: para auditar las herramientas existentes, para revelar opacidades, para detectar patrones discriminatorios, para mostrar los límites y sesgos de las HVR y para fortalecer, en definitiva, los mecanismos de control democrático sobre la toma de decisiones. Ese uso metaevaluativo, orientado a la garantía y no a la automatización, podría constituir una vía razonable de integración de la IA en el ámbito penitenciario, evitando los riesgos más intensos sin renunciar a su potencial analítico.

En suma, quizá la clave esté en comprender que la digitalización, la algoritmización y la IA podrían ser sinónimos de progreso, pero solo cuando se usen en la dirección adecuada. El sistema penitenciario no necesita correr sin rumbo y abrazar tecnologías sin rumbo para acercarse a la modernidad. Tampoco, a nuestro parecer, debiera permitirse aferrarse a lo que ya tiene por puro miedo a lo que vendrá si es consciente de que las técnicas actuales

producen más perjuicios que las emergentes o se encuadran menos en los fines perseguidos. Es necesario seguir avanzando, aunque sea al paso, pero hacerlo con una dirección bien elegida. Solo cuando los fines rehabilitadores y reintegradores orienten realmente las prácticas penitenciarias y la decisión de adoptar nuevas herramientas, cuando los riesgos que las mismas traen consigo estén bien definidos y los mecanismos de supervisión para mitigarlos sean sólidos, tendrá sentido decidir las concretas tecnologías, algoritmos e IA, que pueden ayudarnos en ese camino. Avanzar con prudencia, aplicar los mejores conocimientos científicos y hacerlo orientándolo a los objetivos constitucionalmente establecidos para el castigo penal, debiera ser la única forma de actuar en un ámbito donde cada decisión, cada acción y también cada omisión, afecta a los derechos y a la dignidad de personas concretas.

5. REFERENCIAS

Adams, I. T. (2022). *Modeling Officer Perceptions of Body-worn Cameras: A National Survey.* (Doctoral dissertation, The University of Utah).

Alarie, B., Niblett, A., & Yoon, A. H. (2018). How artificial intelligence will affect the practice of law. *University of Toronto Law Journal*, *68*(Supplement 1), 106-124.

Andrés-Pueyo, A., & Echeburúa, E. (2010). Valoración del riesgo de violencia: instrumentos disponibles e indicaciones de aplicación. *Psicothema, 22*(3), 403-409.

Andrés-Pueyo, A., & Redondo, S. (2007). Predicción de la violencia: entre la peligrosidad y la valoración del riesgo de violencia. Papeles del psicólogo, 28(3), 157-173.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's *software* used across the country to predict future criminals. And it's biased against blacks. *ProPublica.* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Brandariz García, J. A. (2016). *El modelo gerencial-actuarial de penalidad. Eficiencia, riesgo y sistema penal*. Madrid: Dykinson.

Brennen, J. S., & Kreiss, D. (2016). *Digitalization.* En K. B. Jensen, R. T. Craig, J. Pooley, & E. W. Rothenbuhler (Eds.), The international encyclopedia of communication theory and philosophy. Wiley-Blackwell.

Chelioudakis, E. (2020). Risk assessment tools in criminal justice: Is there a need for such tools in Europe and would their use comply with European data protection law? *ANU Journal of Law and Technology, 1*(2), 72-96.

Chouldechova, A., & Lum, K. (2020). *The present and future of AI in pre-trial risk assessment instruments.* MacArthur Foundation Safety and Justice Challenge. https://safetyandjusticechallenge.org/resources/the-present-and-future-of-ai-in-pre-trial-risk-assessment-instruments/

Consejo de Europa, Comité de Ministros. (2024). *Recommendation CM/Rec(2024)5 of the Committee of Ministers to member States regarding the ethical and organisational aspects of the use of artificial intelligence and related digital technologies by prison and probation services*. Council of Europe. https://search.coe.int/cm?i=0900001680b1d0e4

Dodd, S., Antrobus, E., & Sydes, M. (2020). Cameras in corrections: Exploring the views of correctional officers on the introduction of body-worn cameras in prisons. *Criminal Justice and Behavior, 47*(9), 1190-1208.

Dou, L., & Dou, X. (2025). Towards just AI: Challenges and solution framework for algorithmic discrimination in judicial system. *International Journal of Digital Law and Governance, 2(1), 39-81.*

Entcheva, K., & Mazilescu, I. (2024). Artificial intelligence and digitalisation of judicial cooperation: the main provisions in recent EU legislation. *Eucrim: The European Criminal Law Associations' Forum (*3), 202-205.

Fedorczyk, F. (2024). Navigating the dichotomy of smart prisons: between surveillance and rehabilitation. *Law, Innovation and Technology*, *16*(1), 243-260.

Feeley, M. M., & Simon, J. (1992). The new penology: Notes on the emerging strategy of corrections and its implications. *Criminology, 30*(4), 449-474.

Finkel, M., & Bertram, W. (2019). *More states are signing harmful "free prison tablet" contracts. Prison Policy Initiative*.

Gendreau, P., Little, T., y Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology*, 34(4).

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. Psychological Assessment, 12(1), 19-30.

Güerri Ferrández, C., Martí Barrachina, M., & Pedrosa, A. (2021). Abriendo ventanas virtuales en los muros de la prisión. *IDP: Revista de Internet, Derecho y Política = Revista d'Internet, Dret i Política* (32), 5.

Hannah-Moffat, K. (2013). Actuarial Sentencing: An "Unsettled" Proposition. *Justice Quarterly, 30*(2), 270-296.

Hannah-Moffat, K. (2019). Algorithmic risk governance: Big data analytics, race and information activism in criminal justice debates. *Theoretical Criminology, 23*(4), 453-470.

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: a meta-analysis of 118 prediction studies. *Psychological assessment*, *21*(1), 1.

Hanson, R. K., y Bussiere, M. T. (1998). Predicting relapse: a meta-analysis of sexual offender recidivism studies. *Journal of consulting and clinical psychology*, 66(2).

Harcourt, B. E. (2007). *From the asylum to the prison: Rethinking the incarceration revolution. Part II: State-level analysis* (University of Chicago Public Law & Legal Theory Working Paper No. 155). University of Chicago.

Hildebrandt, M. (2015). *Smart technologies and the End(s) of Law*. Edward Elgar Publishing.

Järveläinen, E., & Rantanen, T. (2021). Incarcerated people's challenges for digital inclusion in Finnish prisons. *Nordic Journal of Criminology, 22*(2), 240-259.

Jefferson, B. (2020). *Digitize and punish: Racial criminalization in the digital age*. University of Minnesota Press.

Jewkes, Y., & Reisdorf, B. C. (2016). A brave new world: The problems and opportunities presented by new media technologies in prisons. *Criminology & Criminal Justice*, *16*(5), 534-551.

Kaun, A., & Stiernstedt, F. (2019). Doing time/time done: Exploring the temporalities of datafication in the smart prison. In *Mediated time: Perspectives on time in a digital age* (pp. 129-147). Cham: Springer International Publishing.

Kehl, D. L., & Kessler, S. A. (2017). Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041.

Kim, L. (2012, abril). Meet South Korea's new robotic prison guards. *Digital Trends.* https://www.digitaltrends.com/cool-tech/meet-south-koreas-new-robotic-prison-guards/

Krikorian, J., & Coye, S. (2019). On digital resources in detention (Issue April). Retrieved from http://www.prisonlitigation.org/wp-content/uploads/2019/07/Report-DigitalResources-in-EU.pdf

Krištofík, A. (2025). Bias in AI (supported) decision making: Old problems, new technologies. *International Journal for Court Administration, 16*(1), 3-27.

Leighton, P. (2014). «A Model Prison for the Next 50 Years": The High-Tech, Public-Private Shimane Asahi Rehabilitation Center. *Justice Policy Journal*, *11*(1).

Lindström B., & Puolakka P. (2020). Smart prison: The preliminary development process of digital self-services in Finnish prisons. https://icpa.org/library/smart-prison-the-preliminary-development-process-of-digital-self-services-in-finnish-prisons/

López Lorca, B. (2023). La digitalización de las prisiones y el uso de la inteligencia artificial: Marcadores de última generación para la normalización del entorno penitenciario y la redefinición del proceso de resocialización. *En F. Miró (Coord.),* Digitalización y algoritmización de la justicia. *IDP. Revista de Internet, Derecho y Política* (39).

McKay, C. (2022). The carceral automaton: Digital prisons and technologies of detention. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 100-119.

Miró Llinares, F (2020). Predictive policing: utopia or dystopia? On attitudes towards the use of big data algorithms for law enforcement. IDP: revista de Internet, derecho y política= revista d'Internet, dret i política (30), 5.

Miró Llinares, F, & Castro Toledo, F. J. (2022). ¿Correlación no implica causalidad? El valor de las predicciones algorítmicas en el sistema penal a propósito del debate epistemológico sobre "el fin de la teoría". In *Derecho penal y comportamiento humano. Avances desde la neurociencia y la inteligencia artificial* (pp. 507-530). Tirant lo Blanch.

Miró Llinares, F. (2022). Inteligencia artificial, delito y control penal: nuevas reflexiones y algunas predicciones sobre su impacto en el derecho y la justicia penal. *El Cronista del Estado Social y Democrático de Derecho* (100), 174-183.

Miró Llinares, F. (2023). ¿Sueñan las cárceles con robots celadores? Aproximación al uso de la inteligencia artificial en prisiones. En Comisión de

Estudios de ATIP (Ed.), XI Jornadas de ATIP Almagro 2022 (pp. 71-88). Asociación de Técnicos de Instituciones Penitenciarias.

Miró Llinares, F. (2025). Derecho penal y desafíos de la inteligencia artificial (en el contexto del nuevo marco regulatorio europeo). En M. E. Casas Baamonde (Dir.) & D. Pérez del Prado (Coord.), *Derecho y tecnologías* (pp. 177-212). Editorial Centro de Estudios Ramón Areces.

Mufarreh, A., Waitkus, J., & Booker, T. A. (2022). Prison official perceptions of technology in prison. *Punishment & Society*, *24*(3), 410-432.

Peterson, B., Kizzort, M., Kim, K., & Shukla, R. (2021). Prison contraband: prevalence, impacts, and interdiction strategies. *Corrections*, 1-18.

Puolakka, P. (2020). Smart Prison. *EuroPris*. P. https://www.europris.org/wp-content/uploads/2020/10/Pia-Puolakka-Smart-Prisons-Finland.pdf

Puolakka, P. (2022). Implementing a Smart Prison in Finland. Advancing Corrections Journal, 14. ICPA.

Rizer, A., & Watney, C. (2018). Artificial intelligence can make our jail system more efficient, equitable, and just. *Texas Review of Law & Politics, 23*(1), 181-227.

Schwab, K. (2016). *La cuarta revolución industrial.* Debate.

Sydes, M., Dodd, S., & Antrobus, E. (2022). Body cameras behind bars: Exploring correctional officers' feelings of safety with body-worn cameras. *Criminology & Criminal Justice*, *22*(2), 323-342.

Thaler, O., Barkworth, J., & Howard, M. (2022). Implementing digital technologies in prisons: A qualitative study of inmate experiences. *Sydney: Corrections Research Evaluation and Statistics.*

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society, 4*(2), 1-17.

Završnik, A. (2021). Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of criminology, 18*(5), 623-642.

Zedner, L. (2020). Pre-crime and post-criminology? The future of security and justice. *The British Journal of Criminology, 60*(5), 1089-1106.

Capítulo 2

Algoritmos de reincidencia: ¿Herramienta útil para Euskadi? Revisión sistemática sobre variables, modelos e interpretación en el ámbito penitenciario

Ane Viana Salinas

Instituto Vasco de Criminología (IVAC-KREI)

Leyre Padilla López

Instituto Vasco de Criminología (IVAC-KREI)

SUMARIO: 1. INTRODUCCIÓN: CONTEXTO Y OBJETIVOS. 2. ANÁLISIS DE LA LITERATURA SOBRE ALGORITMOS DE REINCIDENCIA. 3. EVALUACIÓN CRÍTICA DE MODELOS, VARIABLES Y RESULTADOS. 4. DISCUSIÓN: ESTADO DE LA CUESTIÓN ACTUAL E IMPLICACIONES PARA EUSKADI. 5. CONCLUSIONES. 6. REFERENCIAS.

1. INTRODUCCIÓN: CONTEXTO Y OBJETIVOS

Hoy en día, la inteligencia artificial y los algoritmos ya no son promesas del futuro: están presentes día a día en todo, desde recomendaciones en plataformas digitales hasta asistentes virtuales y sistemas automatizados en salud, banca o educación (Perry, McInnis, Price, Smith, & Hollywood, 2013). Nos movemos en una sociedad donde tomar decisiones basadas en datos se ha vuelto una práctica habitual.

El sistema penal y el sistema penitenciario no han quedado al margen de la tendencia hacia la utilización de algoritmos (Miró-LLinares, 2018; Fanega et al., 2024; Portela et al., 2022). En el ámbito de justicia penal, los algoritmos

predictivos y de evaluación de riesgo (RAI) se han utilizado para predecir dónde es más probable que se produzcan delitos, identificar a personas con mayor probabilidad de cometer delitos violentos, prever quiénes podrían no comparecer a sus audiencias judiciales y estimar la probabilidad de reincidencia futura (Dressel & Farid, 2018; Hamilton, 2021).

Cada vez más, los sistemas penitenciarios emplean herramientas de predicción para evaluar el riesgo de reincidencia de las personas privadas de libertad, con el objetivo de apoyar decisiones sobre medidas de seguridad, programas de reinserción así como la concesión de permisos de salida de los centros penitenciarios (Férez-Mangas, y Andrés-Pueyo, 2018).

Se considera que los permisos de salida son una herramienta imprescindible en la intervención penitenciaria y forman parte del programa individualizado de tratamiento para cada interno (Férez-Mangas, y Andrés-Pueyo, 2015). En otras palabras, estos permisos constituyen un pilar o llave en los procesos de reinserción social (Martínez Escamilla, 2002).

Diversos estudios han mostrado que la toma de decisiones sobre la conducta futura de los internos no debe basarse únicamente en la experiencia y el juicio clínico de los profesionales, sino que requiere la incorporación de herramientas y sistemas de valoración más precisos y contrastados. Por ello, los instrumentos de valoración del riesgo de quebrantamiento de permisos de salida resultan esenciales para estimar la probabilidad de incumplimiento o no retorno (Férez-Mangas y Andrés-Pueyo, 2015).

Sin embargo, la adopción de estos sistemas plantea desafíos significativos y, surgen varias cuestiones críticas: ¿Qué tan transparente es la toma de decisiones automatizada? ¿Quién se hace responsable cuando un sistema falla? ¿Podría estar ampliando desigualdades sociales que ya existían? Una de las cuestiones más debatidas es la capacidad predictiva de estos instrumentos; esto es, con qué grado de acierto efectúan los pronósticos sobre la conducta futura de las personas. En otras palabras, qué probabilidad hay de equivocarse cuando los operadores jurídicos toman decisiones basadas en estos pronósticos (Martínez Garay, 2016).

Aunque los algoritmos de predicción ofrecen beneficios claros en términos de eficiencia, estandarización y soporte a la toma de decisiones, su fiabilidad, equidad y transparencia no están garantizadas. Existen riesgos de sesgos, errores de predicción y opacidad en la interpretación de resultados, lo que puede desplazar la responsabilidad humana y, bajo la apariencia de neutralidad técnica, reforzar injusticias. Por ello, según Skeem y Lowenkamp

(2020), estas herramientas deben diseñarse, evaluarse y regularse con rigor, bajo criterios claros y transparentes.

Este debate adquiere una relevancia particular en la Comunidad Autónoma Vasca (CAV, en adelante), que desde 2021 ha asumido la gestión de las instituciones penitenciarias (Real Decreto 474/2021). La transferencia de competencias abre la oportunidad de reflexionar sobre qué herramientas y enfoques resultan más adecuados para la población penitenciaria vasca, considerando sus especificidades sociales, culturales y normativas.

En definitiva, este trabajo se plantea dos preguntas centrales: ¿qué tan fiables, justos y útiles son los algoritmos de evaluación de riesgo de reincidencia? ¿Cuáles son los riesgos asociados a su uso? El objetivo principal es revisar sistemáticamente la literatura existente sobre las diferentes herramientas de evaluación de riesgo de reincidencia, con énfasis en las variables que emplean, su interpretación, los posibles sesgos y los debates ético-legales asociados. A partir de ello, se busca valorar la potencial utilidad de una herramienta de predicción de la reincidencia en el ámbito penitenciario de la CAV, asegurando siempre la primacía de criterios éticos y profesionales en la toma de decisiones.

2. ANÁLISIS DE LA LITERATURA SOBRE ALGORITMOS DE REINCIDENCIA

La reincidencia delictiva, en sentido amplio, abarca toda la sucesión de hechos delictivos cometidos con posterioridad a la condena derivada del primero de ellos (Gobierno de España, Ministerio del Interior, 2022). Por su parte, el Código Penal actual establece en su art. 22 cuáles son las circunstancias agravantes de la responsabilidad criminal, fijando en el apartado octavo de dicho precepto la reincidencia.

Hay reincidencia cuando, al delinquir, el culpable haya sido condenado ejecutoriamente por un delito comprendido en el mismo título de este Código, siempre que sea de la misma naturaleza.

En el ámbito penitenciario, la reincidencia constituye uno de los principales desafíos, ya que condiciona la gestión de permisos de salida, programas de reinserción y medidas de seguridad (Skeem & Lowenkamp, 2020; Medina, 2013). Las intervenciones rehabilitadoras buscan reducir la probabilidad de nuevos delitos, pero los resultados han sido limitados, dada la complejidad del comportamiento delictivo y su carácter multifactorial. Solo algunas características generales de inteligencia y personalidad se han aso-

ciado de manera consistente con la *conducta delicti*va, sin que hasta la fecha sea posible definir un perfil único de delincuente (Chico, 1997; Pelechano, 2008; Sánchez-Teruel y Robles-Bello, 2013).

Diversos estudios han identificado factores de riesgo asociados a la reincidencia, incluyendo cognición antisocial, historial delictivo, redes de apoyo prodelictivas y ciertos rasgos de personalidad (Andrews y Bonta, 2010; Ortega, García y Frías, 2014; Singh et al., 2011; Ortega-Campos, García-García, De la Fuente-Sánchez y Zaldívar-Basurto, 2020).

Estos factores de riesgo plantean un desafío práctico: determinar de manera fiable qué individuos presentan mayor probabilidad de reincidir. La necesidad de responder a esta cuestión ha impulsado el desarrollo de métodos de predicción de la reincidencia, que han evolucionado desde evaluaciones cualitativas basadas en la experiencia profesional hasta algoritmos y herramientas estructuradas utilizadas en la actualidad.

La valoración del riesgo de reincidencia se remonta a la década de 1960 (Borum, 2000). En su origen, la predicción de la violencia respondía a una visión marcadamente clínica y a una valoración no estructurada basada en la libre interpretación de la información del caso que el/la profesional realizaba sobre la peligrosidad del sujeto (León y Bartolomé, 2024). Los jueces y el resto de operadores jurídicos recurrían a los informes de psiquiatras, médicos forenses y especialistas de las juntas de tratamiento penitenciarias, quienes emitían sus conclusiones a partir de la libre interpretación de la información disponible y de su experiencia profesional individual, sin el apoyo de herramientas estandarizadas (Martínez Garay y Montes Suay, 2018).

Esta modalidad tradicional de evaluación de la peligrosidad, basada en el juicio no estructurado (clínico) (Andrés-Pueyo & Echeburúa, 2010), se sostuvo en la práctica incluso cuando ya se reconocían la ambigüedad conceptual y las limitaciones de la «peligrosidad» como predictor de la conducta violenta. Su aplicación respondía, en gran medida, a las exigencias legales en determinados contextos y a la inexistencia, en aquel momento, de procedimientos alternativos que ofrecieran una mayor validez y fiabilidad (Morales, 2021).

Sin embargo, la necesidad de dar respuesta a estas demandas legales impulsó la investigación sobre los factores asociados a un mayor riesgo de violencia y la proliferación de instrumentos más estructurados. En este contexto, el concepto de "peligrosidad" comenzó a ser sustituido por el de valoración del riesgo de violencia *(violence risk assessment)* (Andrés-Pueyo y Redondo, 2007). Mientras que la peligrosidad se concebía como una cualidad estática e inherente al individuo, presente de manera permanente desde la comisión del

primer delito, el enfoque basado en el riesgo se centra en factores personales y contextuales dinámicos que aumentan o disminuyen la probabilidad de reincidencia (Morales, 2021). En consecuencia, a partir de la década de 1970, el juicio clínico comenzó a dar paso al desarrollo de lo que Bonta y Andrews (2007) denominaron instrumentos de evaluación de riesgo de segunda generación. Estos instrumentos, desarrollados sobre la base de análisis estático, superaban de manera consistente al juicio clínico en la predicción de la reincidencia (Brennan, Dieterich y Ehret, 2009; Férez-Mangas y Andrés-Pueyo, 2015).

Los instrumentos de segunda generación se clasifican en dos tipos principales: los instrumentos actuariales y los de juicio clínico estructurado. Ambos se fundamentan en la observación empírica de grupos de sujetos y en la cuantificación y combinación estadística de factores de riesgo —y, en ocasiones, también de protección— que han mostrado una asociación significativa con la aparición de conductas violentas o delictivas (Martínez Garay y Montes Suay, 2018).

En primer lugar, surgieron los instrumentos actuariales, basados principalmente en factores de riesgo estáticos como el historial criminal, la edad o el tipo de delito (Rodríguez, López y Andrés-Pueyo, 2002). Estos instrumentos permitieron construir predicciones más estandarizadas y objetivas, reduciendo el margen de arbitrariedad individual. En otras palabras, los instrumentos actuariales generan una puntuación global mediante un algoritmo que combina los factores de riesgo incluidos, ofreciendo así una estimación automática y reproducible. No obstante, Martínez (2014), tras realizar una exhaustiva revisión, advirtió sobre los peligros de confiar excesivamente en las escalas actuariales, especialmente en casos que presenten características atípicas o factores contextuales no contemplados por el instrumento.

Aunque los instrumentos actuariales mejoraron la objetividad, su dependencia de factores estáticos y la limitada flexibilidad motivaron el desarrollo de los sistemas de juicio profesional estructurado (Pérez, Redondo, Martínez, García y Andrés-Pueyo, 2008). Estas herramientas funcionan como guías que señalan los factores relevantes y orientan su valoración, pero dejan la decisión final al profesional, quien pondera el peso de cada factor e incluso puede incorporar otros no contemplados por la herramienta cuando los considera determinantes para el caso concreto (Loinaz, 2017). Este enfoque combina la sistematicidad de los instrumentos basados en evidencia con la flexibilidad del juicio profesional, sentando las bases para la incorporación posterior de factores dinámicos orientados a la intervención, característicos de la tercera generación de instrumentos de evaluación de riesgo.

El avance científico permitió incorporar factores de riesgo dinámicos —las necesidades criminógenas—, esenciales no solo para mejorar la precisión pronóstica, sino también para orientar las intervenciones penitenciarias hacia la reducción efectiva del riesgo mediante programas de rehabilitación adaptados a cada individuo (Férez-Mangas y Andrés-Pueyo, 2015). Así surgieron los instrumentos de tercera generación, que incorporaban estos factores dinámicos junto con los estáticos, permitiendo no solo predecir la reincidencia, sino también identificar áreas específicas de intervención para guiar programas de rehabilitación personalizados y reducir el riesgo de reincidencia de manera más efectiva (Duwe & Kim, 2015).

Posteriormente, surgieron los instrumentos de cuarta generación, que amplían esta perspectiva al integrar la evaluación del riesgo con la planificación y el monitoreo sistemático de las intervenciones. Más allá de identificar factores de riesgo y necesidades, estos instrumentos permiten diseñar planes de tratamiento individualizados, evaluar continuamente el progreso y ajustar las estrategias de intervención según la evolución del riesgo y las necesidades del sujeto, buscando maximizar la efectividad de la rehabilitación y reducir la reincidencia de manera más integral y adaptativa (Brennan et al., 2009; Duwe & Kim, 2015).

Actualmente, existen más de 100 herramientas de predicción de riesgo, la mayoría orientadas a contextos violentos o conductas delictivas específicas. Estas herramientas ponderan factores de riesgo estáticos (por ejemplo, antecedentes penales, edad, tipo de delito) y dinámicos (por ejemplo, consumo de drogas, situación laboral, entorno social), cuya combinación se correlaciona con la probabilidad de reincidencia (Bravo, 2023).

A nivel internacional, se han desarrollado múltiples instrumentos para la evaluación del riesgo en contextos penitenciarios. En los Estados Unidos, la situación es particularmente heterogénea, ya que cada estado posee atribuciones exclusivas en materia de política de ejecución penal, lo que se traduce en la utilización de distintos protocolos e instrumentos de valoración del riesgo (Férez-Mangas y Andrés-Pueyo, 2015). Entre los más reconocidos se encuentran el *Level of Service Inventory-Revised* (LSI-R; Andrews & Bonta, 1995), el *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS; Brennan, Dieterich, & Ehret, 2009; Brennan, Dieterich, & Oliver, 2007) y el *Historical Clinical Risk Management-20* (HCR-20; Douglas et al., 2014). También se han desarrollado otros modelos, como el OASys en Inglaterra y Gales (Howard et al., 2006; Howard & Dixon, 2012) o el OXREC, propuesto por la Universidad de Oxford (Fazel et al., 2016).

A nivel nacional, los instrumentos de valoración del riesgo de quebrantamiento utilizados en los centros penitenciarios varían según su dependencia administrativa. Se distinguen principalmente dos contextos: los centros dependientes de la Administración General del Estado y los que dependen de la Comunidad Autónoma de Cataluña (Férez-Mangas y Andrés-Pueyo, 2015;2018).

En los centros penitenciarios dependientes de la Administración General del Estado, las decisiones sobre permisos de salida se apoyan en dos instrumentos de evaluación: la Tabla de Variables de Riesgo (TVR) y la Tabla de Concurrencia de Circunstancias Peculiares (CCP) (Férez-Mangas & Andrés-Pueyo, 2015; 2018).

La TVR es un instrumento actuarial que fue introducido por la Instrucción 1/1995, de 10 de enero, sobre el estudio y tramitación de los permisos de salida; aunque no fue hasta la entrada en vigor de la Instrucción 22/1996, de 16 de diciembre, tras la aprobación del Reglamento Penitenciario de 1996, cuando se detallaron realmente las variables que integraban los citados instrumentos de valoración del riesgo de quebrantamiento (Férez-Mangas & Andrés-Pueyo, 2018; Fanega et al., 2024; Chaves, 2025).

La TVR consiste en un protocolo actuarial compuesto por diez factores de riesgo, cada uno puntuado de 0 (ausencia de riesgo) a 3 (máximo riesgo), según la naturaleza y presencia del factor. Las variables que la conforman son: extranjería, drogodependencia, profesionalidad delictiva, reincidencia, quebrantamientos anteriores, artículo 10 de la LOGP, ausencia de permisos, deficiencia convivencial, lejanía y presiones internas a las que puede estar sometido el interno (Fanega et al., 2024; Chaves, 2025).

Junto a la TVR, la Administración Penitenciaria utiliza la Tabla de Concurrencia de Circunstancias Peculiares (CCP). Mientras que la TVR ofrece un protocolo actuarial basado en diez factores de riesgo puntuables, la CCP recoge las denominadas circunstancias peculiares, es decir, variables adicionales que deben considerarse al decidir la concesión de un permiso de salida (Chaves, 2025).

Desde 2009, en las cárceles bajo la competencia penitenciaria de Cataluña —asumida por la administración autonómica en 1984— se aplica RisCanvi (Alemán, 2023). Según Castro (2019), se trata de la herramienta de valoración de riesgos más avanzada, refinada y consolidada de su tipo en España. Por ello, aunque su ámbito de aplicación se limite a Cataluña, su importancia en el contexto español es insoslayable.

El RisCanvi es un protocolo multi-escala de valoración del riesgo que permite pronosticar comportamientos futuros (Férez-Mangas & Andrés-Pueyo, 2018). Según el Manual de aplicación del protocolo (Subdirecció General de Programes de Rehabilitació i Sanitat del Departament de Justícia, 2019), se organiza en dos escalas: RisCanvi Screening y RisCanvi Completa, ambas integrando factores de riesgo estáticos y dinámicos, sin incluir factores de protección (Alemán, 2023). Una vez introducidos los factores de riesgo, el instrumento utiliza una combinación algorítmica para generar valores de riesgo, que actualmente se refieren a cinco tipos de comportamientos: violencia autodirigida, violencia intrainstitucional, reincidencia general, reincidencia violenta y quebrantamiento de condena (Andrés-Pueyo, 2016; Alemán, 2023).

En síntesis, la introducción de este tipo de herramientas se ha justificado por la necesidad de reducir la discrecionalidad y la subjetividad inherente al juicio clínico (Rueda Soriano & Navarro Blasco, 2019), hacia modelos estructurados sustentados por la evidencia empírica (Castro, 2019). La incorporación de herramientas estandarizadas ha respondido a la necesidad de reducir la discrecionalidad y aumentar la objetividad en la toma de decisiones, proporcionando criterios más consistentes y verificables (Alemán, 2023). Además, se considera que dan estabilidad al sistema y permiten tanto agilizar el trabajo de los y las profesionales como reforzar su seguridad en la toma de decisiones (Castro, 2019).

No obstante, el uso de estos sistemas plantea desafíos significativos. Si bien pueden mejorar la precisión, la solidez y la eficiencia en los procesos de toma de decisiones (Kleinberg et al., 2018), también existe el riesgo de que generen sesgos en los resultados y, en consecuencia, produzcan efectos discriminatorios (Angwin et al., 2016; Skeem & Lowenkamp, 2020; Alemán, 2023).

En resumen, el desafío actual consiste en equilibrar la eficacia predictiva de estos modelos con los principios de justicia, equidad y finalidad resocializadora de la pena. Este debate abre el camino al análisis de los sistemas algorítmicos de evaluación del riesgo, que serán objeto de estudio en el siguiente epígrafe, con el objetivo de valorar su eficacia y determinar si conviene aplicarlos en el sistema penitenciario vasco.

3. EVALUACIÓN CRÍTICA DE MODELOS, VARIABLES Y RESULTADOS

El empleo de algoritmos de predicción del riesgo de reincidencia ha transformado en los últimos años la manera en que las instituciones peni-

tenciarias abordan la evaluación de las personas privadas de libertad. Su uso promete mayor objetividad, homogeneidad y eficacia en la toma de decisiones, en contraste con el tradicional juicio clínico no estructurado que dominó durante décadas la práctica penitenciaria (Andrés-Pueyo y Echeburúa, 2010). Sin embargo, la literatura reciente advierte que la fiabilidad y la validez de estos sistemas distan de ser absolutas, y que su adopción implica dilemas éticos, jurídicos y técnicos que deben ser analizados con cautela.

En España, el RisCanvi se ha consolidado como el principal instrumento de valoración del riesgo en los centros penitenciarios catalanes. Diseñado con el objetivo de mejorar la precisión de las predicciones y reducir la discrecionalidad profesional, combina factores estáticos y dinámicos para estimar el riesgo de reincidencia general, violenta o de quebrantamiento de condena (Andrés-Pueyo, 2016; Alemán, 2023). Pese a sus avances metodológicos, las investigaciones empíricas sobre su capacidad predictiva han mostrado resultados moderados. Un estudio retrospectivo reveló un área bajo la curva (AUC) de 0,64 para la reincidencia violenta, lo que indica un nivel de discriminación solo ligeramente superior al azar (Nguyen, Arbach & Andrés-Pueyo, 2019; Martínez, 2014).

En un análisis prospectivo con 684 exreclusos, la herramienta alcanzó una sensibilidad del 77 % y una especificidad del 57 %, con un valor predictivo positivo del 17,94 %, lo que refleja un número considerable de falsos positivos (García & Rivera, 2016). Dicho de otro modo, muchas de las personas clasificadas como de alto riesgo no reincidieron realmente, lo que evidencia la dificultad de trasladar la probabilidad estadística al pronóstico individual. Este resultado se ve influido por la baja tasa general de reincidencia en Cataluña, situada entre el 16,5 % y el 29,4 %, lo que reduce la capacidad discriminativa del modelo (Capdevila et al., 2015; Luque et al., 2005).

Estas limitaciones no son exclusivas del contexto español. En Estados Unidos, el sistema COMPAS -Correctional Offender Management Profiling for Alternative Sanctions— se ha convertido en el paradigma de los algoritmos penitenciarios basados en grandes volúmenes de datos. No obstante, estudios críticos han demostrado que, lejos de eliminar los sesgos humanos, puede reproducir y amplificar desigualdades raciales y socioeconómicas. La investigación de ProPublica (Angwin, Larson, Mattu & Kirchner, 2016) reveló que el algoritmo tendía a sobrestimar el riesgo de reincidencia en personas negras y a subestimarlo en personas blancas, incluso cuando las tasas reales de reincidencia eran similares. Este hallazgo desató un debate internacional sobre la equidad algorítmica y la responsabilidad institucional,

abriendo interrogantes sobre la legitimidad del uso de sistemas automáticos en decisiones que afectan derechos fundamentales (Kleinberg, Ludwig, Mullainathan & Rambachan, 2018; Skeem & Lowenkamp, 2020).

Frente a estos modelos automatizados, instrumentos como el Level of Service Inventory-Revised (LSI-R) o el Historical Clinical Risk Management-20 (HCR-20) se apoyan en un enfoque más equilibrado, que combina la objetividad de los criterios actuariales con la flexibilidad del juicio clínico estructurado. El LSI-R, desarrollado por Andrews y Bonta (1995), incluye 54 factores —36 dinámicos y 18 estáticos— distribuidos en diez áreas relacionadas con la historia criminal, la educación, el empleo, la situación financiera, la familia, el entorno social, el ocio, el consumo de sustancias y las cogniciones antisociales. Este modelo, además de predecir la reincidencia general y violenta, permite orientar la intervención penitenciaria y diseñar programas personalizados de tratamiento, lo que lo diferencia de los algoritmos puramente estadísticos. El HCR-20, por su parte, ha demostrado una alta capacidad para evaluar el riesgo de violencia en contextos forenses y de salud mental, integrando datos empíricos con la interpretación profesional de los técnicos (Douglas et al., 2014).

La fiabilidad de cualquier instrumento de predicción depende en gran medida de la naturaleza de las variables que incorpora. Los factores estáticos, como el tipo de delito o el número de condenas previas, son fácilmente cuantificables, pero no modificables, por lo que resultan útiles para estimar la probabilidad base de reincidencia, aunque limitados para orientar el tratamiento. Los factores dinámicos, en cambio, permiten medir necesidades criminógenas susceptibles de cambio, tales como la actitud hacia la autoridad, el consumo de drogas, el empleo o el entorno familiar (Bonta & Andrews, 2007; Férez-Mangas & Andrés-Pueyo, 2015). Este enfoque dinámico, propio de los instrumentos de tercera y cuarta generación, sitúa la predicción del riesgo al servicio de la intervención y de la reinserción social, en lugar de concebirla únicamente como una herramienta clasificatoria.

No obstante, la literatura especializada advierte sobre la importancia de contextualizar las variables y evitar interpretaciones automáticas. Los algoritmos pueden otorgar un peso excesivo a ciertos factores socioeconómicos o demográficos, como el nivel de ingresos o la residencia en determinados barrios, que pueden reflejar desigualdades estructurales más que propensiones individuales al delito (Bravo Bolado, 2023). Este tipo de sesgos estructurales puede generar una injusticia algorítmica, donde la supuesta neutralidad técnica oculta discriminaciones históricas o culturales. Por

ello, la incorporación de factores de protección -como la participación en programas educativos, el apoyo familiar o la motivación para el cambio— resulta esencial para equilibrar el cálculo del riesgo y ofrecer una visión más completa del proceso de reinserción (Ortega-Campos, García-García, De la Fuente-Sánchez & Zaldívar-Basurto, 2020).

Más allá de su precisión estadística, la transparencia de los modelos es una condición indispensable para su legitimidad jurídica y social. En este sentido, la diferencia entre algoritmos abiertos y cerrados resulta determinante. Mientras que herramientas públicas como RisCanvi o LSI-R permiten conocer las variables y criterios que intervienen en la evaluación, modelos de propiedad privada como COMPAS funcionan mediante sistemas cerrados cuya lógica de decisión no es accesible ni a los operadores jurídicos ni a las personas evaluadas. Ello dificulta la posibilidad de impugnar o auditar los resultados, vulnerando el derecho a la defensa y el principio de motivación de las resoluciones (Angwin et al., 2016; Bravo Bolado, 2023). La literatura reciente subraya que la interpretabilidad -es decir, la capacidad de explicar cómo se genera un resultado— no es solo un requisito técnico, sino una garantía democrática y procesal (Scaria, Subramanian, George & Sengupta, 2024). En contextos judiciales, donde las decisiones deben poder justificarse y revisarse, la opacidad de un algoritmo se traduce en una forma de irresponsabilidad institucional.

El debate sobre la legitimidad del uso de algoritmos de riesgo se enmarca, además, en un contexto normativo cada vez más exigente. El Reglamento General de Protección de Datos (RGPD) reconoce el derecho de las personas a no ser objeto de decisiones automatizadas con efectos jurídicos significativos, salvo que existan garantías adecuadas de transparencia y supervisión humana. En esta misma línea, el futuro Artificial Intelligence Act de la Unión Europea clasifica los sistemas de inteligencia artificial utilizados en justicia penal como de "alto riesgo", imponiendo requisitos de trazabilidad, explicabilidad y control institucional. Desde esta perspectiva, los algoritmos de predicción penitenciaria deben considerarse instrumentos auxiliares, subordinados siempre al juicio profesional y al marco ético de la intervención (Skeem & Lowenkamp, 2020; Hamilton, 2021).

En la práctica cotidiana, los profesionales penitenciarios valoran la utilidad de estas herramientas como sistemas de apoyo estructurado, pero también señalan el peligro de una dependencia excesiva del resultado numérico. En Cataluña, el protocolo RisCanvi se aplica mediante dos escalas —*Screening* y *Completa*— que permiten estimar el riesgo de reincidencia y orientar

la planificación del tratamiento (Subdirecció General de Programes de Rehabilitació i Sanitat, 2019). Si bien este sistema ha mejorado la coherencia de las valoraciones, algunos estudios advierten que el peso excesivo del resultado final puede reducir la atención a factores cualitativos no contemplados en el instrumento, como la evolución conductual o el contexto social del interno (Castro, 2019; Alemán, 2023). De ahí que la combinación entre la evaluación técnica y la experiencia profesional resulte clave para mantener el equilibrio entre evidencia empírica y comprensión individual del caso (Loinaz, 2017).

A nivel internacional, los resultados de la investigación siguen siendo heterogéneos. Mientras algunos estudios confirman que los modelos actuariales mejoran la capacidad predictiva respecto al juicio clínico no estructurado (Brennan, Dieterich & Ehret, 2009), otros advierten que su precisión decrece cuando se aplican fuera del contexto para el que fueron desarrollados (Férez-Mangas & Andrés-Pueyo, 2015). La validez externa de estos instrumentos continúa siendo limitada, lo que evidencia la necesidad de evaluaciones independientes y continuas. Además, la calidad de los datos empleados para entrenar los algoritmos condiciona directamente sus resultados: si las bases de datos contienen sesgos históricos —por ejemplo, un sobrerregistro de ciertos colectivos—, el modelo tenderá a reproducir esas desigualdades (Angwin et al., 2016; Morales, 2021).

Los avances en técnicas de aprendizaje automático (*machine learning*) ofrecen nuevas posibilidades para la predicción del comportamiento delictivo. Estas herramientas pueden procesar grandes volúmenes de información y detectar patrones complejos, superando en algunos casos a los modelos estadísticos tradicionales (Duwe & Kim, 2015). Sin embargo, su opacidad y la dificultad de interpretar sus decisiones plantean importantes retos éticos y jurídicos. La tendencia hacia algoritmos cada vez más sofisticados puede incrementar la distancia entre los operadores jurídicos y la tecnología, dificultando el control institucional y ciudadano sobre las decisiones que afectan a la libertad de las personas (Bravo Bolado, 2023).

De cara al futuro, la investigación apunta hacia modelos híbridos que combinen las ventajas del análisis algorítmico con la flexibilidad y el juicio contextual del profesional. Estos modelos deberían incorporar dimensiones cualitativas, validar sus resultados en poblaciones locales y ser sometidos a auditorías periódicas. La inclusión de la perspectiva de género, la diversidad cultural y los principios de equidad social constituye un reto ineludible para evitar la reproducción de desigualdades dentro del sistema penitenciario (León & Bartolomé, 2024).

En el contexto vasco, la transferencia de competencias penitenciarias abre una oportunidad para reflexionar sobre qué tipo de herramientas resultan más adecuadas. La creación de un sistema propio permitiría adaptar los modelos de predicción a las características demográficas y sociales de la población penitenciaria de Euskadi, garantizando al mismo tiempo la transparencia, la validación científica y la formación continua del personal técnico. Cualquier implementación debería priorizar la función de apoyo y orientación, evitando que el algoritmo sustituya el juicio profesional o desplace la responsabilidad ética de la decisión.

4. DISCUSIÓN: ESTADO DE LA CUESTIÓN ACTUAL E IMPLICACIONES PARA EUSKADI

En los últimos años, la introducción de herramientas algorítmicas en el ámbito penitenciario ha pasado de ser una práctica experimental a consolidarse como parte esencial de la gestión del riesgo. Su expansión ha venido acompañada de un intenso debate sobre su fiabilidad, su legitimidad ética y su impacto en la toma de decisiones judiciales y penitenciarias. Aunque la literatura científica ha demostrado que los instrumentos estructurados superan en consistencia al juicio clínico no sistemático (Andrés-Pueyo & Echeburúa, 2010; Bonta & Andrews, 2007), los retos actuales no radican únicamente en la precisión estadística, sino en su correcta implementación institucional, en la formación de los profesionales y en la adaptación a los contextos locales.

Por un lado, la experiencia catalana con el RisCanvi constituye un referente europeo en la aplicación de sistemas algorítmicos al ámbito penitenciario. Su desarrollo en 2008 marcó un punto de inflexión en la gestión del riesgo delictivo, al sustituir los juicios basados exclusivamente en la experiencia por un modelo de evaluación empírica, estructurada y verificable. Como explica Marian Martínez García, responsable de la *Unitat de Programes d'Intervenció Especialitzada*, el RisCanvi nació de un consenso entre el *Departament de Justícia* y la Universidad de Barcelona tras un caso grave de reincidencia sexual que evidenció la necesidad de "unificar criterios y dotar a los profesionales de herramientas comunes" (Martínez, comunicación personal, 2 de octubre de 2025).

El instrumento se construyó sobre la base de la evidencia científica acumulada sobre predictores de reincidencia y adaptado a la realidad penitenciaria catalana mediante un estudio retrospectivo. Su finalidad no era únicamente predecir, sino gestionar preventivamente. Según Martínez, "el

algoritmo no toma decisiones por sí solo; sirve para organizar la intervención y priorizar la atención", lo que refuerza su carácter de herramienta de apoyo técnico y no de sustituto del juicio humano (Martínez, comunicación personal, 2 de octubre de 2025). Uno de los mayores aportes del modelo catalán fue su integración en el sistema informático penitenciario, lo que permitió una aplicación universal, sistemática y coordinada entre distintos perfiles profesionales. La existencia de dos niveles —*Screening* (10 factores) y *Completa* (43 factores)— facilita una primera detección de riesgo bajo y una segunda fase más detallada para los casos complejos. Esta estructura modular y su validación en equipo multidisciplinar han contribuido a institucionalizar el modelo, aunque en la práctica persisten dificultades: falta de reuniones conjuntas, automatización de las valoraciones y diferencias en la calidad de las evidencias (Martínez, comunicación personal, 2 de octubre de 2025).

Más allá de los aspectos técnicos, el RisCanvi simbolizó un cambio de paradigma hacia el modelo Riesgo-Necesidad-Responsividad (RNR) (Andrews & Bonta, 2010). Este enfoque, que sustituye la lógica de programas estándar por la individualización de las intervenciones, se ha convertido en la base conceptual del tratamiento penitenciario catalán. El nivel de riesgo determina la intensidad del seguimiento y las áreas identificadas como necesidades criminógenas se trasladan al Programa Individualizado de Tratamiento (PIT). Este sistema ha permitido alinear la intervención con los objetivos de reinserción social, favoreciendo la coordinación entre los equipos técnicos y los responsables de los programas de rehabilitación (Subdirecció General de Programes de Rehabilitació i Sanitat, 2019).

Martínez destaca, además, que el RisCanvi ha servido para reducir sesgos implícitos en la práctica profesional, particularmente en la evaluación de mujeres con conductas autolesivas, cuyos riesgos podrían haber llegado a infravalorarse. En este sentido, el algoritmo ha contribuido a reforzar la objetividad y la equidad en la valoración del riesgo, al visibilizar factores que antes podían pasar inadvertidos por prejuicios o interpretaciones subjetivas.

El modelo catalán, sin embargo, no está exento de retos. Entre los principales se encuentran la necesidad de actualizar periódicamente el algoritmo -ya que la población penitenciaria y los patrones delictivos han cambiado desde 2008— y de garantizar una formación continuada para evitar la mecanización del proceso. Como subraya la entrevistada, "necesitamos un equipo estable que revise el instrumento y evite sesgos, porque la sociedad y los delitos evolucionan". Esta dimensión adaptativa resulta crucial para mantener la validez del modelo a lo largo del tiempo.

En todo caso, la situación catalana debe entenderse dentro de un contexto internacional marcado por la expansión de los sistemas de evaluación del riesgo en ámbitos judiciales y penitenciarios. En Estados Unidos, el uso de herramientas como COMPAS ha suscitado un amplio debate sobre los límites de la inteligencia artificial en la justicia penal. Investigaciones como la de ProPublica (Angwin, Larson, Mattu & Kirchner, 2016) demostraron que el algoritmo presentaba sesgos raciales significativos, al sobrestimar el riesgo en personas negras y subestimarlo en personas blancas. Estos hallazgos reforzaron la necesidad de garantizar la transparencia y auditabilidad de los sistemas empleados, dado que las decisiones basadas en algoritmos opacos pueden tener consecuencias graves sobre la libertad y la igualdad ante la ley.

En Europa, la tendencia se ha orientado hacia modelos mixtos, en los que la tecnología complementa, pero no reemplaza, la valoración profesional. Herramientas como el HCR-20 o el LSI-R han demostrado una validez sólida cuando se aplican dentro de un marco institucional estable y con supervisión humana (Douglas et al., 2014; Andrews & Bonta, 1995). Los principios del RNR han sido ampliamente adoptados por los servicios penitenciarios europeos, aunque con distintos niveles de desarrollo y adaptación cultural.

En los últimos años, algunos países han comenzado a experimentar con modelos de cuarta generación, que combinan la evaluación de riesgo tradicional con la actualización continua de datos mediante sistemas de aprendizaje automático (*machine learning*). Estas herramientas buscan captar la evolución dinámica del comportamiento y adaptarse a los cambios en tiempo real, aunque su implementación sigue planteando desafíos éticos y metodológicos (Duwe & Kim, 2015). Por otro lado, los países escandinavos, especialmente Noruega y Finlandia, han demostrado que la tecnología puede coexistir con enfoques penitenciarios humanistas: allí, la evaluación del riesgo se utiliza principalmente para ajustar los programas de acompañamiento social y la planificación del retorno a la comunidad (Dressel & Farid, 2018; Fazel et al., 2016). Este ejemplo ilustra que la aplicación de algoritmos no tiene por qué ser sinónimo de control, sino que puede integrarse en modelos de justicia orientados a la rehabilitación y la confianza social.

El debate ético también ha cobrado especial relevancia con la aprobación del Reglamento General de Protección de Datos (RGPD) y el futuro AI Act de la Unión Europea, que clasifican los sistemas de inteligencia artificial utilizados en justicia penal como de "alto riesgo". Estas normas exigen que las decisiones automatizadas estén sometidas a control humano, sean explicables y respeten los derechos fundamentales (Bravo Bolado, 2023).

La experiencia comparada demuestra que la fiabilidad de un modelo no puede separarse de su gobernanza institucional ni de la capacidad de los profesionales para interpretarlo críticamente.

Finalmente, la transferencia de competencias penitenciarias a Euskadi en 2021 ha abierto un escenario propicio para repensar el papel de la evaluación del riesgo en la gestión penitenciaria vasca. A diferencia de Cataluña, donde el RisCanvi está plenamente implantado, Euskadi parte de un marco más flexible y con la posibilidad de diseñar un sistema propio adaptado a su realidad sociocultural y demográfica. Esta oportunidad implica tanto ventajas como desafíos.

Desde una perspectiva técnica, el modelo catalán ofrece una base sólida de referencia, pero su aplicación directa podría resultar limitada si no se contextualiza adecuadamente. Tal como advierte Martínez, "los algoritmos deben revisarse periódicamente y ajustarse a la población donde se aplican". En Euskadi, la tasa de reincidencia, la composición etaria y la estructura de delitos difieren parcialmente de las catalanas, por lo que sería necesario realizar un estudio empírico específico que permita calibrar los factores de riesgo y los puntos de corte de manera local (Martínez, comunicación personal, 2 de octubre de 2025).

Además, la introducción de un sistema algorítmico debería enmarcarse en un proyecto integral de formación para el personal penitenciario. La experiencia catalana muestra que la resistencia inicial de los profesionales se superó solo cuando comprendieron que el instrumento no era una herramienta de control, sino una guía para planificar las intervenciones. En Euskadi, donde el modelo penitenciario se orienta hacia la reinserción y la humanización del trato, la incorporación de un algoritmo debería acompañarse de procesos de co-creación con los técnicos, garantizando su apropiación y legitimidad (Férez-Mangas & Andrés-Pueyo, 2018). En términos organizativos, la posibilidad de integrar la evaluación del riesgo en un sistema informático unificado —similar al catalán— podría mejorar la trazabilidad de los datos, la coordinación entre centros y la continuidad en el seguimiento postpenitenciario. Ello permitiría también establecer indicadores comunes y generar evidencia sobre la eficacia de los programas de intervención, una de las demandas persistentes en la literatura criminológica.

Sin embargo, el caso catalán también advierte sobre los riesgos de una automatización excesiva. La pérdida del debate en los equipos multidisciplinares y la delegación del juicio al "validador" son ejemplos de cómo la eficiencia técnica puede desplazar la reflexión colectiva. En este sentido,

el modelo vasco podría apostar por un sistema híbrido que mantenga la estructura técnica del algoritmo, pero refuerce la deliberación profesional y la dimensión cualitativa de la valoración.

Por otro lado, la implementación de un sistema algorítmico en Euskadi debería incorporar una evaluación ética y de impacto social, en consonancia con los principios de justicia algorítmica (Bravo Bolado, 2023). Ello implica asegurar que el modelo no reproduzca desigualdades estructurales y que la tecnología se use como medio para reforzar la equidad y la reinserción, no para aumentar el control punitivo. La inclusión de variables de protección, la perspectiva de género y el respeto a la diversidad cultural resultan esenciales para construir un modelo contextualizado y justo.

Finalmente, la experiencia catalana demuestra la importancia de la revisión continua. Como recordó Martínez, la población penitenciaria cambia con el tiempo: "ya no consumen tanta heroína, los delitos evolucionan, las realidades sociales son otras". En consecuencia, cualquier modelo que aspire a ser fiable debe contar con un equipo estable de supervisión que revise periódicamente la base de datos, los factores de riesgo y los resultados obtenidos (Martínez, comunicación personal, 2 de octubre de 2025). En el caso de Euskadi, ello podría traducirse en la creación de un observatorio permanente sobre evaluación del riesgo y reinserción, que combine investigación aplicada, análisis de datos y participación profesional.

5. CONCLUSIONES

El análisis realizado a lo largo del presente trabajo permite afirmar que los algoritmos de predicción de la reincidencia constituyen una herramienta de gran relevancia en la gestión penitenciaria contemporánea. Su incorporación ha supuesto un cambio sustancial en la forma de entender la evaluación del riesgo, introduciendo parámetros de objetividad, uniformidad y evidencia empírica en un ámbito históricamente condicionado por la heterogeneidad de criterios profesionales. Sin embargo, este avance técnico no está exento de limitaciones, y su valor depende fundamentalmente del contexto institucional, del uso que hagan de ellos los profesionales y de las garantías éticas y jurídicas que acompañen su aplicación.

Los modelos de evaluación estructurada, tanto actuariales como mixtos, han demostrado una capacidad predictiva superior al juicio clínico no estructurado (Andrés-Pueyo & Echeburúa, 2010; Bonta & Andrews, 2007). No obstante, su precisión nunca alcanza niveles absolutos, ya que la reinci-

dencia es un fenómeno complejo, multifactorial y condicionado por variables personales, sociales y situacionales. En este sentido, los instrumentos deben entenderse como soportes técnicos de apoyo, no como mecanismos de decisión autónomos. El juicio profesional sigue siendo el elemento insustituible para contextualizar los resultados, interpretar los matices y ponderar la relevancia de cada factor en la historia individual del sujeto evaluado.

La experiencia catalana con el RisCanvi ilustra con claridad este equilibrio entre tecnología y criterio profesional. Su diseño, basado en la evidencia científica y adaptado a la realidad penitenciaria local, ha permitido homogeneizar las valoraciones y orientar de manera más coherente los programas de tratamiento. Como señaló Marian Martínez García, el instrumento no busca determinar sanciones ni resolver automáticamente las decisiones penitenciarias, sino "organizar la intervención y priorizar la atención" (comunicación personal, 2 de octubre 2025). De este modo, el algoritmo se consolida como un recurso de gestión preventiva, que facilita la detección temprana de factores de riesgo y la planificación de estrategias de intervención ajustadas al perfil criminógeno y a las necesidades de cada persona.

A pesar de su consolidación, el caso catalán revela también las tensiones inherentes al proceso de institucionalización de este tipo de herramientas. La automatización parcial de las valoraciones, la disminución de los espacios de deliberación colectiva y la posible rutina en la introducción de evidencias constituyen riesgos reales que pueden afectar la calidad del proceso de evaluación. Ello pone de relieve la necesidad de reforzar la formación continuada de los profesionales y de mantener la discusión interdisciplinar como núcleo del trabajo técnico. La fiabilidad del instrumento no depende solo de su diseño estadístico, sino del modo en que se aplica y se interpreta en la práctica diaria.

Desde una perspectiva comparada, el análisis de otros modelos internacionales —como COMPAS en Estados Unidos o HCR-20 y LSI-R en Europa— muestra que la fiabilidad técnica no garantiza por sí sola la legitimidad social. Los sistemas cerrados y no auditables, basados en algoritmos opacos, tienden a reproducir sesgos estructurales y a erosionar la confianza en las instituciones judiciales (Angwin, Larson, Mattu & Kirchner, 2016; Bravo Bolado, 2023). Por el contrario, los modelos que combinan criterios empíricos con supervisión humana tienden a ofrecer resultados más equilibrados, tanto en precisión como en equidad. En este sentido, la transparencia, la interpretabilidad y la auditabilidad se han convertido en principios rectores para el desarrollo ético de la inteligencia artificial aplicada a la justicia penal.

Las nuevas regulaciones europeas, en particular el Reglamento General de Protección de Datos (RGPD) y el AI Act, insisten en que toda decisión automatizada con impacto jurídico significativo debe estar sujeta a control humano y a criterios de explicabilidad. Este marco normativo constituye un avance decisivo hacia una gobernanza responsable de la inteligencia artificial, que busca conjugar innovación tecnológica y respeto a los derechos fundamentales. La implementación de algoritmos de riesgo debe, por tanto, integrarse dentro de una política penitenciaria que priorice la protección de la dignidad, la igualdad y la reinserción como objetivos irrenunciables.

En el plano operativo, el modelo catalán se ha consolidado como un referente técnico y organizativo, pero su eficacia futura dependerá de su capacidad para actualizarse. La propia administración penitenciaria reconoce la necesidad de revisar periódicamente los factores de riesgo, los pesos estadísticos y las bases de datos que alimentan el sistema, a fin de evitar que el instrumento se vuelva obsoleto o descontextualizado. Esta revisión continua resulta esencial para mantener la validez y la legitimidad del modelo, y debería realizarse desde una perspectiva interdisciplinar, combinando la investigación académica con la experiencia profesional de los equipos penitenciarios.

En este contexto, la transferencia de competencias penitenciarias a Euskadi ofrece una oportunidad estratégica para diseñar un modelo propio, científicamente validado y adaptado a la realidad social y demográfica del territorio. La creación de un sistema algorítmico vasco podría beneficiarse de las lecciones aprendidas en Cataluña, pero también de la posibilidad de partir de una base más flexible y de incorporar desde el inicio mecanismos de evaluación ética y participación profesional. La experiencia demuestra que los procesos de co-creación institucional —en los que los técnicos participan activamente en la definición de las variables, los procedimientos y la interpretación de los resultados— generan mayor compromiso, transparencia y sostenibilidad.

Asimismo, la configuración de un sistema propio permitiría a Euskadi integrar factores de protección y resiliencia que reflejen mejor su realidad sociocultural, así como reforzar la coordinación entre instituciones penitenciarias, servicios sociales y redes comunitarias. Este enfoque holístico, alineado con las tendencias europeas en materia de justicia restaurativa, favorecería una comprensión más integral del riesgo y del proceso de desistimiento delictivo. La creación de un observatorio permanente sobre eva-

luación del riesgo y reinserción podría garantizar la actualización continua de los instrumentos y la rendición de cuentas pública sobre sus resultados.

De forma transversal, la cooperación interautonómica e internacional emerge como un elemento clave para el futuro de estas herramientas. Compartir datos, metodologías y evaluaciones entre territorios no solo aumenta la robustez estadística, sino que permite detectar posibles sesgos estructurales y promover estándares comunes de calidad. En un ámbito tan sensible como el penitenciario, la colaboración entre administraciones —dentro del marco del Estado y con otros países europeos— puede fortalecer la transparencia, la innovación y la comparabilidad de resultados.

Por último, conviene subrayar que la incorporación de tecnología en la gestión penitenciaria no debe interpretarse como un proceso neutro ni puramente técnico. La utilización de algoritmos implica siempre decisiones normativas y éticas sobre qué variables medir, cómo interpretarlas y con qué finalidad aplicarlas. Por ello, la tecnología debe estar al servicio del modelo de justicia que se desea construir, y no a la inversa. El riesgo de despersonalización o de dependencia excesiva de los sistemas automatizados exige mantener una reflexión constante sobre los límites de la predicción y el valor insustituible del juicio humano.

En conclusión, los algoritmos de evaluación del riesgo constituyen un instrumento útil y necesario, pero no suficiente, dentro de una política penitenciaria moderna orientada a la reinserción. Su aportación principal radica en la capacidad de estructurar la información, reducir la arbitrariedad y mejorar la eficiencia en la gestión de recursos. No obstante, su eficacia real depende del equilibrio entre precisión técnica y sensibilidad humana, entre objetividad empírica y comprensión social. En un sistema que aspire a ser justo, preventivo y resocializador, la tecnología debe ocupar un lugar complementario y responsable: un medio al servicio de la persona, nunca un fin en sí misma.

6. REFERENCIAS

Alemán Aróstegui, L. (2023). El uso del RISCANVI en la toma de decisiones penitenciarias. *Estudios Penales y Criminológicos*, 44, 1-43. https://doi.org/10.15304/epc.44.8884.

Andrés-Pueyo, A. (2016). ¿Es técnicamente posible anticipar la reincidencia delictiva? El protocolo RisCanvi en las prisiones de Cataluña. En *IX Jornadas de ATIP Almagro* (pp. 55-78).

Andrés-Pueyo, A. & Redondo, S. (2007). Predicción de la violencia: entre la peligrosidad y la valoración del riesgo de violencia. Papeles del Psicólogo, 28(3), 157-173.

Andrés-Pueyo, A., & Echeburúa, E. (2010). Valoración del riesgo de violencia: instrumentos disponibles e indicaciones de aplicación [Violence risk assessment: available tools and instructions for use]. *Psicothema,* 22, 403-409.

Andrews, D. A., & Bonta, J. (1995). *The level of service inventory-revised*. Toronto, Canada: Multi-Health Systems.

Andrews, D. y Bonta, J. (2010). *The Psychology of Criminal Conduct* (5a Ed.). New Providence, NJ: LexisNexis.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine bias: There is software that is used across the country to predict future criminals. And it is biased against Blacks.* ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bonta, J., & Andrews, D. A. (2007). *Risk-needs-responsivity model for offender assessment and rehabilitation*. Ottawa, Ontario: Public Safety Canada.

Borum, R. (2000). Assessing violence risk among youth. *Journal of Clinical Psychology,* 56, 1263-1288. https://doi.org/10.1002/1097-4679(200010)56:10<1263::AID-JCLP3>3.0.CO;2-D.

Bravo Bolado, A. (2023). Justicia algorítmica: Un enfoque sociotécnico. *Estudios Penales y Criminológicos, 44*(ext), 1-42. https://doi.org/10.15304/epc.44.8838.

Brennan, T., Dieterich, W., & Oliver, W. (2007). *COMPAS: Correctional offender management for alternative sanctioning. Technical manual and psychometric report*. Traverse City, MI: Northpointe Institute for Public Management.

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior*, *36*, 21-40.

Capdevila, M., Blanch, M., Ferrer, M., Andrés, A., Framis, B., Comas, B., Mora, J. (2015). *Tasa de reincidencia penitenciaria 2014* [Prison recidivism rate 2014]. Barcelona, Spain: Centre de Estudis Jurídics i Formacio Especialitzada.

Castro Liñares, D. (2019). *Los instrumentos de valoración y gestión de riesgos en el modelo de penalidad español*. Madrid: Reus.

Chaves Pedrón, C. (2025). Valoración penitenciaria del riesgo en los permisos de salida (*Penitentiary risk assessment in exit permits*). *Cuaderno de Res publica en Derecho y Criminología, 6*, 1-19. https://doi.org/10.46661/respublica.12050.

Chico, E. (1997). The invariance in the factorial structure of the Raven in groups of criminals and non-criminals. Psicothema, 9(1), 47-55.

Douglas, K. S., Hart, S. D., Webster, C. D., Belfrage, H., Guy, L. S., & Wilson, C. M. (2014). Historical-Clinical-Risk Management-20, Version 3 (HCR-20^{V3}): Development and Overview. *International Journal of Forensic Mental Health*, *13*(2), 93-108. https://doi.org/10.1080/14999013.2014.906519.

Dressel, J. y Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism, *Science Advance,* 4, 1-5. DOI: 10.1126/sciadv.aao5580.

Duwe, G., & Kim, K. (2015). Out with the old and in with the new? An empirical comparison of supervised learning algorithms to predict recidivism. *Criminal Justice Policy Review, 28*(6), 1-31. https://doi.org/10.1177/0887403415604899.

Fanega, M., Fresneda Portillo, C. y Beltrán Camacho, A. (2024). Publicación del algoritmo penitenciario TVR. *Revista Española de Investigación Criminológica,* 22(1), 1-22. https://doi.org/10.46381/reic.v21i1.917.

Fazel, S., Chang, Z., Fanshawe, T., Långström, N., Lichtenstein, P., Larsson, H., & Mallett, S. (2016). Prediction of violent reoffending on release from prison: Derivation and external validation of a scalable tool. *The Lancet Psychiatry, 3*(6), 535-543. https://doi.org/10.1016/S2215-0366(16)00103-6.

Férez-Mangas, D. y Andrés-Pueyo, A. (2015). Predicción y prevención del quebrantamiento de los permisos penitenciarios. *Revista Española de Investigación Criminológica*, 7(13), 1-28. https://doi.org/10.46381/reic.v13i.92.

Férez-Mangas, D. y Andrés-Pueyo, A. (2018). Eficacia predictiva en la valoración del riesgo del quebrantamiento de permisos penitenciarios. *La Ley Penal, 134,* 1-14.

García, P. M. M., & Rivera, R. M. B. (2016). Psychometric prediction of penitentiary recidivism. *Psicothema*, 194-200.

Gobierno de España, Ministerio del Interior. (2022). *Estudio de reincidencia penitenciaria 2009-2019* (Documentos penitenciarios, 30). Secretaría General de Instituciones Penitenciarias. https://www.interior.gob.es/opencms/pdf/archivos-y-documentacion/documentacion-y-publicaciones/publica-

ciones-descargables/instituciones-penitenciarias/Estudio_de_reincidencia_penitenciaria_2009-2019_DP-30_126220415.pdf.

Hamilton, M. 2021. Algorithmic Risk Assessment: A Progressive Policy in Pretrial Release. *Idaho Law Review* 57 (3), 615-634. https://ssrn.com/abstract=4024142.

Howard, P., Clark, D., & Garnham, N. (2006). *An Evaluation of the Offender Assessment System (OASys) in Three Pilots 1999-2001*. London: National Offender Management Service.

Howard, P. D., & Dixon, L. (2012). The Construction and Validation of the OASys Violence Predictor: Advancing Violence Risk Assessment in the English and Welsh Correctional Services. *Criminal Justice and Behavior*, *39*(3), 287-307.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. *AEA Papers and Proceedings, 108*, 22-27.

León, C.M. y Bartolomé, R. (2024). Valoración del riesgo de reincidencia en jóvenes infractoras: ¿Es el SAVRY un instrumento realmente neutro? *Boletín Criminológico,* 30, 1-25. https://doi.org/10.24310/bc.30.2024.20683.

Loinaz, I. (2017). *Manual de evaluación del riesgo de violencia*. *Metodología y ámbitos de aplicación*. Pirámide.

Luque, E., Ferrer, M., & Capdevila, M. (2005). *La reincidencia penitenciaria a Catalunya (1997-2002)* [Penitentiary recidivism in Catalunya 1997-2002]. Barcelona, Spain: Centre d'Estudis Jurídics i Formació Especialitzada.

Martínez García, M. (2025). *Entrevista personal realizada por L. Padilla* [Transcripción]. En **Anexo 1**, *Algoritmos de reincidencia: ¿Herramienta útil para Euskadi? Revisión bibliográfica sobre variables, modelos e interpretación en el ámbito penitenciario.*

Martínez, M. L. (2014). La incertidumbre de los pronósticos de peligrosidad: consecuencias para la dogmática de las medidas de seguridad [The uncertainty of forecasts of dangerousness: consequences for dogmatic security measures]. *InDret*, 2, 1-77.

Martínez Escamilla, M. (2002). *Los permisos ordinarios de salida: Régimen jurídico y realidad* [Ordinary prison leaves: Legal framework and reality]. Edisofer.

Martínez Garay, L. (2016). Errores conceptuales en la estimación de riesgo de reincidencia. La importancia de diferenciar sensibilidad y valor predictivo, y estimaciones de riesgo absolutas y relativas. *Revista Española de Investigación Criminológica*, 3(14), 1-31. https://doi.org/10.46381/reic.v14i.97.

Martínez Garay, L., y Montes Suay, F. (2018). El uso de valoraciones del riesgo de violencia en Derecho Penal: algunas cautelas necesarias. *Revista para el Análisis del Derecho, 2*, 2-47.

Medina, P. M. (2013). *Evaluación experimental de la eficacia de los programas psicológicos de tratamiento penitenciario [Experimental evaluation of the effectiveness of psychological programs for prison treatment*]. Madrid, Spain: Ministerio del Interior, Secretaría General Técnica.

Miró-LLinares, F. (2018). Inteligencia Artificial y justicia penal: más allá de los resultados lesivos causados por robots [Artificial Intelligence and criminal justice: Beyond the harmful results caused by machines]. *Revista de Derecho Penal y Criminología, 20*, 87-130. https://doi.org/10.5944/rdpc.20.2018.26446.

Morales Moreno, A. M. (2021). Algoritmos en el estrado, ¿realmente los aceptamos? Percepciones del uso de la inteligencia artificial en la toma de decisiones jurídico-penales. *Ius Et Scientia,* 7(2), 57-87. https://dx.doi.org/10.12795/IETSCIENTIA.

Nguyen, T., Arbach-Lucioni, K., & Andrés-Pueyo, A. (2019). Factores de riesgo de la reincidencia violenta en población penitenciaria. *Revista de Derecho Penal y Criminología* (6), 273-294. Recuperado a partir de https://revistas.uned.es/index.php/RDPC/article/view/24608.

Ortega, E., García, J. y Frías, M. (2014). Meta-análisis de la reincidencia criminal en menores: Estudio de la investigación española. *Revista Mexicana de Psicología*, *31* (2), 111-123.

Ortega-Campos, E., García-García, J., De la Fuente-Sánchez, L., & Zaldívar-Basurto, F. (2020). Predicting risk of recidivism in Spanish young offenders: Comparative analysis of the SAVRY and YLS/CMI. *Psicothema, 30(*2), 221-228. doi:10.7334/psicothema2019.275.

Pelechano, V. (2008). Delincuencia, personalidad y psicopatología [Delinquency, personality and psychopathology*]. Análisis y Modificación de Conducta*, 34(150-151), 13-66. 10.33776/amc.v34i150-151.701.

Pérez Ramírez, M., Redondo Illescas, S., Martínez García, M., García Forero, C., & Andrés Pueyo, A. (2008). Predicción de riesgo de reincidencia en agresores sexuales. *Psicothema, 20*(2), 205-210.

Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. RAND Corporation.

Portela, M., Castillo, C., Tolan, S., Karimi-Haghighi, M., & Andrés-Pueyo, A. (2022). A Comparative User Study of Human Predictions in Algorithm-Supported Recidivism Risk Assessment. *Computers and Society-Ar-Xvid.* https://doi.org/10.48550/arXiv.2201.11080.

Real Decreto 474/2021, de 29 de junio, de traspaso de funciones y servicios de la Administración del Estado a la Comunidad Autónoma del País Vasco sobre ejecución de la legislación del Estado en materia penitenciaria. *BOE*, núm. 161, 7 de julio de 2021. https://www.boe.es/eli/es/rd/2021/06/29/474.

Rodríguez, A., López, J.M., y Andrés Pueyo, A. (2002). Personalidad y comportamiento penitenciario. *Psicothema,* 14 (Supl.), 90-100.

Sánchez-Teruel, D., & Robles-Bello, M. (2013). Model "Big Five" personality and criminal behavior. *International Journal of Psychological Research*, 6(1), 102-109. https://doi.org/10.21500/20112084.709.

Scaria, A. G., Subramanian, V., George, N. K., & Sengupta, N. (2024). Algorithms and Recidivism: A Multi-disciplinary Systematic Review. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*(1), 1292-1305. https://doi.org/10.1609/aies.v7i1.31724.

Singh, J., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. Clinical Psychology Review, 31(3), 499-513. https://doi.org/10.1016/j.cpr.2010.11.009.

Skeem, J., & Lowenkamp, C. (2020). Using algorithms to address trade-offs inherent in predicting recidivism. *Behavioral Sciences & the Law, 38*(3), 259-278. https://doi.org/10.1002/bsl.2465

Subdirecció General de Programes de Rehabilitació i Sanitat, Departament de Justícia. (2019). *Manual d'aplicació del protocol de valoració RisCanvi*. https://repositori.justicia.gencat.cat/bitstream/handle/20.500.14226/1320/manual-aplicacio-protocol-avaluacio-riscanvi.pdf.

Capítulo 3

La valoración predictiva del riesgo en la ejecución penal: límites, garantías y propuestas para el ámbito penitenciario vasco

Izaskun Ibabe

Universidad del País Vasco UPV/EHU

SUMARIO: 1. INTRODUCCIÓN. 2. MODELOS DE HERRAMIENTAS PREDICTIVAS DE VALORACIÓN DEL RIESGO. 2.1. *El modelo Riesgo-Necesidad-Responsividad como marco teórico.* 2.2. *Tipologías de instrumentos de valoración del riesgo.* 2.2.1. Juicio clínico no estructurado. 2.2.2. Juicio clínico estructurado. 2.2.3. Métodos actuariales. 2.2.4. Modelos mixtos básicos y avanzados. 2.3. *Límites predictivos, validez externa y garantías jurídicas.* 3. USO DE HERRAMIENTAS PREDICTIVAS EN EL ÁMBITO ESPAÑOL. 4. MULTI-ESCALA RISCANVI. 5. PROPUESTA DE MARCO PROSPECTIVO PARA LA EVENTUAL INCORPORACIÓN DE UNA HERRAMIENTA EN EL SISTEMA PENITENCIARIO VASCO. 6. CONCLUSIONES: BASES PARA UN MODELO AVANZADO Y GARANTISTA. 7. REFERENCIAS.

1. INTRODUCCIÓN

La incorporación de herramientas predictivas de valoración del riesgo en el ámbito penal y penitenciario constituye uno de los desarrollos más relevantes —y controvertidos— de la política criminal contemporánea. Estas herramientas están diseñadas para estimar la probabilidad de reincidencia o de comisión de conductas violentas, se utilizan de forma habitual como apoyo a decisiones administrativas y judiciales con un impacto directo sobre derechos fundamentales, tales como la clasificación penitenciaria, la concesión de permisos o la libertad condicional (Andrés-Pueyo y Echeburúa, 2010; Fazel et al., 2019).

Se estima que en todo el mundo hay más de 11 millones de personas que se encuentran actualmente privadas de libertad (Fair y Walmsley, 2021). Los estudios que han analizado la reincidencia en periodos de seguimiento de dos años informan tasas de nuevo arresto situadas entre el 26% y el 60%, de nueva condena entre el 20% y el 63%, y de reingreso en prisión entre el 14% y el 45% (Yukhnenko et al., 2020). Esta variabilidad pone de relieve la complejidad del fenómeno de la reincidencia y la dependencia de los resultados del indicador empleado (arresto, condena o encarcelamiento). En cualquier caso, la reducción de la reincidencia constituye una de las prioridades centrales del sistema de justicia penal, en consonancia con los objetivos de seguridad ciudadana y prevención del delito.

Durante las últimas décadas se han desarrollado nuevas herramientas para la predicción de la conducta violenta, lo que ha supuesto un cambio de paradigma desde la estimación de la peligrosidad hacia la valoración probabilística del riesgo de reincidencia. Este enfoque incorpora tanto factores de riesgo estáticos como dinámicos, e incluso explora indicadores neurobiológicos y factores de protección específicos según el tipo de reincidencia (Andrés-Pueyo y Echeburúa, 2010; Van Dongen et al., 2025). En este contexto, el uso de las herramientas de evaluación del riesgo se ha promovido como un enfoque basado en la evidencia científica para valorar la probabilidad de reincidencia y orientar decisiones relativas a la condena y la rehabilitación, con intervenciones dirigidas a factores de riesgo modificables (Fazel et al., 2019; Van Ginneken, 2019;).

En cualquier caso, una parte apreciable de la doctrina jurídica se ha mostrado crítica respecto a la utilización de este tipo de instrumentos, al señalar limitaciones en su validez científica y advertir sobre su posible impacto en derechos fundamentales, como la presunción de inocencia o el principio de culpabilidad (Martínez y Montes, 2018). En la medida en que el derecho penal se basa en hechos ya ocurridos y en la responsabilidad personal del autor/a, la introducción de enfoques orientados a anticipar y gestionar riesgos puede generar tensiones con los principios de legalidad y proporcionalidad del sistema penal. Desde el punto de vista criminológico, estas tensiones se explican por las limitaciones estructurales de la predicción del comportamiento delictivo y por el impacto que la gestión del riesgo tiene sobre los procesos de intervención y reinserción.

En el caso del sistema penitenciario vasco, tras la asunción de las competencias en materia de ejecución penitenciaria en 2021, la eventual implantación de alguna herramienta predictiva avanzada plantea la necesidad de

examinar críticamente la transferibilidad de esos instrumentos, dado que la mayoría de ellos han sido desarrollados en contextos jurídicos, institucionales y poblacionales distintos. Por ello, sería necesario valorar tanto el rendimiento predictivo de las herramientas como las condiciones jurídicas, culturales y organizativas necesarias para su implantación.

En este marco, el primer objetivo de este trabajo es analizar los diferentes modelos de herramientas predictivas de valoración y gestión del riesgo en el ámbito de la ejecución penal, desde una aproximación crítica e interdisciplinar. El segundo objetivo consiste en examinar los principales instrumentos que se utilizan actualmente en el contexto español, considerando sus características, ámbitos de aplicación y los principales retos derivados de su uso práctico. Por último, el estudio tiene como tercer objetivo formular una propuesta de marco prospectivo para la eventual incorporación de una herramienta de valoración y gestión del riesgo en el sistema penitenciario vasco.

2. MODELOS DE HERRAMIENTAS PREDICTIVAS DE VALORACIÓN DEL RIESGO

En la valoración y gestión del riesgo de violencia y reincidencia en el ámbito penal y penitenciario se utilizan una pluralidad de herramientas técnicas. La selección de herramientas depende del propósito de la evaluación, del contexto institucional y de las propiedades psicométricas de la herramienta. En general, el objetivo último de estos sistemas no es la predicción del comportamiento delictivo en sí misma, sino la prevención de la violencia y la reducción de la reincidencia mediante la identificación temprana de personas con mayor nivel de riesgo y la orientación de intervenciones proporcionales. En este marco, el análisis comparado de los distintos modelos de valoración del riesgo resulta imprescindible para comprender tanto su alcance predictivo como los límites metodológicos y jurídicos que condicionan su utilización en la toma de decisiones que afectan a derechos fundamentales.

2.1. EL MODELO RIESGO-NECESIDAD-RESPONSIVIDAD COMO MARCO TEÓRICO

El marco teórico más aceptado en la evaluación y gestión del riesgo en el entorno penal y penitenciario es el modelo Riesgo-Necesidad-Responsividad (Andrews y Bonta, 2010; Polaschek, 2012), que se articula en torno a tres principios básicos: proporcionalidad del nivel de intervención al riesgo estimado, atención prioritaria a las necesidades criminógenas dinámicas y adaptación de

las medidas a las características personales del penado/a. En este ámbito, la valoración del riesgo se concibe como un instrumento orientado a garantizar intervenciones individualizadas y basadas en evidencia empírica.

La valoración del riesgo de violencia o reincidencia mediante herramientas estructuradas y tecnologías predictivas es un procedimiento muy complejo, cuya interpretación requiere conocimiento técnico y comprensión estadística. Además, esta valoración plantea desafíos jurídicos en cuanto a la correcta calibración del riesgo, la validez externa de los instrumentos y la prevención de sesgos, porque los resultados de dichas evaluaciones pueden incidir en decisiones que afectan a derechos fundamentales (Beaudry et al., 2023; Van Dongen et al., 2025). Las consecuencias derivadas de los errores predictivos —falsos positivos, que implican restricciones injustificadas de derechos, y falsos negativos, que suponen un riesgo de revictimización de personas concretas—, obligan a los/as profesionales responsables de la valoración del riesgo a afrontar este reto con la máxima rigurosidad científica y ética.

Desde una perspectiva jurídico-penitenciaria, la integración del modelo RNR en los instrumentos de valoración y gestión del riesgo exige examinar críticamente su compatibilidad con las exigencias de motivación, proporcionalidad y control jurisdiccional, lo que implicaría garantizar la transparencia del proceso evaluativo, la supervisión humana efectiva y la posibilidad de revisión y contradicción de las valoraciones de riesgo.

2.2. TIPOLOGÍAS DE INSTRUMENTOS DE VALORACIÓN DEL RIESGO

En el ámbito de la evaluación del riesgo de violencia y de reincidencia penitenciaria, los instrumentos disponibles se organizan en cuatro tipologías diferenciadas, que se describen en la Tabla 1. Esta clasificación permite analizar sus fundamentos teóricos y metodológicos, su eficacia predictiva y las principales ventajas y limitaciones, así como los instrumentos más utilizados tanto a nivel nacional como internacional en cada categoría. A través de esta clasificación se observa la evolución histórica y técnica de la valoración del riesgo, desde la subjetividad pura hasta el procesamiento masivo de datos.

2.2.1. Juicio clínico no estructurado

Este enfoque es el más tradicional y se basa en el uso de información clínica relevante, como la historia personal, la entrevista y los datos diagnósticos. La valoración depende fundamentalmente de la interpretación

libre que realizan los/as profesionales —psicólogos/as, criminólogos/as, trabajadores/as sociales o educadores/as— a partir de la información disponible, sin apoyarse en guías estructuradas ni en protocolos basados en evidencia empírica.

Los resultados de los estudios empíricos indican que este modelo ofrece una baja fiabilidad inter-evaluador y una limitada transparencia en el proceso de toma de decisiones (Andrés-Pueyo y Echeburúa, 2010; Redondo, 2017), y un rendimiento predictivo inferior al de los modelos estructurados —tanto actuariales como de juicio clínico estructurado— (Ægisdóttir et al., 2006; Viljoen et al., 2025). Además, su aplicación en el ámbito penitenciario no garantiza la homogeneidad de criterios en la adopción de decisiones con impacto jurídico relevante, como la concesión de la libertad condicional, los permisos penitenciarios, la clasificación o la planificación de intervenciones y tratamientos, lo que plantea serias objeciones en cuanto a las garantías propias del Estado de Derecho (Hilterman, 2019).

2.2.2. Juicio clínico estructurado

El juicio clínico estructurado es un enfoque solvente para la valoración del riesgo de violencia, al combinar la sistematicidad de las guías basadas en evidencia con la flexibilidad del juicio profesional. Este método parte de una lista estandarizada de factores de riesgo y protección, identificados por la investigación empírica, que deben ser evaluados de manera rigurosa y documentada por la persona experta (Douglas et al., 2013; Douglas y Otto, 2021). En esta modalidad no se utiliza ningún algoritmo y el profesional tiene la libertad de dar un peso diferente a cada factor según el caso individual e incluir variables idiosincrásicas no previstas en la herramienta. Su principal aportación radica en la gestión del riesgo, ya que permite elaborar recomendaciones individualizadas, diseñar intervenciones y planificar medidas de control y protección ajustadas a los factores dinámicos identificados. Así, este tipo de herramientas ofrece un marco de evaluación estructurado, transparente y adaptable, que reduce la subjetividad inherente al juicio clínico tradicional sin renunciar a la contextualización necesaria en la práctica forense. Es algo más dependiente de la calidad del profesional y menos transparente estadísticamente que un modelo estrictamente actuarial. Estos instrumentos no requieren, para su aplicación, estudios de validación empírica externa en cada contexto en el mismo sentido que los instrumentos actuariales; aunque resulta altamente recomendable para reforzar su correcta interpretación, coherencia metodológica y legitimidad jurídica.

2.2.3. Métodos actuariales

Estos métodos utilizan instrumentos estructurados y basados en modelos estadísticos previamente validados. Utilizan algoritmos para predecir el riesgo de reincidencia basándose en grupos de población con características similares, siendo este tipo de instrumentos muy objetivos, pero poco flexibles. Su objetivo es estimar de manera cuantificada la probabilidad de que una persona cometa una conducta violenta en el futuro. La persona evaluadora introduce datos en un algoritmo o tabla de puntuación fija que genera un resultado automático de peligrosidad o probabilidad. Una limitación de este tipo de instrumentos es priorizar factores **estáticos** (históricos e inmutables, o la edad del primer delito), lo cual hace que tenga una alta eficacia predictiva, pero son factores de riesgo difíciles de modificar mediante intervención (Viljoen et al., 2025), y otra es que no sirven para la identificación de las necesidades criminógenas **dinámicas**, tan importantes en el tratamiento y la intervención penitenciaria en aras de la rehabilitación de los reclusos (Andrews y Bonta, 2010; Férez-Mangas y Andrés-Pueyo, 2018; Harris et al., 2015). Lo que define a un instrumento actuarial no es solamente que los factores de riesgo sean estáticos, sino que la suma de puntos sea rígida. En general, este tipo de instrumentos son más precisos que el juicio clínico no estructurado y que varios métodos de juicio clínico estructurado para la predicción de reincidencia y violencia, con un nivel de precisión de aceptable a bueno (Hanson y Morton-Bourgon, 2009; Singh et al., 2013). Algunos instrumentos pueden considerarse actuariales puros, en la medida en que la estimación del riesgo se obtiene exclusivamente mediante algoritmos estadísticos basados en factores estáticos y sin posibilidad de ajuste clínico, como STATIC-2000-R, VRAG-R, SORAG o ODARA, mientras que otros no lo son (DV-MOSAIC, SVRA-I, EPV-R, VPR).

2.2.4. Modelos mixtos básicos y avanzados

Los modelos mixtos se orientan a integrar la robustez empírica de los predictores actuariales con la capacidad interpretativa del juicio profesional. Sin embargo, pueden distinguirse versiones básicas y avanzadas en función de su complejidad técnica y de su orientación principal. Por un lado, los modelos mixtos básicos se centran fundamentalmente en la clasificación del riesgo mediante criterios estadísticos, incorporando el juicio clínico como elemento interpretativo complementario. Y, por otro lado, los modelos mixtos avanzados cuentan con herramientas tecnológicas y los factores estáticos son elementos de referencia para la estimación inicial del riesgo, pero los integran con factores dinámicos y evaluaciones periódicas, de modo que el

nivel de riesgo no queda fijado de forma rígida, sino que puede ajustarse en función de la evolución individual y de la intervención aplicada.

En ambos casos, la evaluación del riesgo se apoya en instrumentos que integran factores estáticos de base estadística, así como factores dinámicos y contextuales, examinados clínicamente, de modo que la persona evaluadora combina la información derivada del algoritmo con el juicio profesional para formular una estimación final del nivel de riesgo y orientar la toma de decisiones (Andrés-Pueyo y Redondo, 2007; Douglas y Otto, 2021). Este enfoque busca conjugar precisión predictiva y flexibilidad clínica, y se encuentra estrechamente alineado con el modelo R-N-R, al permitir vincular la estimación del riesgo con la planificación de estrategias de gestión individualizadas orientadas a la reducción del riesgo mediante la intervención sobre factores dinámicos modificables (Andrews y Bonta, 2010; Andrews et al., 2006).

La correcta aplicación de los modelos mixtos exige una protocolización rigurosa del proceso evaluativo, así como una adecuada calidad y actualización de los datos actuariales. Al mismo tiempo requiere una formación especializada de los/as profesionales, con el fin de minimizar la variabilidad inter-evaluador y preservar la fiabilidad, la coherencia metodológica y la transparencia del procedimiento. La literatura advierte que a pesar de todo este equilibrio puede verse comprometido cuando la intervención clínica introduce ajustes sustanciales no justificados ni documentados sobre los resultados actuariales, lo que reduce la explicabilidad del proceso, debilita su coherencia interna y puede afectar negativamente tanto a la precisión predictiva como a su control externo, especialmente en contextos de toma de decisiones con relevancia jurídica (Douglas y Otto, 2021; Fazel et al., 2019). En este sentido, sería deseable que pudiera verse el modo en que el algoritmo llega al resultado, y en caso de identificar algún sesgo, poder corregirlo (Hogan Doran, 2021).

En este marco, la inteligencia artificial (IA) debe entenderse como una extensión tecnológica de los modelos predictivos, pero no como un equivalente a los modelos mixtos clínico-actuariales, ya que no garantiza la integración efectiva del juicio profesional. Los sistemas basados en el aprendizaje automático (machine learning) representan la generación más reciente de tales instrumentos en el ámbito criminológico y penitenciario, con capacidad para identificar patrones complejos y alcanzar niveles de precisión similares o ligeramente superiores a los modelos actuariales tradicionales (Berk, 2017). Aun así, plantea relevantes desafíos metodológicos y jurídicos, especialmente en relación con el volumen de datos que necesitan, la opacidad de los modelos, los riesgos de sesgo, la validación externa y la explicabilidad de las predicciones, lo que exige mecanismos reforzados de supervisión,

transparencia y control institucional (Dressel y Farid, 2018; Rudin, 2019). El progreso futuro de la IA exige avanzar hacia modelos de explicabilidad intrínseca y auditorías estandarizadas que refuercen la transparencia, el control institucional y la compatibilidad de estos sistemas con las exigencias propias del Estado de Derecho en decisiones de alto impacto.

2.3. LÍMITES PREDICTIVOS, VALIDEZ EXTERNA Y GARANTÍAS JURÍDICAS

La eficacia predictiva de los instrumentos de valoración del riesgo se evalúa habitualmente mediante el área bajo la curva ROC (AUC, Area Under the Curve). El modelo asigna una puntuación de riesgo mayor a un caso que presenta mayor probabilidad de que el evento se produzca. Conforme a los puntos de corte propuestos en la literatura especializada, valores de AUC ≤ 0,63 indican una capacidad discriminativa baja, valores entre 0,64 y 0,78 una capacidad moderada, y valores > 0,78 una capacidad alta (Harris et al., 2015; Rice y Harris, 2005).

Estudios recientes muestran que, aunque la capacidad predictiva de los instrumentos de valoración del riesgo sigue oscilando en rangos moderados, la evidencia actual continúa respaldando estas estimaciones en revisiones sistemáticas y meta-análisis recientes (AUC entre aproximadamente 0,60 y 0,80, en función del tipo de instrumento, población y contexto (Ling et al., 2025; Viljoen et al., 2025). Los instrumentos actuariales y los de juicio clínico estructurado presentan un rendimiento predictivo superior al del juicio clínico no estructurado en la predicción de conductas violentas (Viljoen et al., 2025). En cuanto a modelos de machine learning, Kozhevnikova et al. (2025) identificaron 38 estudios que desarrollaron y validaron 40 modelos de predicción de violencia, con valores de AUC generalmente elevados, pero con importantes limitaciones metodológicas. La mayoría de los estudios presentó elevado riesgo de sesgo, escasa evaluación de la calibración y muy pocas validaciones externas, lo que se traduce en una baja utilidad clínica y limitada generalizabilidad. Aunque los modelos de aprendizaje automático de tipo caja negra muestran potencial para identificar individuos de alto riesgo, su aplicación práctica exige mejoras sustanciales en calidad metodológica, explicabilidad, predicciones dinámicas y validación externa.

Tanto los sistemas actuariales como los basados en la IA aplicados a la valoración del riesgo se fundamentan en probabilidades estadísticas derivadas de datos poblacionales, que estiman la frecuencia esperada de conductas violentas o delictivas en grupos de personas con características compartidas.

Aunque estas estimaciones se proyectan sobre casos individuales, no constituyen predicciones individualizadas en sentido estricto ni garantizan una anticipación precisa del comportamiento de una persona concreta. Como ya señalaba Allport (1942), aun cuando un elevado porcentaje de personas procedentes de un determinado entorno resultara reincidente, ello no permite inferir que un individuo concreto con ese mismo origen presente idéntica probabilidad de reincidir, dado que las probabilidades grupales no pueden extrapolarse directamente al comportamiento individual. Esta limitación es inherente a cualquier modelo probabilístico, incluidos los modelos avanzados de IA, y pone de relieve que estas herramientas, aunque útiles para la gestión del riesgo a nivel poblacional, no sustituyen el juicio clínico ni el análisis individualizado del caso concreto. De hecho, la conducta humana no es plenamente predecible, sino contingente y sensible a múltiples factores contextuales, sociales y personales.

La validez externa es un criterio metodológico clave en la evaluación de instrumentos de valoración del riesgo, pues determina si el rendimiento predictivo observado en los estudios originales se mantiene en poblaciones y contextos distintos. Estudios empíricos han documentado variaciones importantes en la discriminación y calibración de herramientas en aplicaciones reales dependientes del contexto, lo que subraya la necesidad de validaciones externas amplias antes de su implementación en un determinado contexto jurídico (Fazel et al., 2022; Spivak et al., 2024; Zhu et al., 2025). Las instituciones que contemplen la inclusión de una herramienta predictiva por primera vez deberían someterla a estudios de validación independientes como parte integral de su estrategia de implementación. Su rendimiento predictivo constituirá solo uno de los criterios de evaluación, junto con su viabilidad de implementación y sostenibilidad a gran escala, la transparencia y las consideraciones éticas.

Al aplicar una herramienta hay que considerar que el desempeño de los modelos predictivos avanzados depende críticamente de la calidad y cantidad de los datos. Aunque determinadas herramientas actuariales y algorítmicas puedan alcanzar valores de validez predictiva relativamente elevados —con AUC que, en algunos estudios, se sitúan en torno a 0,70-0,80—, su uso exclusivo no resulta aconsejable en la práctica judicial. La dependencia predominante de factores estáticos limita su utilidad para la gestión dinámica del riesgo y la planificación de intervenciones orientadas a la reinserción social. Además, la opacidad en la toma de decisiones y la posible reproducción de sesgos históricos comprometen principios fundamentales como la transparencia y la igualdad ante la ley (UNESCO, 2025).

La aplicación de herramientas de predicción del riesgo en el ámbito penal entraña el riesgo de vulnerar los principios de proporcionalidad y no discriminación, en la medida en que puede reforzar desigualdades estructurales y perpetuar estereotipos sobre determinados colectivos, especialmente cuando se basan en datos históricos sesgados o carecen de mecanismos adecuados de control y revisión (Cueto-Santa Eugenia, 2025). Si bien la función jurisdiccional aspira a ejercerse con imparcialidad y objetividad, no puede ignorarse que los operadores del sistema penal pueden verse influidos por prejuicios inconscientes; sin embargo, la introducción de herramientas basadas en IA no elimina este riesgo, sino que puede reproducir o amplificar sesgos estructurales preexistentes, en la medida en que dichos sistemas se alimentan de datos históricos que reflejan desigualdades y prácticas discriminatorias previas. Desde el marco normativo europeo, el Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de IA, prohíbe expresamente la adopción de decisiones que supongan juzgar o sancionar a las personas sobre la base de comportamientos predichos mediante sistemas de IA fundamentados en la elaboración de perfiles. Se considera que este tipo de prácticas vulnera principios esenciales como la dignidad humana, la presunción de inocencia y la prohibición de discriminación.

El uso de herramientas algorítmicas automatizadas en el sistema penal exige la protección de datos personales. En el caso de que el uso de algoritmos con impacto sobre los derechos de las personas habría que informar a las personas afectadas sobre la lógica de funcionamiento del sistema, como garantía de protección de sus datos y de control de la decisión (Castagnedi-Ramírez, 2024). Sin embargo, esta exigencia se ve comprometida cuando se emplean modelos opacos, lo que plantea relevantes problemas éticos y jurídicos en términos de transparencia y rendición de cuentas.

Además, la literatura reciente subraya que la interpretabilidad constituye un requisito esencial para garantizar la legitimidad, la equidad y el control democrático de los sistemas predictivos utilizados en el ámbito penal y penitenciario (Guidotti et al., 2019; Rudin, 2019; Wachter et al., 2018). En esta línea, se insiste en la necesidad de priorizar modelos intrínsecamente explicables y de complementarlos con mecanismos capaces de justificar predicciones individuales, facilitar su revisión técnica y jurídica y detectar posibles sesgos ocultos. La evidencia disponible indica, además, que las mejoras obtenidas mediante modelos algorítmicos avanzados respecto a las escalas actuariales tradicionales suelen ser moderadas y no siempre clínicamente relevantes. Por ello, se recomienda integrar estos instrumen-

tos con modelos clínicos estructurados o enfoques mixtos, que aportan flexibilidad interpretativa y permiten incorporar información contextual relevante para la gestión del riesgo y la reinserción social (Cueto-Santa Eugenia, 2025).

Desde el contexto criminológico, la valoración del riesgo se enfrenta al desafío estructural de intentar anticipar conductas extremadamente infrecuentes —como, afortunadamente, el asesinato—, lo que introduce severas limitaciones predictivas derivadas del problema de la tasa base e incrementa de forma significativa la probabilidad de falsos positivos y falsos negativos (Fazel et al., 2012). Puesto que las tasas de reincidencia no son constantes ni comparables entre países, periodos o muestras, resulta más adecuado interpretar los resultados de las herramientas de evaluación del riesgo en términos de posición percentil, esto es, atendiendo al nivel de riesgo relativo de una persona en comparación con otras evaluadas mediante el mismo instrumento (Weber et al., 2025).

Por todo ello, los instrumentos de valoración del riesgo sirven de apoyo al juicio de los operadores jurídicos, y no como sustitutos de este, garantizando su control y supervisión por parte de las autoridades jurisdiccionales, que mantienen la responsabilidad última en la toma de decisiones. En conjunto, la evidencia disponible indica que ningún modelo resulta óptimo en términos absolutos, siendo los enfoques estructurados y mixtos los que ofrecen, en la actualidad, un mejor equilibrio entre capacidad predictiva, gestión del riesgo y respeto a las garantías jurídicas en el ámbito penitenciario.

Tabla 1. *Tipologías de instrumentos de valoración del riesgo en el ámbito penal y penitenciario.*

Eficacia predictiva	Tipo de herramienta	Ventajas y limitaciones	Instrumentos	Utilidad
BAJA	Juicio clínico no estructurado	Basada en intuición profesional *Baja fiabilidad y transparencia, alto sesgo*	No se aplica ninguno	—
MODERADA	Juicio clínico estructurado (dinámicos)	Juicio profesional guiado por evidencia, factores protectores, gestión riesgo *Exigencia de formación, subjetividad residual, laborioso*	HCR-20 V3 SARA SAVRY SVR-20	Violencia general Violencia de pareja Violencia juvenil Violencia sexual

Eficacia predictiva	Tipo de herramienta	Ventajas y limitaciones	Instrumentos	Utilidad
ALTA	Herramientas actuariales (estáticos/ dinámicos)	Algoritmo, alta validez predictiva teórica, predicción automática, elimina sesgos subjetivos *Prioriza factores estáticos, baja utilidad preventiva*	STATIC-2000-R LSI-R SORAG ODARA	Reincidencia sexual Reincidencia general Violencia sexual Violencia de pareja
MODERADA/ ALTA	Modelos mixtos básicos/ avanzados (actuarial + juicio clínico)	Algoritmo/IA, objetividad, contexto, elevada utilidad preventiva, aceptación institucional y ética *Exigencia elevada de formación, conflicto de criterios, riesgo de opacidad, mayor carga de trabajo*	TRV (básico) OASYS (avanzado) COMPAS (avanzado) RISCANVI (avanzado)	Reincidencia y violencia Reincidencia general Reincidencia general Reincidencia y violencia

Nota: Baja: AUC 0,60-0,63; Moderada: AUC 0,65-0,78; Alta: AUC > 0,78.

3. USO DE HERRAMIENTAS PREDICTIVAS EN EL ÁMBITO ESPAÑOL

España no ha sido ajena a la progresiva incorporación de herramientas de valoración y gestión del riesgo en el ámbito penitenciario. Si bien su implantación no ha alcanzado aún el grado de generalización propio de los sistemas penales anglosajones, puede apreciarse una tendencia paulatina hacia la introducción de este tipo de instrumentos en la práctica penitenciaria (Pereira-Puigvert, 2024).

En el sistema penitenciario español, y en particular en los centros dependientes de la Administración General del Estado, se emplean fundamentalmente dos instrumentos de valoración y gestión del riesgo que operan de forma complementaria: la Tabla de Riesgo de Violencia (TRV) y la Tabla de Circunstancias Concurrentes del Permiso (CCP) (Alemán-Aróstegui, 2023). En la actualidad, transcurridos más de veinticinco años desde su creación y entrada en vigor, ambos instrumentos continúan plenamente operativos sin haber sido objeto de una revisión o actualización integral. Las modificaciones introducidas se han limitado al CCP, mediante la incorporación de nuevas circunstancias específicas a través de distintas instrucciones dictadas por la

Secretaría General de Instituciones Penitenciarias, sin que ello haya supuesto una reforma estructural del sistema (Férez-Mangas y Andrés-Pueyo, 2018).

La TRV es un instrumento de valoración del riesgo de naturaleza esencialmente actuarial, con una intervención profesional de carácter complementario. No puede calificarse como un modelo mixto clínico-actuarial en sentido estricto, en la medida en que no integra el juicio clínico en el cálculo del nivel de riesgo ni permite ajustes razonados y estructurados de la estimación final. La herramienta está diseñada para clasificar el riesgo de reincidencia mediante la evaluación de diez factores de riesgo, cada uno de los cuales se puntúa entre 0 (ausencia de riesgo) y 3 (presencia elevada de riesgo), en función de su intensidad y relevancia. La puntuación total permite establecer a la persona evaluada en siete niveles de riesgo, que oscilan desde muy bajo hasta riesgo máximo. Se trata, por tanto, de un instrumento orientado a la clasificación del riesgo en categorías relativamente estables, que no incorpora tecnología algorítmica avanzada, sistemas de seguimiento dinámico ni procedimientos automatizados de reevaluación periódica. En consecuencia, la TRV no integra mecanismos propios de los modelos avanzados de gestión del riesgo ni permite la modificación estructurada del nivel de riesgo en función de cambios evolutivos en la situación de la persona evaluada. Tras la aplicación de la TRV, la valoración del riesgo debe complementarse necesariamente con la CCP, que incorpora variables adicionales relevantes para la toma de decisiones administrativas, especialmente en relación con la concesión o denegación de permisos penitenciarios. Este instrumento permite contextualizar el nivel de riesgo previamente estimado a la luz de factores institucionales y personales no contemplados de forma directa en la TRV.

Junto a estos instrumentos, el Ministerio del Interior implantó el sistema VioGén en 2007, destinado a la clasificación del riesgo de las víctimas de violencia de género mediante procedimientos algorítmicos avanzados, con el objetivo de clasificar el nivel de riesgo de que una mujer sufra nuevas agresiones por parte de su agresor, a fin de orientar la adopción de medidas policiales y judiciales de protección. Se trata de cuestionarios estructurados y reglas de decisión automatizadas; sin embargo, no constituye un modelo mixto clínico-actuarial avanzado, dado que la estimación del riesgo no es modificable mediante juicio clínico y la intervención profesional se produce en una fase posterior, vinculada a la gestión policial del riesgo. Esta herramienta, por un lado, tiene las limitaciones metodológicas y jurídicas derivadas de los factores estáticos, con riesgo de reproducir sesgos estructurales, y limitada explicabilidad del algoritmo para las personas afectadas. Y, por

otro, plantea tensiones relevantes con principios fundamentales, en particular la proporcionalidad, la no discriminación y el derecho a una decisión motivada. Aunque el sistema VioGén se concibe formalmente como una herramienta policial orientada a la protección de la víctima, la clasificación del nivel de riesgo puede tener efectos jurídicos indirectos relevantes sobre el agresor, al influir en la adopción de medidas cautelares, en la intensidad de la intervención policial y en la construcción de una presunción fáctica de peligrosidad, con potencial incidencia sobre derechos fundamentales como la libertad personal, la presunción de inocencia y el derecho de defensa.

La Comunidad Autónoma del País Vasco cuenta con su propia plataforma, Etxekoen eta emakumeen BAbesa (EBA, Protección de la familia y de las mujeres), destinada a la protección integral frente a la violencia doméstica y de género. Este sistema unifica la información relativa al maltrato en el ámbito familiar, facilita la evaluación del riesgo y permite la coordinación de los distintos servicios implicados, pudiendo además interconectarse con el sistema estatal VioGén-2, cuya versión tecnológicamente avanzada fue implantada en 2025. En todo caso, la coexistencia de EBA y VioGén plantea retos relevantes en materia de coordinación e interoperabilidad de la información. Asimismo, en el ámbito autonómico se han desarrollado otros modelos de evaluación del riesgo, como el protocolo multi-escala RisCanvi, aplicado en la Comunidad Autónoma de Cataluña, que se configura como uno de los instrumentos más consolidados en el contexto español.

4. MULTI-ESCALA RISCANVI

La Dirección General de Asuntos Penitenciarios de Cataluña implementó el protocolo de evaluación RisCanvi en 2009 como herramienta universal de valoración del riesgo y de las necesidades criminológicas de las personas internas. En el segundo semestre de 2024, de un total de 8.488 personas de la población penitenciaria catalana, el 81% fue evaluado, y el 23% de dicha población mostraba un perfil de riesgo bajo en todos los criterios de predicción (Battle, 2025). El algoritmo funciona por medio de factores de riesgo tanto estáticos como dinámicos y ofrece unos resultados que se dividen en cinco tipos de comportamiento de riesgo: violencia autodirigida, violencia intrainstitucional, reincidencia general, reincidencia violenta y quebrantamiento de condena (Andrés-Pueyo, 2016). La aplicación de esta herramienta constituye un procedimiento habitual de trabajo de los/as profesionales penitenciarios, que comparten información entre distintos técnicos, garantizan su actualización constante y consolidan el uso de instrumentos técnicos de validez y utilidad contrastada empíricamente (Andrés-Pueyo, 2016). Uno

de los objetivos de esta herramienta es la individualización de las intervenciones atendiendo al riesgo, las posibilidades de cambio personal y las necesidades criminógenas de las personas internas, siguiendo el modelo RNR (Andrews y Bonta, 2007).

Según el Manual de aplicación del protocolo RisCanvi (2019), la herramienta cuenta con una versión de cribado o screening (RisCanvi-S) con 10 factores de riesgo, mientras que la versión completa (RisCanvi-C) incluye 43 factores de riesgo agrupados en cinco categorías. La versión completa se aplica si la persona evaluada obtiene con el RisCanvi-S una puntuación de riesgo alto. Dado que se trata de una herramienta que corresponde al modelo mixto avanzado, presenta las mismas virtudes y carencias, por lo que es susceptible de recibir las mismas críticas que se han expuesto anteriormente. Sin embargo, una de las cuestiones más relevantes es la aplicación en la práctica en el ámbito penitenciario, si sus resultados pueden acabar siendo determinantes del sentido de la decisión, ya que legalmente no existen previsiones que especifiquen estos extremos (Alemán-Aróstegui, 2023).

Este instrumento se aplica de forma periódica con un intervalo habitual recomendado de 6 meses o bien, a criterio de los responsables, cuando se considere oportuno como por ejemplo ante un cambio de centro penitenciario, o nueva situación personal. La administración del RisCanvi se realiza por medio de un programa informático (e-RisCanvi) que está incluido en el SIPC (Sistema de Información Penitenciaria de Cataluña). En la codificación de los factores de riesgo participa el equipo multidisciplinar (juristas, educadores/as sociales, criminólogos/as, psicólogos/as y pedagogos/as) supervisados generalmente por un/a subdirector/a del centro penitenciario que revisa y valida las valoraciones finales del RisCanvi de las personas internas a su cargo.

El estudio cualitativo realizado por Alemán-Aróstegui (2023), basado en entrevistas a profesionales que aplicaban el protocolo RisCanvi, pone de relieve diversos aspectos positivos del instrumento, entre los que se encuentran la posibilidad de obtener un retrato más preciso de la situación criminógena de la persona evaluada, así como la detección de errores en la intervención y su posterior corrección. Pese a ello, los/as profesionales señalan como principal aspecto negativo la considerable carga de trabajo que supone la aplicación del instrumento y la subjetividad de algunos ítems. A continuación, se describen las cinco fases de la aplicación del protocolo RisCanvi junto a las limitaciones que destacan los profesionales.

a) *Recogida de información*. Los equipos multidisciplinares son los que tienen contacto directo con las personas privadas de libertad y se encargan

de realizar entrevistas y consultar la documentación sobre los factores de riesgo. Los/as profesionales advierten la dificultad de obtener información fiable y suficiente acerca de los factores de riesgo.

b) *Registro de las evidencias*. Los/as profesionales del equipo multidisciplinar registran en el SIPC las denominadas evidencias correspondientes a 43 ítems del RisCanvi-C, distribuidos por áreas de especialización profesional. La mayoría de los/as profesionales entrevistados/as subrayan la ambigüedad y el carácter subjetivo de diversos factores de riesgo, ya que las evidencias requeridas tienen un carácter valorativo. Por su parte, los/as jefes de servicio señalan la necesidad de mejorar esta fase del proceso, debido a que se detectan contradicciones entre la información aportada por distintos profesionales y la existencia de datos no actualizados.

c) *Valoración de los factores de riesgo*. A partir de las evidencias registradas en la fase anterior, se determina la presencia o ausencia de los distintos factores de riesgo conforme a la definición de cada escala, habitualmente en el marco de reuniones de los equipos multidisciplinares. Los/as profesionales manifiestan dificultades para interpretar ítems de carácter más abstracto o contextual, como el desajuste infantil, los procesos de socialización problemática en la familia de origen, así como los valores antisociales. La falta de tiempo disponible provoca en ocasiones que esta valoración sea realizada de forma individual por las personas validadoras —mandos de los equipos o profesionales específicamente acreditados—, en detrimento del carácter colegiado que formalmente se atribuye a esta fase del proceso.

d) *Cierre de la evaluación y generación de resultados.* La persona validadora se encarga del cierre de la evaluación y validación. Posteriormente, se generan los resultados del RisCanvi en cuanto a los niveles de riesgo de los cinco tipos de comportamientos. Los resultados obtenidos tienen una vigencia de seis meses, siendo la evaluación periódica una característica de los modelos avanzados. Esta herramienta prevé la posibilidad de que el equipo, pero no la persona validadora, realice correcciones en el nivel de riesgo reflejado por el sistema.

e) *Aplicación del resultado RisCanvi*. El resultado de la herramienta actúa como un elemento de apoyo en los procesos de toma de decisiones, sin limitar la decisión del equipo, sino orientándola. Así, la asignación de un nivel de riesgo alto no impide, por sí sola, la aprobación de una salida o permiso penitenciario, siempre que dicha decisión

se acompañe de un plan adecuado de gestión y control del riesgo. En aquellos casos en los que el equipo discrepa del nivel de riesgo asignado, procede a revisar y corregir la evaluación. Aunque es posible apartarse del resultado del RisCanvi cuando el equipo lo considera, en los delitos que generan una mayor alarma social o tienen una elevada visibilidad mediática —como los delitos sexuales—, el peso del resultado del instrumento aumenta considerablemente, pudiendo llegar a ser determinante en la decisión final.

5. PROPUESTA DE MARCO PROSPECTIVO PARA LA EVENTUAL INCORPORACIÓN DE UNA HERRAMIENTA EN EL SISTEMA PENITENCIARIO VASCO

La eventual incorporación de una herramienta de valoración y gestión del riesgo en el contexto vasco requiere un enfoque especialmente prudente, atento a las especificidades del sistema vasco, al reducido tamaño de su población penitenciaria y a la experiencia previa en el uso de instrumentos de apoyo a la intervención. Estas características condicionan de manera directa la utilidad, los límites y los riesgos de las citadas herramientas, reforzando la necesidad de garantías metodológicas y jurídicas.

La introducción de este tipo de instrumentos no puede entenderse como una transferencia automática de modelos desarrollados en otros contextos, sino como una decisión de política penitenciaria que debe ser evaluada a la luz del modelo vasco de ejecución penal, orientado a la individualización de la intervención y a la reinserción social. El marco prospectivo que se presenta a continuación se concibe, por tanto, como una hoja de ruta gradual y garantista, orientada a explorar la eventual adopción de un modelo mixto avanzado, basado en la combinación de herramientas técnicas y supervisión humana significativa.

Fase 1. Análisis preliminar y delimitación del marco jurídico-institucional.

En esta primera fase, resulta imprescindible llevar a cabo un análisis preliminar que no se limite al marco jurídico e institucional, sino que incorpore de manera expresa las especificidades contextuales del País Vasco, entre ellas el tamaño reducido de la población, el volumen de casos gestionados y las características del modelo penitenciario propio. En contextos poblacionales reducidos, la baja tasa base de los delitos graves acentúa las limitaciones predictivas de los instrumentos de evaluación del riesgo y refuerza la nece-

sidad de cautelas metodológicas y garantías jurídicas reforzadas. Para ello sería necesario plantearse los siguientes objetivos:

- Identificación de los ámbitos de aplicación posibles (penitenciario, medidas alternativas, seguimiento post-penitenciario).
- Análisis de competencias y necesidad de habilitación jurídica expresa.
- Evaluación de la compatibilidad con el modelo penitenciario vasco, los principios de legalidad, proporcionalidad e individualización, los estándares de protección de derechos fundamentales.
- Elaboración de un informe sobre población, volumen de casos y tasa base.
- Creación de un grupo de trabajo interdisciplinar (juristas, criminólogos/as, profesionales penitenciarios, expertos/as en derechos fundamentales).

Un resultado objetivable de esta fase sería la elaboración de un informe jurídico-estratégico que determine si resulta viable la incorporación o desarrollo de una nueva herramienta de valoración del riesgo de última generación, en qué *ámbitos podría utilizarse y con* qué finalidades concretas, atendiendo a su compatibilidad con el marco jurídico, institucional y garantista vigente.

Fase 2. Selección o diseño del instrumento

En esta fase se procedería a la selección, adaptación o diseño de una herramienta de valoración del riesgo, atendiendo a criterios de transparencia, explicabilidad, supervisión humana y compatibilidad con el modelo institucional vasco. Esta fase tendría un carácter estrictamente técnico y exploratorio, sin efectos jurídicos directos, y podría incluir la evaluación de instrumentos existentes, la adquisición de entornos de prueba o el desarrollo de prototipos propios. La eventual habilitación jurídica expresa quedaría reservada a una fase posterior, previa a cualquier uso operativo del instrumento en la toma de decisiones que afecten a derechos fundamentales. En esta fase correspondería realizar la selección razonada de un instrumento existente o decidir el diseño de un instrumento nuevo. A tal efecto, esta fase se articula en torno a los siguientes objetivos:

- Evaluación comparada de instrumentos existentes a nivel nacional o internacional.

- Análisis de transparencia del modelo, tipo de predictores utilizados, posibilidad de supervisión humana significativa.
- Decisión estratégica: adaptación, desarrollo propio o descartes justificados.

Un resultado objetivable de esta fase sería una decisión motivada sobre el instrumento más adecuado, la adaptación de algún instrumento existente o la decisión de diseñar un nuevo instrumento *ad hoc*.

Fase 3. Validación empírica externa y adaptación contextual

La validación externa y la adaptación contextual constituyen una fase imprescindible para verificar la idoneidad de la herramienta en el contexto vasco antes de cualquier uso operativo. Para ello, es necesario realizar un:

- Estudio piloto con datos del contexto vasco.
- Evaluación rigurosa de discriminación (capacidad para diferenciar reincidencia/no reincidencia), calibración (ajuste entre riesgo estimado y real), estabilidad de los predictores en población vasca.
- Revisión crítica de sesgos potenciales (clase, género, origen, tipo delictivo).
- Evaluación del proceso dirigida a las personas responsables implicadas.

La prueba piloto puede realizarse en unidades seleccionadas o con responsables administrativos concretos. El objetivo de esta fase es identificar y corregir deficiencias en el procedimiento, optimizar los formularios y poner en marcha el proceso de monitorización de datos. Para ello es necesario desarrollar políticas y protocolos que faciliten la integración de la evaluación del riesgo en la práctica ordinaria. La implementación piloto debería ser evaluada mediante encuestas dirigidas a las personas responsables implicadas, abordando posteriormente los problemas y dificultades detectados. Este tipo de evaluación contribuye a incrementar la implicación y aceptación del personal. Además, pueden realizarse auditorías de casos con el fin de evaluar la fiabilidad inter-evaluador a nivel local y el grado de adherencia al protocolo establecido.

El resultado esperado de esta fase es la obtención de evidencia empírica sólida y contextualizada que permita valorar la idoneidad del instrumento

para su aplicación en Euskadi, en términos de validez predictiva, adecuación al contexto institucional y proporcionalidad de su uso. Dicha evidencia debería servir de base para decidir, de forma informada y motivada, si procede avanzar hacia su eventual implantación operativa, su modificación o su descarte.

Fase 4. Diseño de garantías reforzadas

En esta fase, el objetivo central es evitar cualquier forma de automatización de las decisiones y reforzar la legitimidad democrática del uso del instrumento mediante garantías jurídicas, institucionales y procedimentales adecuadas. Para ello es preciso seguir los siguientes pasos:

- Definición normativa del carácter no vinculante del resultado.
- Obligación de motivación reforzada cuando se sigue el resultado del instrumento o se aparta de él.
- Regulación de supervisión humana significativa, mecanismos de revisión y corrección.
- Establecimiento de protocolos de protección de datos personales.
- Garantizar la trazabilidad de las decisiones y la posibilidad de reclamación efectiva por parte de las personas penadas.

El resultado esperado de esta fase es configurar un marco jurídico de garantías que limite el poder punitivo y evite decisiones opacas.

Fase 5. Formación especializada y profesionalización

La consolidación del modelo exige una formación especializada y continuada de los/as profesionales, acompañada de la evaluación periódica de la fiabilidad inter-jueces, como instrumento clave para detectar desviaciones, mejorar la aplicación homogénea de la herramienta y reforzar la calidad y legitimidad de las decisiones adoptadas.

- Formación (técnica, jurídica, ética y criminológica) obligatoria y continuada para equipos multidisciplinares, personas validadoras, y responsables de toma de decisiones.
- Programas de entrenamiento orientados a reforzar la consistencia y la fiabilidad entre evaluadores/as en la aplicación del instrumento.

- Evaluación periódica del impacto en el volumen de trabajo de los/as profesionales.

El resultado esperado de esta fase es garantizar un uso informado, crítico y responsable del instrumento, sustentado en una adecuada formación y profesionalización de los equipos implicados.

Fase 6. Implementación gradual y evaluación continua

El objetivo de esta fase es garantizar una implantación progresiva y revisable del instrumento, orientada a la evaluación continuada de su funcionamiento y a la adopción, cuando proceda, de medidas de ajuste, suspensión o retirada. Para ello, se establecen:

- Implementación progresiva y limitada.
- Evaluaciones periódicas independientes sobre el impacto en decisiones, tasas de error, y efectos en derechos fundamentales.
- Mecanismos expresos de revisión, suspensión o retirada del instrumento.

El resultado esperado de esta fase es un sistema de valoración del riesgo flexible, revisable y controlado democráticamente.

Por último, la eventual utilización del protocolo RisCanvi u otro instrumento avanzado existente por parte del Gobierno Vasco exigiría una habilitación jurídica expresa antes de la implantación. Además, la sostenibilidad económica debería analizarse de forma progresiva y subordinada a la validez empírica y a las garantías jurídicas del instrumento. En una fase inicial, el análisis de costes cumple una función orientativa, destinada a descartar opciones inviables. En fases posteriores, el coste debe evaluarse en relación con su valor añadido real, incluyendo los gastos de implantación, mantenimiento, formación y supervisión, así como el coste de oportunidad frente a alternativas no algorítmicas. Solo tras acreditarse la idoneidad y proporcionalidad del instrumento puede el factor económico incorporarse como elemento decisorio.

6. CONCLUSIONES: BASES PARA UN MODELO AVANZADO Y GARANTISTA

En la gestión penitenciaria contemporánea se ha generalizado el uso de protocolos de valoración del riesgo de reincidencia como instrumento para

la aplicación de medidas preventivas adaptadas a las necesidades criminógenas de las personas privadas de libertad, en sustitución de enfoques centrados exclusivamente en el endurecimiento de las respuestas penales (Andrés-Pueyo, 2016). La evaluación diferenciada del riesgo permite una asignación más eficiente y proporcional de los recursos públicos, orientando las intervenciones más intensivas hacia los casos con mayores necesidades criminógenas. Asimismo, cuando estas herramientas se utilizan de forma transparente y no automatizada, pueden reforzar la motivación y legitimidad de las decisiones administrativas o judiciales, al aportar criterios explícitos, verificables y revisables. Finalmente, su carácter dinámico favorece una mejor individualización de la intervención, mediante el seguimiento periódico de la evolución del riesgo y la adaptación de las medidas adoptadas a los cambios relevantes en la situación de la persona evaluada. No obstante, debe subrayarse que tanto los algoritmos como los propios profesionales pueden introducir sesgos y generar efectos discriminatorios, lo que hace imprescindible la existencia de mecanismos adecuados de control y supervisión.

A pesar de las ventajas señaladas, la incorporación de herramientas algorítmicas de la valoración predictiva del riesgo plantea relevantes desafíos éticos, jurídicos y sociales (Cueto-Santa Eugenia, 2025). Por eso resulta necesario identificar a los grupos en situación de riesgo que puedan ser tratados injustamente desde el punto de vista ético por el sistema, como pueden ser las minorías étnicas con bajos niveles de alfabetización o con dificultades lingüísticas, así como las personas con trastornos mentales o con discapacidades físicas que puedan dificultar la adecuada interacción con los/as profesionales del sistema judicial. En cualquier caso, resultaría de interés analizar de manera específica el impacto psicológico de la valoración del riesgo en las personas internas, incluyendo variables como el estrés, la autoimagen o la ansiedad directamente atribuibles al proceso evaluativo.

El modelo de supervisión humana resultaría eficaz, si el funcionamiento del algoritmo fuera explicable, de modo que permitiera comprender la lógica subyacente a los resultados que genera y, en su caso, detectar y corregir posibles sesgos, errores o disfunciones (Hogan-Doran, 2021). Sin embargo, este requerimiento puede plantear dificultades prácticas cuando las personas que desarrollan la plataforma tecnológica solicitan la protección de su propiedad intelectual, tanto en relación con el *software* como con el diseño técnico de las herramientas empleadas (Zouave y Marquenie, 2017). En la medida en que este tipo de instrumentos se utilizan para la prestación de servicios públicos y puedan incidir en la adopción de decisiones con relevancia jurídica, la ausencia de mecanismos que permitan conocer y supervisar su

funcionamiento resultaría incompatible con los principios de transparencia y rendición de cuentas (Borges-Blázquez, 2024).

En definitiva, la eventual incorporación de una herramienta de valoración y gestión del riesgo en el ámbito penitenciario vasco —en un contexto en el que ya se emplean otros instrumentos de apoyo a la intervención— solo resulta defendible si se articula como un modelo mixto avanzado, basado en la combinación de herramientas técnicas y una supervisión humana significativa. La formación especializada de los/as profesionales es una condición necesaria para evitar usos mecánicos o acríticos del instrumento y para reforzar el control humano efectivo. Este modelo debe desarrollarse como un proceso gradual, evaluable y garantista, sustentado en una validación empírica externa rigurosa y en mecanismos efectivos de transparencia, trazabilidad y control, a fin de evitar procesos de automatización opaca del poder punitivo. La propuesta de marco prospectivo aquí formulada pretende contribuir a futuras decisiones normativas e institucionales compatibles con la innovación tecnológica y el respeto a las garantías jurídicas.

7. REFERENCIAS

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G. y Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34*(3), 341-382. https://doi.org/10.1177/0011000005285875.

Alemán-Aróstegui, L. (2023). El uso de RisCanvi en la toma de decisiones penitenciarias. *Estudios Penales y Criminológicos, 44*(Ext.), 1-43. https://doi.org/10.15304/epc.44.8884

Allport, G. W. (1942). *The use of personal documents in psychological science*. Social Science Research Council.

Andrés-Pueyo, A. (2016). ¿Es técnicamente posible anticipar la reincidencia delictiva? El protocolo RisCanvi en las prisiones de Cataluña. En *IX Jornadas de ATIP Almagro* (pp. 55-78).

Andrés-Pueyo, A. y Echeburúa, E. (2010). Valoración del riesgo de violencia: instrumentos disponibles e indicaciones de aplicación. *Psicothema, 22*, 403-409.

Andrews, D. A. y Bonta, J. (2010). *The psychology of criminal conduct* (5ª ed.). Routledge.

Beaudry, G., Yu, R., Miller, O., Prescott-Mayling, L., Fanshawe, T. R. y Fazel, S. (2023). Prediction of violent reoffending in people released from prison in England: external validation study of a risk assessment tool (OxRec). *Journal of Criminal Justice, 86*, 102061. https://doi.org/10.1016/j.jcrimjus.2023.102061.

Borges-Blázquez, R. (2024). Algoritmización de la concesión de medidas cautelares en el proceso penal para la protección de víctimas de violencia de género. ¿Es capaz VioGén de interpretar el *periculum in mora*? *Actualidad Jurídica Iberoamericana, 21*, 384-407.

Castagnedi-Ramírez, A. E. (2024). La construcción de un algoritmo "ético". Ius et Scientia, 10(2), 123-151. https://doi.org/10.12795/IESTSCIENTIA.2024.i02.06.

Cueto-Santa Eugenia, E. (2025). *Herramientas de predicción criminal y su impacto en nuestro proceso penal*. *Ius et Scientia, 11*(1), 14-31. https://doi.org/10.12795/IESTSCIENTIA.2025.i01.01

Douglas, K. S. y Otto, R. K. (Eds.). (2021). *The handbook of violence risk assessment* (2nd ed.). Routledge.

Douglas, K. S., Hart, S. D., Webster, C. D. y Belfrage, H. (2013). HCR-20 V3: *Assessing risk for violence — User guide*. Mental Health, Law, and Policy Institute.

Fair, H. y Walmsley, R. (2021). *World Prison Population List* (13th ed.). Institute for Crime & Justice Policy Research. https://www.prisonstudies.org/sites/default/files/resources/downloads/world_prison_population_list_13th_edition.pdf.

Fazel, S., Singh, J. P., Doll, H. y Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24,827 people: systematic review and meta-analysis. The BMJ, 345, e4692. https://doi.org/10.1136/bmj.e4692.

Fazel, S., Wolf, A., Vazquez-Montes, M. D. L. Á., Chang, Z., Lichtenstein, P., Larsson, H., & Lundström, S. (2019). Prediction of violent reoffending in prisoners and individuals on probation: a Dutch validation study (OxRec). *Scientific Report, 9*, 841. https://doi.org/10.1038/s41598-018-37539-x

Férez-Mangas, D. y Andrés-Pueyo, A. (2018). Eficacia predictiva en la valoración del riesgo del quebrantamiento de permisos penitenciarios. *La Ley Penal, 134*, 1-14.

Generalitat de Catalunya Departament de Justícia (2019). *Manual d'aplicació del protocol de valoració RisCanvi*. https://repositori.justicia.gencat.cat/bitstream/handle/20.500.14226/1320/manual-aplicacio-protocol-avaluacio-riscanvi.pdf.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. y Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1-42. https://doi.org/10.1145/3236009.

Hanson, R. K. y Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21*(1), 1-21. https://doi.org/10.1037/a0014421.

Harris, G. T., Rice, M. E., Quinsey, V. L. y Cormier, C. A. (2015). *Violent offenders: appraising and managing risk*. American Psychological Association.

Hilterman, E.D. (2019). *Predicción, valoración y gestión del riesgo de reincidencia.* Universitat Oberta de Catalunya. https://openaccess.uoc.edu/server/api/core/bitstreams/03539e1d-416a-4862-90e9-2a9a8ff60843/content.

Hogan-Doran, D. (2021). Computer says "no": automation, algorithms and artificial intelligence in Government decision-making. En *Handbook for Judicial Officers* (pp. 1113-1143). Judicial Commission of New South Wales.

Martínez, L. y Montes, F. (2018). El uso de valoraciones del riesgo de violencia en Derecho Penal: algunas cautelas necesarias. *InDret Revista para el Análisis del Derecho, 2*, 1-47.

Pereira-Puigvert, S. (2024). La justicia y el proceso penal en clave algorítmica: Nuevos enfoques, nuevos riesgos. *Ius et Scientia, 10*(2), 66-79. https://doi.org/10.12795/IESTSCIENTIA.2024.i02.03.

Polaschek, D. L. (2012). An appraisal of the risk-need-responsivity (RNR) model of offender rehabilitation and its application in correctional treatment. *Legal and Criminological Psychology, 17*(1), 1-17.

Redondo, S. (2017). *Evaluación y tratamiento de delincuentes jóvenes y adultos*. Pirámide.

Rudin, C. (2019). *Stop* explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206-215. https://doi.org/10.1038/s42256-019-0048-x.

Singh, J. P., Desmarais, S. L. y Van Dorn, R. A. (2013). Measurement of predictive validity in violence risk assessment studies: A meta-review. *Behavioral Sciences & the Law, 31*(1), 55-73. https://doi.org/10.1002/bsl.2052.

UNESCO (2025). Ética de la inteligencia artificial y justicia algorítmica. https://www.unesco.org/es/articles/etica-de-la-inteligencia-artificial-y-justicia-algoritmica

Van Dongen, J. D., Haveman, Y., Sergiou, C. S. y Choy, O. (2025). Neuroprediction of violence and criminal behavior using neuro-imaging data: From innovation to considerations for future directions. *Aggression and Violent Behavior, 80*, 102008. https://doi.org/10.1016/j.avb.2024.102008.

Van Ginneken, E.F.J.C. (2019). The use of risk assessment in sentencing. En J.W. de Keijser, J.V. Roberts y J. Ryberg (eds.), *Predictive sentencing: Normative and empirical perspectives* (pp. 9-32). Oxford: Hart Publishing.

Viljoen, J. L., Goossens, I., Monjazeb, S., Cochrane, D. M., Vargen, L. M., Jonnson, M. R., Adam, J. E. B., Shanna M.Y. L. & Jackson, J. R. (2025). Are risk assessment tools more accurate than unstructured judgments in predicting violent, any, and sexual offending? A meta-analysis of direct comparison studies. *Behavioral Sciences & The Law, 43*(1), 75-113. https://doi.org/10.1002/bsl.2698.

Wachter, S., Mittelstadt, B. y Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841-887.

Weber, M. A., Schnyder, N., Kirschstein, M. A., Graf, M., Endrass, J. y Rossegger, A. (2025). The key role of base rates: systematic review and meta-analysis of the predictive value of four risk assessment instruments. *Swiss Medical Weekly, 155*(2), 3517-3517. https://doi.org/10.57187/s.3517.

Yukhnenko, D., Sridhar, S. y Fazel, S. (2020). A systematic review of criminal recidivism rates worldwide: 3-year update. *Wellcome Open Research, 3*, 4-28. https://doi.org/10.12688/wellcomeopenres.14970.3.

Zouave, E. T. y Marquenie, T. (2017). An inconvenient truth: Algorithmic transparency & accountability in criminal intelligence profiling. *2017 European Intelligence and Security Informatics Conference (EISIC)*, 17-23. https://doi.org/10.1109/EISIC.2017.12.

Capítulo 4

Comparativa europea de herramientas algorítmicas e inteligencia artificial para predecir la reincidencia de la población penitenciaria y recomendaciones para su potencial implementación en la CAPV

Aitor Martínez González

Instituto Vasco de Criminología (IVAC-KREI)

SUMARIO: 1. INTRODUCCIÓN. 2. HERRAMIENTAS ALGORÍTMICAS PREDICTORAS DE REINCIDENCIA EN LA UE. 2.1. *Finlandia*. 2.2. *Países Bajos*. 2.3. *Estonia*. 2.4. *Grecia, Chipre, Portugal y Bulgaria*. 2.5. *Suiza.* 3. CONSIDERACIONES ÉTICAS Y LEGALES A TENER EN CUENTA PARA SU IMPLEMENTACIÓN EN LA CAPV. 4. CONCLUSIONES. 5. REFERENCIAS.

1. INTRODUCCIÓN

La inteligencia artificial promete apoyar la prevención del delito mediante la identificación algorítmica de patrones, brindando así la oportunidad de predecir, anticipar y prevenir mejor el crimen (Rolland, 2021). La reflexión de Puolakka et al., 2021, sobre el futuro del uso de la IA en prisiones hace hincapié en los dilemas morales y éticos relacionados con la relación de los humanos con las máquinas, y en la medida en que los humanos depositarán su confianza en algoritmos y robots en el futuro.

Uno de los principales objetivos detrás del uso de la IA en el contexto penitenciario es la mejora de los procesos de identificación de recursos que den respuesta a las necesidades de la población penitenciaria, así como la minimización de los riesgos de reincidencia (Puolakka P. &., 2021).

A este respecto, el artículo 22 del Reglamento General de Protección de Datos de la Unión Europea[6] determina que "*El interesado tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar*." Además, el Reglamento Europeo de Inteligencia Artificial[7] considera que el uso de la IA en el apoyo a decisiones en la administración de justicia, así como herramientas del riesgo criminal o de reincidencia conlleva un "*alto riesgo*" por lo que estas prácticas están sujetas a una estricta supervisión humana. Teniendo en cuenta el marco legislativo de la UE, las decisiones que se toman sobre el riesgo de reincidencia apoyadas por herramientas algorítmicas no son tomadas por el propio algoritmo, sino que requieren de supervisión y validación profesional (Portela, 2024).

A la hora de entender las herramientas algorítmicas para predecir la reincidencia delictiva, resulta imprescindible definir a grandes rasgos dos modelos de IA, el estadístico y el aprendizaje automático (machine learning) (Van Dijck, 2022):

- Modelo de aprendizaje automático: es el proceso mediante el cual se construye un modelo a partir de datos de muestra —denominados datos de entrenamiento— con el fin de realizar predicciones o tomar decisiones. El propio Sistema "descubre patrones y entrena los algoritmos de forma autónoma. Una de las principales herramientas de este modelo es el COMPAS (Correctional Management Profiling for Alternative Sanctions), utilizado en los EE.UU, o el Casandra en Ucrania[8].

- Modelo estadístico: depende de conceptos matemáticos para hallar patrones en los datos, tal como son definidos por el investigador que diseña el modelo. Como por ejemplo el OxRec en Suecia y en Países Bajos.

---

6. Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos.
7. Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial.
8. Cassandra es un sistema desarrollado en 2020 por el Ministerio de Justicia de Ucrania que incorpora elementos de *machine learning* para automatizar la evaluación del riesgo de reincidencia y elaborar informes previos al juicio y a la sentencia. Actualmente en fase inicial, combina datos introducidos por profesionales con análisis automatizado, con la intención de integrar en el futuro toda la información disponible del infractor y ampliar su uso a la revisión de decisiones judiciales (Fair Trials, 2021).

Uno de los principales desafíos a la hora de investigar los sistemas de predicción del riesgo de reincidencia en Europa ha sido la dificultad para acceder a información detallada y actualizada sobre su diseño, implementación y evaluación. Esta limitación responde, por un lado, a la opacidad y falta de transparencia generalizada que rodea el funcionamiento de estas herramientas, muchas de las cuales son desarrolladas en contextos institucionales con escasa rendición de cuentas pública y sin evaluaciones independientes disponibles. Por otro lado, debe señalarse que uno de los límites metodológicos de este artículo es el idioma, ya que la búsqueda de literatura se ha restringido a fuentes disponibles en castellano e inglés, lo que ha podido excluir documentación relevante publicada en otros idiomas europeos como el alemán, francés, neerlandés, sueco, etc. donde se están implementando sistemas predictivos de interés. Esta combinación de factores dificulta una comprensión completa del panorama europeo y pone de relieve la necesidad de fomentar la transparencia, el acceso abierto a datos y estudios técnicos, así como la cooperación transnacional en la evaluación de estas tecnologías.

2. HERRAMIENTAS ALGORÍTMICAS PREDICTORAS DE REINCIDENCIA EN LA UE

La reincidencia es un fenómeno complejo que carece de una definición estandarizada a nivel europeo, ya que varía entre jurisdicciones y puede medirse mediante nueva detención, nueva condena o nuevo ingreso en prisión (Innovative Criminal Justice Solutions, 2025). Esta ambigüedad dificulta la estandarización de políticas públicas europeas eficaces.

Entre los factores de riesgo, se distinguen los estáticos —como edad, género o historial penal—, que aunque predicen bien, no son modificables, y los dinámicos —como necesidades de salud mental, abuso de sustancias, relaciones antisociales, problemas de empleo, ser soltero o divorciado, bajo nivel de ingresos—, que pueden ir cambiando y en muchos casos abordarse con programas de intervención (Yukhnenko, 2020).

Por otra parte, la investigación de Blanchette, K. (2002) identificó que las mujeres infractoras presentan necesidades específicas vinculadas al trauma, la salud mental y la parentalidad, lo que requiere un enfoque sensible al género en las herramientas predictivas.

En el contexto europeo, el uso de sistemas de inteligencia artificial (IA) para la predicción del riesgo de reincidencia presenta una tendencia creciente, aunque su adopción requiere salvaguardas robustas en materia ética, de

transparencia y de rendición de cuentas (Wieringa, 2020). Estos sistemas, junto con herramientas de toma de decisiones automatizadas (ADM), ya están siendo implementados por autoridades policiales y judiciales para apoyar resoluciones en materia de sentencia y libertad condicional, evaluando el "supuesto riesgo" de criminalidad o reincidencia en el futuro" (Fair Trials, 2021). Además, en este punto se recogen herramientas actuariales de aplicación práctica en países de la UE, así como iniciativas académicas orientadas a la mejora del uso de la IA en el contexto europeo.

2.1. FINLANDIA

El sistema finlandés contempla un proceso mixto de valoración del riesgo de reincidencia. Las decisiones son tomadas por profesionales penitenciarios y se basan en dos modelos: "Risk-Need-Responsivity" y el "Good Lives Model". En este caso, la IA se utiliza como apoyo para tomar las decisiones y reforzar la capacidad técnica de los profesionales, pero sin sustituir su criterio (Puolakka P. , 2023).

En mayo de 2022 introdujeron un nuevo Sistema de gestión penitenicaria (Offender Management System) llamado ROTI. Este Sistema utiliza la Inteligencia Artificial para evaluar y analizar los factores de riesgo y las necesidades de los presos para reducir la reincidencia delictiva. ROTI sirve como repositorio centralizado de datos de internos, incluye evaluaciones de necesidades, historial penal, estado de salud y otros aspectos relevantes para la planificación de condenas.

Además de ROTI, en Finlandia utilizan el RISE AI. Un sistema que simula el proceso de evaluación del infractor orienta la condena y recomienda servicios rehabilitadores a los presos durante su tiempo en prisión. Este sistema utiliza la información almacenada sobre el preso, así como varios factores de riesgo ciminógeno. Se concibe como una nueva herramienta de apoyo para los coordinadores superiores responsables de la planificación de las condenas, no como una herramienta de toma de decisiones automatizada, ya que esto iría en contra del uso ético de la inteligencia artificial en el ámbito penitenciario. Desde este servicio entienden la precisión predictiva como la satisfacción de las necesidades psico-sociales de los presos, así como la reducción del riesgo de reincidencia (Puolakka P. &., 2021). Aunque no se detallan auditorías técnicas específicas para ROTI o RISE AI, las directrices éticas nacionales recomiendan que los sistemas de IA en la administración pública estén sujetos a mecanismos de revisión y control, incluyendo la participación de los funcionarios responsables.

Estas prácticas buscan reforzar la confianza institucional y garantizar que las decisiones impulsadas por la IA respeten los derechos fundamentales y los principios del Estado de derecho (Chambers and Partners, 2025), (Ministry of Finance Finland, 2025).

Además, el Servicio Penitenciario y de Libertad Condicional de Finlandia ha participado recientemente en el piloto de Aurora AI, un programa nacional de inteligencia artificial que funciona como sistema de recomendación de intinerarios de inserción social, ayudando a la ciudadanía a encontrar servicios públicos necesarios y relevantes a través de una plataforma común de proveedores de servicios públicos (Prison and Probation Service of Finland, 2023). El desarrollo de Aurora AI ha sido posible gracias a la colaboración del Instituto VTT de Investigación Técnica de Finlandia y diversas universidades, lo que garantiza que los algoritmos se basen en los últimos avances científicos y se alineen con prácticas fundamentadas en la evidencia científica (Ministry of Finance Finland, 2019).

En línea con los principios de la Estrategia Nacional de Inteligencia Artificial de Finlandia y con las directrices del Consejo de Europa, el uso de estas herramientas debe ser siempre transparente, centrado en la persona y sometido a evaluaciones continuas. Esto incluye evitar cualquier forma de discriminación automatizada, garantizar la supervisión humana y establecer mecanismos de revisión ética (Comisión Europea, 2021).

En el marco de un proyecto piloto desarrollado por la Criminal Sanctions Agency de Finlandia, personas privadas de libertad participaron activamente en tareas de etiquetado de datos para el desarrollo de sistemas de inteligencia artificial, como parte de una iniciativa para modernizar el trabajo penitenciario. Esta participación no solo les permitió adquirir habilidades digitales relevantes para el mercado laboral, sino que también ofreció una experiencia más significativa y motivadora en comparación con las tareas carcelarias tradicionales. El proyecto subraya el potencial de incluir a las personas presas en procesos vinculados al desarrollo y aplicación de tecnologías emergentes, promoviendo un enfoque más participativo e inclusivo (Criminal Sanctions Agency , 2023).

2.2. PAÍSES BAJOS

En Países Bajos los servicios de libertad condicional (probation) utilizan desde el 2019 el Oxford Risk of Recividism (OxRec) como herramienta principal para medir el riesgo de reincidencia de individuos en el sistema penal,

que forma parte de una serie de calculadores del riesgo desarrollados por la Universidad de Oxford, llamados OxRisk[9].

OxRec forma parte del "Risk Assessment Scales instrument" (RISc), una herramienta diagnóstica de estos servicios que valora la probabilidad de reincidencia, necesidades criminógenas y permite formular planes de supervisión (Maša Galic, 2023).

La OxRec fue desarrollada en la Universidad de Oxford, y externamente validada en Suecia, y en Paises Bajos. Además también se ha adaptado a contextos como Noruega y Tayikistan. Los desarrolladores del proceso de validación holandesa de esta herramienta (Fazel S. S., 2022) destacan que la OxRec está pensada para complementar la valoración profesional y toma de decisiones de los agentes socio-jurídicos, no sustituirla. Comenzaron con la validación del contexto sueco, entre otros factores, por la disponibilidad de los datos. Suecia cuenta con grandes muestras de datos, que contemplan factores de riesgo de diferentes dominios como el socio-demográfico, clínico, criminógeno (experiencia criminal), familiar, etc. (Fazel S. , 2025)[10].

OxRec permite que cada caso se valore en igualdad de condiciones utilizando la misma serie de variables estáticas y dinámicas, correlacionadas con la reincidencia violenta.

Esta herramienta utiliza conjuntos de datos (datasets) muy grandes. Predefine los predictores y los resultados, analiza los predictores mediante modelos multivariables, y reporta un amplio rango de medidas de rendimiento (performance measures), como las medidas de discriminación (lo

9. Vid. en https://oxrisk.com/. OxRisk es un proyecto del grupo de psiquiatría y psicología forense de la Universidad de Oxford, dedicado al desarrollo de calculadores del riesgo de acceso gratuito y de uso online. En este epígrafe nos hemos centrado en el OxRec, pero también han desarrollado los siguientes calculadores del riesgo: OxMIV — Herramienta sobre Trastorno Mental y Violencia, OxMIS — Herramienta sobre Trastorno Mental y Suicidio, FoVOx — Herramienta de Psiquiatría Forense y Violencia, FoxWeb — Herramienta de monitorización de la violencia en pacientes ingresados, OxRIS — Herramienta para el riesgo de reincidencia en delincuentes sexuales, OxDOV — Herramienta para el riesgo de reincidencia en personas detenidas por violencia doméstica, OxSATS — Herramienta para el riesgo de suicidio tras autolesiones y OxSET –Herramienta para la repetición de autolesiones.

10. El 17-18 de noviembre de 2025 un miembro del equipo de investigación del IVAC-KREI participó en el "Simposio internacional sobre evaluación del riesgo de violencia", organizado por UIK, en Donostia-San Sebastián. Los contenidos y debates desarrollados en dicho simposio se incorporan en este artículo como complemento a la revisión bibliográfica realizada.

bien que distingue la herramienta entre los que reinciden y los que no) y medidas de calibración (la cercanía de las puntuaciones entre las estimaciones de reincidencia y los datos reales). Los predictores incluyen aspectos como la edad, salario, estado civil, historial delictivo, salud mental, consumo de drogas, y factores sociodemográficos (Fazel S. , 2025).

La OxRec se basa en 14 factores, y se lleva a cabo a la hora de pensar en la excarcelación u obtención de permisos de la persona privada de libertad, y una de sus finalidades es facilitar la vinculación de estas personas con servicios comunitarios que faciliten su reinserción social, y por tanto disminuyan las posibilidades de reincidencia delictiva, como son servicios de tratamiento de drogodependencias y alcohol, atención en salud mental, seguimiento social, etc. (Fazel S. W.-M., 2019).

OxRec incluye probabilidades de conducta violenta y reincidencia en un plazo de 1 a 2 años, en intervalos de riesgo bajo, medio o alto. Consideran que estas categorías resultan más robustas que las numéricas, ya que cuando la herramienta no está correctamente calibrada se corre el riesgo de sobreestimar el riesgo de reincidencia de aquellos con alto riesgo (Fazel S. S., 2022). Aunque el propio Fazel, señaló en 2025 la importancia de cómo se comunican los resultados de las herramientas actuariales, o del juicio clínico. Desde esta perspectiva, advierte que el uso de categorías como "riesgo bajo", "medio" o "alto" no está exento de problemas, ya que dichas etiquetas pueden ser interpretadas de manera diferente por los distintos actores implicados en la toma de decisiones, favoreciendo lecturas reificadas del riesgo y efectos prácticos no deseados (Fazel S. , 2025).

Una de las principales ventajas de la OxRec es que fue desarrollada basándose en un protocolo predefinido en el cual se describieron las variables predictoras y cómo estas iban a ser testeadas y categorizadas antes de que se realizase el estudio estadístico (Fazel S. W.-M., 2019). Este enfoque aporta un mayor nivel de transparencia metodológica, lo que permite conocer cómo las variables contribuyen a la predicción de la reincidencia, fortaleciendo así la legitimidad técnica del instrumento (Van Dijck, 2022).

En la práctica, esta herramienta se aplica en unos 10-15 minutos, y sus desarrolladores manifiestan no ser necesaria una formación intensa, sino que con un video explicativo es suficiente, ni la realización de una entrevista con el preso. Los profesionales de los servicios de libertad condicional realizan una estimación del riesgo de reincidencia en paralelo a la estimación de la OxRec, en la que tienen en cuenta las circunstancias individuales del preso. Es el juez el que toma la decisión final, la cual puede diferir de la valoración aportada por la OxRec.

La OxRec es un instrumento basado en la evidencia científica, que cuenta con un buen performance cuya implementación en un nuevo contexto se considera simple. Así, la OxRec está diseñada para que sea fácil de usar, escalable, de uso online, gratuita y transparente. De esta forma, el proceso de implementación de la OxRec debería incluir una validación externa rigurosa hasta estar adecuadamente dimensionada, empleando los indicadores de predicción reales de la herramienta en lugar de marcadores indirectos, y adaptándose a diversos contextos del sistema penal. A esto debería seguirle una serie de estudios complementarios que aborden el análisis coste-beneficio (incluyendo tanto el tiempo como los costes directos, como el coste unitario de su aplicación), el estudio de posibles daños colaterales, y una evaluación de cómo las puntuaciones de riesgo pueden traducirse en intervenciones concretas o ajustes en la gestión del riesgo individual. Estos procesos son fundamentales para garantizar que las estimaciones algorítmicas no solo sean técnicamente precisas, sino también socialmente útiles y legítimas dentro de la práctica penal (Fazel et al., 2022; Van Dijck, 2022). Para una adecuada implementación de estas herramientas se necesitan referentes locales (local champions) que apuesten y lideren el proceso de adaptación e implementación local (Fazel S. , 2025).

2.3. ESTONIA

Según la "Confederation of European Probation" (2017) el sistema de predicción de reincidencia utilizado en Estonia es un modelo revisado del OASyS (Ofender Assessment System) utilizado en el Reino Unido. En Estonia se implementó en los servicios de libertad condicional en el año 2007, con posteriores procesos de estandarización en el 2011 y categorización de los presos según el riesgo de reincidencia en el año 2014, cuando también se implementó el modelo de cooperación británico MAPPA (Multy-Agency Public Protection Arrangements).

OASys combina métodos estadísticos predictivos con el "juicio profesional estructurado" para realizar evaluaciones estandarizadas del riesgo y las necesidades de los infractores, con el fin de generar planes individualizados de sentencia y gestión del riesgo, e identificar intervenciones adecuadas. Evalúa la probabilidad de reincidencia, el riesgo de daño a otras personas o a sí mismas, y las necesidades individuales, produciendo una puntuación de riesgo: bajo, medio, alto o muy alto (Fair Trials, 2021). Incluye dos predictores algorítmicos clave:

- OGP (Predicción General de Reincidencia): riesgo de delitos no violentos.

- OVP (Predicción de Violencia): riesgo de delitos violentos.

Estos predictores se construyen a partir de la evaluación OASys, junto con información del historial delictivo y variables sociodemográficas. Las evaluaciones disponibles sugieren la existencia de diferencias en el rendimiento predictivo por género y origen étnico, con una mayor precisión en mujeres, personas blancas y de mayor edad, y un rendimiento comparativamente inferior en personas negras y de origen étnico mixto.

Como alternativa, también se usa OGRS (Offender Group Reconviction Scale), que predice la probabilidad de reincidencia en dos años (en general o violenta), usando edad, género, sanciones previas, edad al primer delito y tasa delictiva estimada (COPAS— Criminal Offending Profile and Assessment System).

En los últimos años, Estonia ha priorizado la transformación digital de su sistema penitenciario y judicial, desarrollando una estrategia nacional de justicia digital basada en el uso de IA, interoperabilidad entre sistemas judiciales y evaluación algorítmica de riesgos (Confederation of European Probation, 2017).

Entre sus principales retos se encuentra la necesidad de asegurar la calidad de los datos penitenciarios, evitar la discriminación algorítmica, y garantizar la responsabilidad (acknowledgement) en las decisiones automatizadas. Aunque el sistema OASyS ha sido adaptado localmente, su implementación sigue generando debate sobre la fiabilidad y equidad de los resultados, particularmente en poblaciones vulnerables. Estonia continúa evaluando la posibilidad de incorporar evaluaciones del impacto algorítmico y mecanismos de revisión humana sistemática como parte de sus futuras reformas (Fair Trials, 2021).

2.4. GRECIA, CHIPRE, PORTUGAL Y BULGARIA

Dado el uso creciente de la IA en la toma de decisiones dentro del ámbito de la justicia penal, se han ido detectando necesidades de desarrollar sistemas de IA libre de sesgos para predecir el riesgo de reincidencia. FAIR-PReSONS es un proyecto europeo de innovación aplicada y fortalecimiento institucional, financiado en el marco del Programa de Justicia de la Unión Europea, Dirección General de Justicia (DG JUST), y coordinado por la Universidad del Egeo (Grecia) que comenzó en junio de 2024 y tiene previsto finalizar en mayo de 2026 (FAIR — PReSONS, 2025). Su orientación es eminentemente transnacional, colaborativa y basada en la evidencia científica,

ya que implica la cooperación de distintos actores judiciales, penitenciarios, tecnológicos y académicos de varios países europeos.

Esta iniciativa reconoce los riesgos que suponen los sesgos inherentes a los datos y algoritmos. Variables como el género, la raza y la nacionalidad pueden influir de forma significativa en las decisiones impulsadas por IA, lo que plantea preocupaciones en torno a la equidad, la justicia y el respeto a la perspectiva de género (Innovative Criminal Justice Solutions, 2025). Asimismo, se busca asegurar la equidad y la no discriminación en todo el proceso, diseñar e implementar actividades de fortalecimiento de capacidades para los actores implicados y promover la difusión de los resultados junto con acciones de sensibilización que fomenten una aplicación ética y transparente de la tecnología.

Para conseguir dichos objetivos, el proyecto se centra en la recolección y digitalización de datos procedentes de prisiones y sistemas de gestión de delitos (Offense Management Systems — OMS) en los países participantes —Grecia, Chipre, Portugal y Bulgaria—, estructurando dicha información en gráficos de conocimiento (knowledge graphs) y haciéndola accesible a través de un portal de datos especializado. FAIR-PReSONS combina metodologías basadas en redes neuronales artificiales (ANNs), gráficos de conocimiento y análisis multivariante, integrando auditorías algorítmicas y evaluación de sesgos desde una perspectiva de género (Innovative Criminal Justice Solutions, 2025).

La recopilación de esta amplia gama de datos nacionales, que incluyen diversas variables, permitirá comprender la relación entre estas variables y el riesgo de reincidencia, así como identificar los factores más influyentes en cada contexto. En última instancia, este proceso servirá como base para determinar qué intervenciones de rehabilitación deben priorizarse para reducir de manera más efectiva el riesgo de reincidencia.

Se llevará a cabo un análisis preliminar con enfoque de género para identificar y abordar posibles sesgos de género en los sistemas de predicción de reincidencia. Este enfoque destaca los requisitos del Reglamento Europeo sobre Inteligencia Artificial, aplicando principios clave como trazabilidad, explicabilidad, supervisión humana significativa y evaluación de impacto algorítmico, que son esenciales para sistemas catalogados como "de alto riesgo" (Comisión Europea, 2021).

Finalmente, los estándares generados por el proyecto buscan extenderse más allá de los países socios, con el objetivo de promover una convergencia europea en la evaluación justa del riesgo de reincidencia mediante IA

(FAIR — PReSONS, 2025). Este enfoque holístico de la "IA justa" aspira a fomentar decisiones no sesgadas, brindando apoyo a jueces y profesionales del derecho en su esfuerzo por tomar decisiones imparciales y mitigar los estereotipos de género.

2.5. SUIZA

Aunque Suiza no forma parte de la Unión Europea, se ha decidido incluir este país en el análisis debido a la longevidad y solidez de su sistema de evaluación del riesgo de reincidencia. En concreto, el programa ROES (Risk-Oriented Enforcement of Sentence) ha sido pionero en la región y ha influido en la práctica penal de otros países de habla germana como Austria y Alemania, lo que justifica su relevancia comparativa. De aplicación obligatoria en los cantones germanoparlantes de Suiza, con planes de extensión a las zonas francófonas y al resto del país (Goncalves, 2017).

El ROES ayuda a orientar la ejecución de las penas privativas de libertad en función del riesgo individual de reincidencia (Rossegger, 2008). Dentro de este marco hay una serie de recursos penitenciarios, y el encargado de valorar el riesgo de reincidencia es el FOTRES (Forensic Operationalized Therapy/Risk Evaluation System). Esta herramienta es utilizada para valorar el riesgo de reincidencia (RNA, Risk-need assessment) y evaluar el progreso del proceso de reinserción del preso (RM, Risk Management). Mientras que la valoración RNA solo se realiza una vez, normalmente cuando el preso entra en prisión, la valoración RM se realiza de forma periódica cada vez que el riesgo de reincidencia tiene que ser valorado (se recomienda una vez al año), como por ejemplo en caso de solicitud de la libertad condicional.

Es un instrumento profesional estructurado que ofrece valores de riesgo de reincidencia respecto a tipos delictivos específicos, pero no ofrece temporalidad del riesgo. El módulo de la RNA analiza 3 grupos de variables; 1— Riesgo de base (perfil de riesgo y relevancia del perfil de riesgo), 2— Verosimilitud del riesgo de base (patrón de la conducta criminal y disposiciones de la personalidad en relación con la conducta criminal), y 3— Tratabilidad de base (perspectivas de tratamiento y recursos personales). En total forman 175 ítems, y ofrecen una valoración de riesgo del 0 (riesgo muy bajo) al 4 (riesgo muy alto) (Räz T. , 2022). El módulo RM incluye la valoración del riesgo actual (autocontrol, perfil de riesgo y relevancia del perfil de riesgo), tratabilidad actual (perspectivas de tratamiento y recursos personales) y los factores situacionales actuales (factores protectores y de riesgo situacional). Los resultados son calculados automáticamente por la

aplicación web, su aplicación requiere un tiempo estimado de 60 minutos. No ofrece información sobre ratios de reincidencia penitenciaria general.

Una característica distintiva de FOTRES es su vocación terapéutica: no solo identifica riesgos, sino que establece un marco para planificar intervenciones específicas orientadas al tratamiento, la supervisión y la reducción del riesgo. Además, el sistema busca evitar automatismos: requiere la interpretación clínica de profesionales especializados, lo que contribuye a mantener una lógica de evaluación personalizada y adaptativa en función del progreso del recluso (Räz C. , 2022).

3. CONSIDERACIONES ÉTICAS Y LEGALES A TENER EN CUENTA PARA SU IMPLEMENTACIÓN EN LA CAPV

La competencia en materia penitenciaria se traspasó a la Comunidad Autónoma del País Vasco (CAPV) en octubre de 2021, por lo que se consolidó un nuevo modelo de gestión autonómica a través de la Ley 7/2021 y su desarrollo organizativo posterior (Gobierno Vasco, 2022).

Las Juntas de Tratamiento (equipo multidisciplinar) elaboran informes y propuestas sobre permisos, progresiones de grado y libertad condicional/suspensión, que son resueltas por la autoridad penitenciaria o por el Juzgado de Vigilancia Penitenciaria según el caso. Estas decisiones son individualizadas y se fundamentan en la LOGP (arts. 72-76, 100) y el Reglamento Penitenciario (arts. 102 y ss.), valorando evolución tratamental, apoyo social, conducta intramuros y riesgo de reincidencia mediante juicio profesional estructurado, pudiendo emplear escalas validadas como apoyo pero sin sustituir la deliberación clínica y jurídica (Gobierno Vasco, 2021).

A la hora de pensar en herramientas para predecir el riesgo de reincidencia de una población penitenciaria, como puede ser la de la CAPV, hay que conocer a dicha población para diseñar las herramientas acordes a las características de esta, así como lo que se está haciendo en los contextos más cercanos. Según los datos estadísticos proporcionados por instituciones penitenciarias del gobierno vasco, en mayo de 2025 la población penitenciaria de la CAPV era de 1785 personas privadas de libertad, un 33% extranjeros. En libertad condicional se encontraban 161 personas, 88% hombres y 12% mujeres. 481 personas en tercer grado. Menores de 25 (10%), entre 26 y 40 (41%), entre 41 y 60 (43%) y mayores de 61(5%)

La CAPV, no es completamente ajena al uso de herramientas algorítmicas para la predicción de conductas criminales, ya que la Ertzaintza utiliza la

plataforma algorítmica EBA (Emakumeen eta Etxekoen Babesa) para calcular el riesgo de revictimización de una víctima de violencia de género y así asignar medidas de protección adecuadas a dicho riesgo. La herramienta utilizada es la EPV-R, formada por 20 preguntas. Una puntuación mayor de 24 supone riesgo severo, mayor de 18 riesgo alto, mayor de 10 riesgo medio y menor de 9 riesgo bajo.

En un estudio preliminar llevado a cabo en 2022 (Valdivia, 2022) sobre la EBA, uno de los riesgos detectados es la presencia de falsos negativos, es decir, obtener una puntuación de riesgo bajo en casos severos. Además, se han encontrado limitaciones tanto en su diseño como en su aplicación práctica. Por un lado, incorpora factores subjetivos y culturalmente sesgados, como la nacionalidad entendida desde una perspectiva occidental, y juicios poco estructurados como la percepción del "grado de celos" del agresor, lo que introduce un riesgo considerable de discriminación. La alta tasa de falsos negativos pone en peligro la seguridad de las víctimas al no garantizar un nivel de protección adecuado. Esta combinación de subjetividad, opacidad y baja precisión plantea serias dudas sobre la fiabilidad del algoritmo y su impacto en los derechos fundamentales de las personas evaluadas (Bellio, 2023).

En otro ejemplo cercano, en Cataluña, el sistema judicial emplea el algoritmo RisCanvi para evaluar el riesgo de reincidencia y asistir en decisiones sobre permisos penitenciarios y libertades condicionales. Aunque se subraya que la decisión final recae en profesionales humanos, en el 95 % de los casos coincide con el resultado del algoritmo, lo que genera preocupación por una posible dependencia acrítica. El sistema es utilizado diariamente por unos 750 profesionales, de los cuales 100 actúan como validadores. Aunque no se han detectado sesgos evidentes en el funcionamiento del algoritmo, se critica la falta de una Evaluación de Impacto Algorítmico (AIA), especialmente porque la normativa europea clasificaría este tipo de sistemas como de alto riesgo (Bellio, 2021).

Los resultados del estudio llevado a cabo con respecto a la efectividad del RisCanvi (Portela, 2024), arrojan que la precisión de la valoración del riesgo determinada por humanos mejora tras conocer la sugerencia de la herramienta algorítmica. Esta mejora es visible a lo largo del tiempo, después de evaluar un número alto de casos, lo que sugiere que el apoyo algorítmico ayuda a fijar puntos de referencia para las predicciones humanas. Los participantes del estudio manifiestan preferir una asistencia parcialmente automatizada, pero con un grado amplio de juicio humano, un elemento de negociación que debe tratarse con cuidado y sin considerar la valoración

algorítmica como un dato objetivo. Los participantes prevén que la automatización de este tipo de procesos va a ser cada vez más prevalente, pero prefieren el criterio humano.

Como parte de esta investigación se ha realizado, entre otras, una entrevista a una persona Técnica de la Secretaría General de Instituciones Penitenciarias españolas, donde se reconoce el valor de estas herramientas como instrumentos de estandarización. La técnica entrevistada señala que la sistematización de procedimientos puede tener efectos positivos en términos de garantías y coherencia institucional, destacando que *"sistematizar los procedimientos hace el sistema más garantista"*. Además, pone de manifiesto diversas limitaciones estructurales del sistema penitenciario que condicionan el uso eficaz de estas herramientas, entre ellas la falta de diagnósticos clínicos fiables, la colinealidad entre variables y la escasa coordinación entre áreas profesionales. Como resume la técnica, *"la realidad de los centros es que normalmente las personas vienen sin diagnóstico, con diagnósticos que se solapan…"*. Todo ello refuerza la necesidad de un rigor metodológico estricto y de una revisión crítica constante en el diseño y uso de las herramientas de evaluación del riesgo, especialmente en un contexto de creciente confianza institucional en los algoritmos. En palabras de la propia entrevistada, *"tenemos que ser rigurosos para que la construcción de la herramienta sea metodológicamente pura"*. Al mismo tiempo, la técnica reconoce el valor del juicio clínico sobre los resultados automatizados (de la TVR), sobre todo en casos en los que se dan variables atípicas que aunque tienen un alto valor pronóstico, suelen quedar excluidas de los modelos actuariales debido a su baja frecuencia estadística, *"hay variables que son muy atípicas […] pero cuando se dan, tienen mucho peso pronóstico, […] por ejemplo el abuso sexual en prisión […] no está contemplado en los modelos, lo que limita la sensibilidad de la herramienta ante situaciones especialmente graves"*. Otro aspecto relevante es la necesidad de incorporar factores de protección en los modelos de evaluación. La técnica critica que la mayoría de las herramientas se centren exclusivamente en factores de riesgo y defiende una aproximación más equilibrada, señalando que *"hay que incluir factores de protección que compensen y que en el algoritmo sumen o resten respecto del resultado"* y que la mera ausencia de factores de riesgo no significa la presencia de factores de protección.

La responsabilidad (acknowledgement) algorítmica representa un aspecto clave en la implementación de herramientas predictivas del riesgo de reincidencia en el ámbito penal. Tal como señala Wieringa (2020), esta responsabilidad debe abordarse desde una perspectiva sistémica, considerando

a múltiples actores (desarrolladores, operadores y responsables de decisiones) que deben justificar y explicar su papel ante distintos foros —institucionales, legales, sociales— a lo largo de todo el ciclo de vida del sistema. Utilizando el marco propuesto por Bovens, se identifican cinco elementos fundamentales: el actor, el foro, la relación, el contenido y las consecuencias. Si bien la transparencia y la explicabilidad de los algoritmos son necesarias, resultan insuficientes sin mecanismos efectivos de supervisión, evaluación y sanción. Entre los riesgos identificados se encuentran la ambigüedad en las responsabilidades, la falta de control efectivo, y fenómenos como el "virtue-washing" —la apariencia de responsabilidad ética sin mecanismos reales de control. Para garantizar un uso legítimo y ético de estos sistemas en la CAPV, se requiere una arquitectura de gobernanza clara, interdisciplinar y con mecanismos de evaluación pública, que respete tanto los derechos fundamentales de las personas evaluadas como las obligaciones institucionales de transparencia y equidad.

4. CONCLUSIONES

El diseño de algoritmos en el contexto penitenciario requiere de una reflexión deliberada sobre qué es exactamente lo que se quiere conseguir y una comprensión sólida sobre qué fallas humanas se pretende suplir (Puolakka et al., 2021).

Todas las herramientas algorítmicas predictivas analizadas en este artículo —OxRec, RISc, OASys, OGRS, FOTRES y RisCanvi— son modelos estadísticos basados en variables predefinidas y métodos multivariables. No se ha identificado en la Unión Europea un uso consolidado de modelos basados en *machine learning* para predecir la reincidencia, la única excepción identificada en un contexto cercano a la UE es el sistema Cassandra del Ministerio de Justicia de Ucrania, que incorpora elementos de *machine learning*, analizando datos de múltiples fuentes estatales para aprender patrones y estimar el riesgo de reincidencia. Una de las principales cuestiones que surgen en este contexto es la fiabilidad de estos sistemas. Mayowa et al. (2023) analizan la fiabilidad "*trustworthiness*" de la IA aplicada a la evaluación post-condena, destacando que la legislación europea —en especial el Reglamento de IA de la UE— clasifica este tipo de herramientas como de alto riesgo, exigiendo supervisión humana significativa para cumplir con la normativa. Esto implica que el juicio profesional no puede ser sustituido por la valoración algorítmica, sino que debe integrarse como un apoyo sujeto a revisión crítica.

El análisis comparado ofrece aprendizajes relevantes para la CAPV:

- Finlandia: entiende la precisión no solo como exactitud estadística, sino como capacidad de satisfacer necesidades psicosociales de reinserción del preso; incorpora mecanismos de control y revisión; y ha desarrollado proyectos piloto de co-participación con personas privadas de libertad.
- Proyecto FAIR-PReSONS: incluye perspectiva de género y estrategias de mitigación de sesgos en todas las fases del ciclo de vida del algoritmo.
- Países Bajos: destaca la transparencia metodológica y la vinculación directa de la predicción con planes de supervisión individualizados.
- Suiza: el sistema FOTRES combina evaluación de riesgo con vocación terapéutica, orientando intervenciones de rehabilitación más allá de la medición numérica del riesgo.
- Estonia: la adaptación local del OASys combina juicio profesional estructurado con predictores algorítmicos (OGP, OVP y OGRS). Sus retos incluyen garantizar la calidad y actualización de los datos, evitar sesgos por género o etnia y establecer mecanismos sistemáticos de revisión humana.

En este sentido, resulta crucial actualizar periódicamente la matriz de datos utilizada para el diseño del algoritmo, asegurando una calidad optima de los datos que refleje la realidad penitenciaria actual y evitando la perpetuación de sesgos históricos. Esta tarea debe incluir mecanismos de auditoría de datos y procesos de validación externa.

Asimismo, la gobernanza de estos sistemas debe contemplar una arquitectura clara de responsabilidades (Wieringa, 2020), que asegure que desarrolladores, operadores y responsables de decisiones puedan rendir cuentas en foros institucionales, legales y sociales. La colaboración con instituciones académicas —como en el caso del proyecto *FAIR PReSONS*— añade rigor metodológico y una revisión independiente, lo que aumenta la legitimidad y transparencia del sistema. La técnica de la Secretaría General de Instituciones penitenciares señala la importancia de la colaboración entre la administración penitenciaria y el ámbito académico. La institución debe abrirse a los avances científicos y metodológicos, afirmando que *"la institución tiene que ser permeable […] las aplicaciones prácticas se tienen que nutrir de los desarrollos académicos"*.

En línea con los principios de *Fair AI*, resulta imprescindible garantizar una transparencia efectiva sobre el funcionamiento interno de los algoritmos, de forma que se pueda conocer con claridad cómo y por qué se toman las decisiones que afectan a las personas privadas de libertad. Esta transparencia no solo es un requisito ético, sino una condición para salvaguardar la seguridad jurídica de los presos y para que las decisiones puedan ser auditadas, impugnadas y corregidas cuando sea necesario. El acceso a explicaciones comprensibles sobre los mecanismos de cálculo, las variables utilizadas y el peso asignado a cada una de ellas favorece un control institucional más sólido, reduce el riesgo de automatismos punitivos y fortalece la legitimidad democrática de estos sistemas.

En el caso vasco, la implementación de una herramienta de predicción del riesgo de reincidencia podría plantearse como un proyecto piloto con evaluación externa, análisis coste-beneficio y participación de las Juntas de Tratamiento. Debería adaptarse al tamaño reducido de la población penitenciaria y a sus características demográficas y criminógenas incorporando protocolos de revisión periódica, mecanismos de participación de las personas internas y planes de supervisión que trasciendan el simple resultado numérico del riesgo. El objetivo final sería asegurar que el uso de herramientas actuariales contribuya a los fines de reinserción social y no derive en automatismos punitivos.

5. REFERENCIAS

Bellio, N. (30 de marzo de 2021). *In Catalonia, an algorithm helps decide who can leave prison temporarily — and professionals follow its lead.* Obtenido de Algorithm Watch: https://algorithmwatch.org/en/riscanvi/.

Bellio, N. (30 de marzo de 2023). *Basque Country: how an algorithm to assess the risk of gender-based violence sees people from "different cultures".* Obtenido de Algorithm Watch: https://algorithmwatch.org/en/epv-violence-algorithm-basque-country-2023/

Blanchette, K. (2002). Classifying female offenders for effective intervention: Application of the case-based principles of risk and need. *Forum on Corrections Research*, 31-35.

Chambers and Partners. (2025, mayo 22). *Trends and Developments: Artificial Intelligence 2025 — Finland*. Retrieved from Chambers and Partners: httpshttpshttps://valtioneuvosto.fi/en/-/study-now-is-the-time-to-ad-

dress-algorithmic-discrimination-while-the-deployment-of-ai-systems-in-finland-is-still-at-a-modest-level.

Comisión Europea. (2021). *AI Watch: National strategies on Artificial Intelligence — A European perspective. 2021 edition.* Bruselas: Publications Office of the European Union.

Comisión Europea. (2021). *Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts..* Bruselas: Comisión Europea.

Confederation of European Probation. (16 de agosto de 2017). *Professionalising probation in Estonia*. Obtenido de Confederation of European Probation: https://www.cep-probation.org/professionalising-probation-in-estonia/.

Criminal Sanctions Agency. (2023). *risoners Involved in the Development of Artificial Intelligence — Piloting New Prison Work.* Helsinki: Criminal Sanctions Agency of Finland.

FAIR — PReSONS. (2025). *FAIR — PReSONS*. Retrieved from https://fair-presons.aegean.gr/

Fair Trials. (2021). *Automating injustice: The use of artificial intelligence and automated decision-making systems in criminal justice in Europe.* Londres: Fair Trials.

Fanega, M. F. (2024). PPublicación del algoritmo penitenciario TVR. *Revista Española de Investigación Criminológica*, 22(1), e917.

Farayola, M. M. (2023). Fairness of AI in predicting the risk of recidivism: Review and phase mapping of AI fairness techniques. *Proceedings of the 18th International Conference on Availability, Reliability and Security (ARES 2023)* (p. https://doi.org/10.1145/3600160.3605033). Benevento: Association for Computing Machinery (ACM).

Fazel, S. (2025). The OxRisk project and how to improve management in clinical context. Seena Fazel, University of Oxford. *Simposio internacional sobre evaluación del riesgo de violencia. 17-18 De noviembre 2025.* Donostia-San Sebastián: UIK.

Fazel, S. S. (2022). Towards a more evidence-based risk assessment for people in the criminal justice system: the case of OxRec in the Netherlands. *European Journal on Criminal Policy and Research* (https://doi.org/10.1007/s10610-022-09520-y), 397-406.

Fazel, S. W.—M. (2019). Prediction of violent reoffending in prisoners and individuals on probation: a Dutch validation study (OxRec). *Scientific Reports, 9*(https://doi.org/10.1038/s41598-018-37539-x).

Gobierno Vasco. (2021). *Modelo penitenciario de Euskadi: Una gestión penitenciaria desde la comunidad, con la persona como centro.* Vitoria-Gazteiz: Dirección de Justicia, Departamento de Igualdad, Justicia y Políticas Sociales.

Gobierno Vasco. (2022). *Estrategia Vasca de Justicia Restaurativa 2022-2025.* Vitoria-Gazteiz: Departamento de Igualdad, Justicia y Políticas Sociales.

Gobierno Vasco. (mayo de 2025). *Tablas estadísticas de la Estadística Mensual de la Población Reclusa — mayo 2025*. Obtenido de Euskadi.eus: https://www.euskadi.eus/estadistica/tablas-estadisticas-de-la-estadistica-mensual-de-la-poblacion-reclusa-mayo-2025/web01-s2enple/es/.

Goncalves, L. C. (2017). Forensic Operationalized Therapy/Risk Evaluation System (FOTRES). En A. J. A. R. Beech, *The Wiley Handbook on the Theories, Assessment and Treatment of Sexual Offending* (pág. https://doi.org/10.1002/9781119184256.ch12).

Innovative Criminal Justice Solutions. (6 de febrero de 2025). *Understanding recidivism and exploring AI's potential in reducing reoffending*. Obtenido de Innovative Criminal Justice Solutions: https://prisonsystems.eu/understanding-recidivism-and-exploring-ais-potential-in-reducing-reoffending/.

Maša Galic, A. D. (2023). *AI and administration of criminal justice: Report on the Netherlands.* Rotterdam: Erasmus University Rotterdam.

Ministry of Finance Finland. (2019). *AuroraAI development and implementation plan 2019-2023.* Helsinki: Ministry of Finance Finland.

Ministry of Finance Finland. (2025). *Guidelines on using generative AI in public administration.* Helsinki: Ministry of Finance Finland.

Portela, M. C.—H. (2024). Comparative user study of human predictions in algorithm-supported recidivism risk assessment. *Artificial Intelligence and Law* (https://doi.org/10.1007/s10506-024-09393-y), 471-517.

Prison and Probation Service of Finland. (2 de marzo de 2023). *Development of the Prison and Probation Service of Finland focused on underage, young and female prisoners in 2022*. Obtenido de RISE: https://www.rikosseuraamus.fi/en/index/topical/pressreleasesandnews/2023/developmentofthepri-

sonandprobationserviceoffinlandfocusedonunderageyoungandfemaleprisonersin2022.html

Puolakka, P. &. (2021). Artificial Intelligence in Prisons in 2030: An exploration on the future of AI in prisons. *Advancing Corrections Journal, 11*.

Puolakka, P. (17 de abril de 2023). *Smart prisons and Artificial Intelligence systems expand in Finland*. Obtenido de Justice Trends Magazine: https://justice-trends.press/smart-prisons-and-artificial-intelligence-systems-expand-in-finland/.

Räz, C. (2022). *FOTRES: Grundlagen und praktische Anwendung in der forensischen Risikodiagnostik.* Zurich: Zürcher Hochschule für Angewandte Wissenschaften (ZHAW).

Räz, T. (2022). Understanding risk with FOTRES? *AI and Ethics, 3*(https://doi.org/10.1007/s43681-022-00223-y), 1153-1167.

Rolland, A. (26 de julio de 2021). *The Security Distillery.* Obtenido de https://thesecuritydistillery.org/all-articles/ethics-artificial-intelligence-and-predictive-policing

Rossegger, A. E. (2008). Risk-based allocation in the Swiss correctional system: From assessment to enforcement. *International Journal of Law and Psychiatry, 31*(5), 484-491.

Valdivia, A. H.—V.—M. (2022). Judging the algorithm: A case study on the risk assessment tool for gender-based violence implemented in the Basque Country. *AI & Society*, https://doi.org/10.1007/s00146-022-01486-w.

Van Dijck, G. (2022). Predicting Recidivism Risk Meets AI Act. *European Journal on Criminal Policy and Research, 28(3)*, 407-423.

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. *2020 Conference on Fairness, Accountability, and Transparency (FAT «20)* (págs. 1-18). Barcelona: Association for Computing Machinery.

Yukhnenko, D. B. (2020). Risk factors for recidivism in individuals receiving community sentences: a systematic review and meta-analysis. *CNS Spectrums, 25*(2), 252-263.

PARTE II: RisCanvi

¿Qué es? ¿Cómo funciona?
¿Cómo se ha aplicado y aplica? ¿Qué normas
ético-jurídicas y técnico-estadísticas deben aplicarse?
¿Qué riesgos tiene?

Capítulo 5

El Protocolo RisCanvi para la Gestión del Riesgo en el Sistema Penitenciario de Cataluña

Dr. A. Andres-Pueyo

Universitat de Barcelona

SUMARIO: 1. INTRODUCCIÓN. 2. RISCANVI EN EL MODELO REHABILITADOR DE LAS PRISIONES EN CATALUÑA. 3. APLICACIONES DEL RISCANVI EN LA GESTIÓN PENITENCIARIA. 4. CONCLUSIONES. 5. REFERENCIAS.

1. INTRODUCCIÓN

La ejecución de la pena de prisión ha experimentado una transformación conceptual y práctica en las últimas décadas en numerosos países (Andrews y Bonta, 2010). Tradicionalmente, el sistema penitenciario se movía entre dos polos más o menos antagónicos. Por un lado, el castigo retributivo, heredero directo de las concepciones clásicas del derecho penal y por otro, el ideal rehabilitador, que aspiraba a transformar al penado mediante el tratamiento y la reeducación. A finales del siglo XX emerge con fuerza un tercer paradigma que redefine los objetivos y la metodología de la institución penitenciaria: el modelo de la gestión del riesgo (Bonta y Andrews, 2017). Bajo este enfoque, que algunos autores han denominado "nueva penología" o "justicia actuarial" (Garland, 2001; Hannah-Moffat, 2005), el foco se desplaza desde la transformación moral y comportamental del individuo hacia la identificación, clasificación y gestión rehabilitadora de los delincuentes en prisión según su probabilidad de reincidencia futura.

La valoración del riesgo se ha convertido en un elemento central de la política criminal contemporánea, sobre todo en contextos anglosajones, pero también con una expansión creciente en Europa continental (Andres-Pueyo et al., 2018). En este marco de cambio paradigmático, el sistema penitenciario catalán se ha posicionado como pionero en el ámbito hispanohablante al introducir, en 2009, el protocolo RisCanvi. Se trata de una herramienta integral para la evaluación y gestión del riesgo de violencia y de reincidencia, fruto de la colaboración entre el Grup de Estudis Avançats en Violència (GEAV) de la Universidad de Barcelona y los servicios penitenciarios de la Generalitat de Cataluña (Andrés-Pueyo y Redondo, 2007; Andrés-Pueyo, Arbach-Lucioni y Redondo, 2018).

RisCanvi nació con una finalidad muy concreta, dotar a la administración penitenciaria de una herramienta de valoración adaptada a la población penitenciaria catalana, empíricamente validada en ese contexto y capaz de orientar tanto las decisiones de clasificación y tratamiento como la planificación de intervenciones individualizadas de los internos en las prisiones de Cataluña. Antes de su implantación, los profesionales para tomar decisiones de gestión sobre los internos, especialmente en todo aquello relacionado con el paso de éstos por la prisión, dependían en gran medida del juicio clínico no estructurado, un modelo —muy extendido en los entornos jurídicos y penales— que la investigación ha demostrado repetidamente inferior en términos de validez predictiva y consistencia (Ægisdóttir et al., 2006). El RisCanvi es un protocolo multiescala que se inserta así en un movimiento más amplio de mejora, profesionalización y estandarización de la práctica pericial y penitenciaria.

Para comprender el alcance del RisCanvi es necesario situarlo sobre el telón de fondo de grandes desafíos estratégicos en la política criminal, el mas importante, reducir y gestionar la reincidencia delictiva, pero también otros aspectos como la violencia intrainstitucional que se produce dentro de los centros penitenciarios (autodirigida e interpersonal) y el quebrantamiento de los permisos de los internos. El indicador central de la eficacia de la intervención penitenciaria se sitúa en el control y reducción de la reincidencia (Andrews y Bonta, 1990). Siendo la reincidencia penitenciaria violenta el primer objetivo de los sistemas de rehabilitación, esta lógica también se ha generalizado a la reincidencia general y otros indicadores de funcionamiento adecuado de la gestión penitenciaria. Así, es bien conocido que las instituciones penitenciarias son entornos altamente complejos, donde confluyen múltiples factores criminógenos y situacionales que favorecen la emergencia de conductas violentas en el entorno penitenciario. La violencia en prisión

se manifiesta, al menos, en tres dimensiones que conviene distinguir: la violencia interpersonal, que comprende las agresiones entre internos o hacia el personal penitenciario; la violencia intrapersonal, que se expresa en forma de autolesiones y tentativas de suicidio; y la violencia colectiva, que adopta la forma de motines, disturbios y otros episodios de alteración grave del orden (Cooke, 1991). Estudios europeos estiman que aproximadamente una cuarta parte de la población reclusa sufre algún tipo de victimización violenta cada año (Wolff et al., 2007). En Cataluña, los análisis de incidentes registrados muestran que una proporción relevante de los sucesos se corresponde con agresiones entre internos, lo que evidencia la persistencia de este fenómeno pese a los esfuerzos preventivos (Ferez-Mangas, D. y Andres-Pueyo, A., 2024; Capdevila et al., 2015).

La violencia penitenciaria no es aleatoria ni homogénea. Presenta patrones relativamente estables asociados a características individuales, a factores situacionales del entorno y a estilos de gestión institucional. La criminología ha desarrollado tres grandes perspectivas teóricas para explicar este fenómeno (Blevins et al., 2010).

- La teoría de la importación (Irwin y Cressey, 1962) sostiene que las conductas violentas observadas en prisión son, en gran medida, un reflejo de patrones preexistentes: los internos importan al establecimiento sus antecedentes de violencia, sus actitudes antisociales, sus déficits de habilidades sociales y sus consumos problemáticos. Desde esta óptica, la prisión actúa como espejo que reproduce trayectorias anteriores más que como origen de nuevas conductas.

- La teoría de la privación, formulada clásicamente por Sykes (1958), desplaza la explicación hacia las condiciones "socio-ecológicas" del internamiento. El hacinamiento, la pérdida de autonomía, la ruptura de vínculos significativos, el aburrimiento, la falta de oportunidades y la frustración sostenida generan un clima de estrés crónico que favorece la aparición de respuestas agresivas de distinto nivel de gravedad y tipología. La violencia, según esta aproximación, es aquí una forma de afrontamiento ante una situación percibida como intolerable.

- Finalmente, la teoría de la gestión o control administrativo (DiIulio, 1987) pone el foco en la organización: el estilo de liderazgo del equipo directivo, la calidad de la supervisión, la consistencia en la aplicación de normas, la legitimidad percibida por

la población interna y la capacidad de anticiparse a los conflictos tienen una incidencia directa sobre los niveles de violencia registrados.

La evidencia disponible sugiere que ninguna de estas perspectivas, por sí sola, resulta suficiente. La violencia en prisión emerge de la interacción compleja entre características individuales, factores de privación y estilos de gestión, y esta comprensión multifactorial constituye el fundamento conceptual del propio RisCanvi (Gendreau, Goggin y Law, 1997).

2. RISCANVI EN EL MODELO REHABILITADOR DE LAS PRISIONES EN CATALUÑA

El primer objetivo de las prisiones es, entre otros, evitar la reincidencia delictiva tras la excarcelación y esta visión —un poco reduccionista si se quiere— se ha consolidado como el indicador más visible del éxito o fracaso del proyecto rehabilitador. Diversos estudios en España han documentado tasas de reingreso en prisión que oscilan, según el período y la metodología, entre el 20 % y el 40 % en seguimientos de cinco años (Capdevila et al., 2015). Estas cifras, aunque relativamente mejores que las observadas en sistemas más punitivos como el estadounidense, donde la reincidencia puede superar el 50 % (Durose, Cooper y Snyder, 2014), ponen de manifiesto la dificultad de romper los ciclos delictivos.

La experiencia catalana aporta una serie histórica especialmente valiosa para el análisis empírico de la reincidencia penitenciaria. Desde 2008, el Centre d'Estudis Jurídics i Formació Especialitzada (CEJFE) sigue cohortes de excarcelados, midiendo la tasa de reincidencia penitenciaria, entendida como el porcentaje de personas que reingresan en prisión por una nueva condena en los cinco años posteriores a su salida (Capdevila et al., 2023). Los resultados muestran un descenso sostenido y notable, desde un 40,3 % en la cohorte de 2008 se pasa a un 21,1 % en la cohorte de 2015 (Capdevila et al., 2015). En términos comparativos, se trata de una de las reducciones más pronunciadas y bien documentadas en la literatura criminológica internacional. Expresado positivamente, alrededor de ocho de cada diez personas excarceladas en 2015 no volvieron a ingresar en prisión durante el período de seguimiento de cinco años (Capdevila, M. et al., 2023). Esta mejora no puede atribuirse a un único factor, pero coincide temporalmente con la implantación del RisCanvi, la ampliación de programas de tratamiento, el impulso de las salidas progresivas y el desarrollo de dispositivos de apoyo postpenitenciario.

Ahora bien, la reincidencia penitenciaria es un fenómeno complejo y se utiliza, con ciertas restricciones, como un indicador de la eficacia del sistema penitenciario en su conjunto (Yukhnenko, D., et al; 2023). Junto a la reincidencia penitenciaria —que, por su naturaleza, captura únicamente los casos más graves— los estudios más recientes incorporan la reincidencia judicial, definida como la apertura de una nueva causa penal por hechos posteriores a la excarcelación, con independencia de su desenlace. Esta ampliación metodológica revela un "lado oculto" del problema: para la cohorte de 2015, la tasa de reincidencia judicial se sitúa en el 53,1 %, muy por encima del 21,1 % penitenciario (Capdevila, M. et al.; 2023). La brecha de 32 puntos porcentuales indica que muchos delitos posteriores a la salida son gestionados mediante medidas alternativas o no culminan en una nueva pena de prisión. Desde un punto de vista evaluativo, la reducción de la reincidencia penitenciaria es una buena noticia; sin embargo, la magnitud de la reincidencia judicial recuerda que el objetivo último de prevención del delito está lejos de haberse alcanzado plenamente.

En este escenario, el modelo dominante que orienta las políticas de tratamiento es el conocido como modelo "Riesgo-Necesidad-Responsividad (RNR)", desarrollado por Andrews y Bonta (1990, 2010; Bonta y Andrews, 2017). El RNR surge como respuesta al célebre diagnóstico escéptico de Martinson (1974), quien concluyó que "nada funciona" en rehabilitación penitenciaria. Frente a este cierto escepticismo simplón, Andrews y Bonta mostraron, a partir de múltiples metaanálisis, que el tratamiento sí funciona, pero solo si se ajusta a ciertos principios (Andrews y Bonta, 1990, 2010). El principio de riesgo establece que la intensidad de la intervención debe ser proporcional al nivel de riesgo de reincidencia, así los programas intensivos aplicados sobre individuos de bajo riesgo son ineficaces e incluso pueden resultar contraproducentes por efectos de contagio criminal, mientras que dirigidos a personas de alto riesgo logran reducciones significativas de la reincidencia (Andrews y Dowden, 2006). El principio de necesidad distingue entre necesidades criminógenas —factores dinámicos que guardan una relación causal con la conducta delictiva, como el consumo de sustancias, las actitudes antisociales o los déficits laborales— y necesidades no criminógenas, cuya intervención puede ser valiosa en otros aspectos sociales o individuales, pero no incide directamente en el riesgo delictivo (Dowden y Andrews, 1999).

Por último, el principio de responsividad subraya que no basta con intervenir sobre los factores adecuados sino que también importa el modo de intervención. La responsividad general apunta a la superioridad de los programas cognitivo-conductuales estructurados frente a otros enfoques

en población delincuente, mientras que la responsividad específica reclama adaptar los métodos a las características individuales de cada interno —nivel cognitivo, estilo de aprendizaje, motivación, género, cultura— para maximizar el impacto (Landenberger y Lipsey, 2005; Bonta y Andrews, 2017). En conjunto, estos principios han transformado la práctica rehabilitadora en aquellas jurisdicciones que los han asumido de forma coherente.

El sistema penitenciario catalán ha ido transitando desde un modelo tradicional centrado en la contención hacia un modelo rehabilitador que se esfuerza por integrar el RNR en sus decisiones. La evidencia comparada resume con claridad esta evolución, el modelo puramente punitivo no reduce la reincidencia y puede resultar iatrogénico; por el contrario los sistemas que incorporan programas estructurados, atención sanitaria, educativa y laboral, evaluación sistemática del riesgo y apoyo postpenitenciario pueden alcanzar reducciones en torno o superiores al 10 % de las tasas de reincidencia delictiva. Aunque esta mejora pueda parecer modesta en términos absolutos, cada punto porcentual representa víctimas evitadas y trayectorias vitales reorientadas (Lipsey, M. y Cullen, F.T.; 2007).

Es precisamente en este marco donde se incorpora, en 2010, el RisCanvi a la gestión penitenciaria en los centros de internamiento penal de Cataluña. Conceptualmente, se trata de un sistema computerizado y basado en la técnica del Juicio Profesional Estructurado (Structured Professional Judgment), no de un algoritmo de inteligencia artificial que opera de manera autónoma (Hart, Douglas y Guy, 2016). Su diseño combina una base empírica sólida —factores de riesgo validados estadísticamente— con la flexibilidad del juicio profesional contextualizado final. El protocolo permite, mediante una misma aplicación evaluar el riesgo de reiteración de cinco resultados distintos: violencia auto-dirigida, violencia intrainstitucional, reincidencia general y violenta post-penitenciaria y quebrantamiento de permisos se presenta en dos versiones muy interrelacionadas en su aplicación cotidiana. Se trata del llamado RisCanvi Screening, una versión abreviada para cribado y compuesta por diez factores seleccionados por su capacidad discriminativa y facilidad de codificación, que permite una primera clasificación rápida. Cuando el resultado indica un riesgo elevado, se aplica la versión completa, el RisCanvi Completo —que incluye 43 factores de riesgo y cuatro variables moduladoras— que constituye el núcleo principal del sistema de gestión del riesgo en las prisiones de Cataluña.

La versión completa incluye cuarenta y tres factores de riesgo organizados en cinco grandes bloques: factores criminales y penitenciarios, que reco-

gen la historia delictiva, la edad al primer delito, la diversidad delictiva, la existencia de condenas previas y el comportamiento penitenciario anterior; factores biográficos, que incluyen problemas escolares precoces, conducta antisocial juvenil y experiencias de victimización en la infancia; factores sociales y familiares, referidos a la calidad de las relaciones familiares, la existencia de redes prosociales o antisociales, la situación laboral y económica; factores clínicos, que recogen trastornos mentales, consumo de sustancias, impulsividad y control de impulsos, así como actitudes favorables a la violencia; y, finalmente, factores de personalidad, que abarcan rasgos relativamente estables como el narcisismo, la frialdad emocional, la manipulación o la ausencia de empatía, vinculados al constructo de psicopatía.

Junto a estos cuarenta y tres factores, el protocolo incorpora cuatro factores moduladores que tienen un papel particularmente relevante desde la perspectiva del cambio: la situación penitenciaria (preventivos/penados), la edad (jóvenes/seniors), el sexo/género (hombre/mujer) y la nacionalidad (español/extranjero). Estos elementos permiten modular el riesgo en el algoritmo de cálculo inicial de los niveles de riesgo de los cinco resultados propios del RisCanvi. Esta modulación ayuda al traslado ponderal de los riesgos propios de los grupos de internos a los individuos en particular.

Una característica distintiva del RisCanvi, ya lo hemos indicado anteriormente, es su capacidad para estimar no un único riesgo global, sino cinco categorías diferentes, en el mismo proceso evaluador (gracias al programa computerizado de uso del RisCanvi) cada una con implicaciones prácticas específicas: el riesgo de violencia autodirigida (autolesiones y conducta suicida), crucial para la prevención del suicidio y la planificación de ubicaciones; el riesgo de violencia intrainstitucional, esencial para organizar la convivencia, asignar regímenes de vida y decidir sobre programas de control de la agresión; el riesgo de quebrantamiento de condena, especialmente relevante para la concesión de permisos de salida y para la gestión de la libertad condicional; el riesgo de reincidencia violenta tras la excarcelación, que es el que suscita mayor preocupación desde la óptica de la seguridad pública; y el riesgo de reincidencia general, incorporado en las revisiones de 2017 para ofrecer una visión más amplia de la probabilidad de un nuevo delito, cualquiera que sea su naturaleza (Andrés-Pueyo et al., 2018; Capdevila et al., 2015).

La aplicación diaria del protocolo está integrada en el marco habitual y en el contexto procedimental del trabajo propio de los equipos multidisciplinares de rehabilitación —psicólogos, juristas, trabajadores sociales y

educadores— en los centros penitenciarios. En la primera fase de aplicación del RisCanvi se recopila información procedente de múltiples fuentes para cada interno y en relación con el momento temporal en que se realiza la evaluación: sentencias y documentos judiciales, informes policiales, historiales penitenciarios anteriores, entrevistas semiestructuradas con la persona interna, informes de otros profesionales y observación directa de la conducta institucional. En una segunda fase, el equipo codifica los factores de riesgo siguiendo criterios operativos detallados en el manual operativo del RisCanvi (incorporado al programa informático), mayoritariamente en forma variable según sea en Screening (dicotómica presencia/ausencia) o Completa, con tres niveles de codificación: presencia/presencia probable o ausencia del factor de riesgo correspondiente. La codificación está concebida, en su diseño original, como un ejercicio colegiado que busca reducir la subjetividad individual mediante el contraste de puntos de vista para cada factor de riesgo (como hemos dicho 43 en el caso del RiscanVi Completo).

Una vez valorados cada factor de riesgo y validada la codificación (este es el input del RisCanvi a nivel computacional), el sistema informático penitenciario (el denominado e-RisCanvi que está integrado en el SIPC) calcula automáticamente la estimación del riesgo en tres niveles: bajo, medio o alto, para cada uno de los cinco resultados antes mencionados. El algoritmo se construyó en base a modelos de regresión logística, derivados de análisis realizados sobre la base de datos histórica de validación del propio sistema penitenciario de Cataluña. El resultado, expresado en tres categorías de riesgo —bajo, medio o alto— no es jurídicamente vinculante y es susceptible de modificaciones profesionalmente justificadas cuando sea oportuno. La técnica del Juicio Profesional Estructurado reserva a los profesionales la facultad de ratificar o modificar la clasificación, al alza o a la baja, cuando disponen de información específica e idiográfica relevante que no ha quedado identificada y valorada por los factores de riesgo estandarizados del modelo completo del RisCanvi. Incluso se pueden incluir —fuera de la fórmula algorítmica— informaciones codificadas de factores de riesgo relevantes del caso y el momento de la evaluación del RisCanvi completo. Cada modificación debe ser justificada por escrito, lo que cumple una doble función: preserva la flexibilidad y documenta la lógica de la decisión para facilitar su revisión, en caso de ser necesario, *a posteriori*.

Lejos de ser un acto puntual y aislado, la valoración del riesgo con el RisCanvi está diseñado para ser un proceso reiterativo y continuado. El protocolo establece una periodicidad recomendada de seis meses, con posibilidad de adelantar la revisión si se producen cambios significativos en la

situación del interno, a consideración de los profesionales del equipo de rehabilitación. Esta dimensión longitudinal permite monitorizar la evolución de los factores dinámicos, detectar incrementos o disminuciones de riesgo que requieran intervención inmediata, documentar reducciones que justifiquen progresiones de grado y, al mismo tiempo, generar trayectorias individuales que son de enorme valor para la toma de decisiones pronósticas. Desde su implantación, en 2010, se han acumulado en Cataluña en torno a 55 000 valoraciones completas de RisCanvi que están archivadas adecuadamente para su uso en tareas diversas, entre otras la de control de calidad o de recalibración del propio RisCanvi. Esto convierte a la base de datos asociada al RisCanvi, en un recurso excepcional para analizar trayectorias de riesgo, evaluar programas de tratamiento, identificar factores protectores y auditar las propias políticas penitenciarias del que nunca anteriormente se había podido disponer por razones diversas (archivos en papel, informes muy distintos según el profesional y el momento histórico, etc…).

3. APLICACIONES DEL RISCANVI EN LA GESTIÓN PENITENCIARIA

La extensión y facilidad con la que el RisCanvi se ha integrado en la práctica diaria de los equipos de rehabilitación en las prisiones catalanas se refleja en su papel en las decisiones penitenciarias sustanciales. Las clasificaciones iniciales de grado, las revisiones periódicas y las propuestas de progresión o regresión se apoyan en las valoraciones (y sus detalles analíticos) más recientes. El Programa Individual de Tratamiento (PIT) toma como punto de partida los factores de riesgo dinámicos identificados; estos se traducen en objetivos de intervención prioritaria que orientan la participación en programas específicos —tratamiento de adicciones, control de la violencia, habilidades prosociales, programas para violencia de género o delitos sexuales— de acuerdo con los principios del modelo RNR. La concesión de permisos ordinarios exige una valoración favorable del riesgo de quebrantamiento, y si este es elevado se diseñan estrategias de gestión adicionales —por ejemplo, acompañamiento, destinos vigilados, condiciones específicas—. La libertad condicional, que en el ordenamiento español exige un pronóstico individualizado de reinserción social, se fundamenta en gran medida en estos informes, sin perjuicio de que la decisión final corresponde al Juzgado de Vigilancia Penitenciaria. Incluso en situaciones especialmente delicadas, como la excarcelación definitiva de personas clasificadas de alto riesgo que agotan íntegramente su condena, la información acumulada en el protocolo sirve de base para articular dispositivos de coordinación con Fiscalía, Mossos d'Esquadra y servicios comunitarios.

La validez del sistema ha sido evaluada tanto desde la administración como desde la academia. En términos de capacidad predictiva, las distintas escalas alcanzan valores del área bajo la curva ROC (AUC)[11] entre aproximadamente 0,68 y 0,85 según el tipo de riesgo considerado, cifras comparables a las de instrumentos de referencia internacional como COMPAS u OASys (Andrés-Pueyo et al., 2018; Capdevila et al., 2015). En reincidencia penitenciaria, el *Odds Ratio* de 4,47 indica que la probabilidad de reingreso es casi cuatro veces y media mayor entre quienes son clasificados como de alto riesgo en comparación con quienes se encuentran en niveles bajos o medios. En el ámbito de la violencia intrainstitucional, el análisis específico de Férez-Mangas y Andrés-Pueyo (2024) reportó un AUC de 0,674 para la escala correspondiente, y muestra que en los incidentes de agresión grave a funcionarios, más de un 60 % de los internos autores de esas agresiones habían sido previamente clasificados como de riesgo alto; en incendios o agresiones graves entre internos, los porcentajes de alto riesgo también se incrementan de forma notable (Férez-Mangas y Andrés-Pueyo, 2024).

No obstante, la adecuación de los parámetros de calidad del RisCanvi, la validación empírica también descubre las tensiones inherentes a cualquier instrumento de este tipo. Los estudios sobre reincidencia de 2020 ponen de manifiesto, junto a una exactitud global elevada, una combinación peculiar de sensibilidad y especificidad. El modelo catalán se caracteriza por una especificidad muy alta —en torno al 94 %—, lo que significa que identifica correctamente a la mayoría de los internos que no reincidirán, pero presenta una sensibilidad más baja, lo que implica que deja de señalar como de alto riesgo a una proporción no despreciable de quienes finalmente reinciden. La calibración actual prioriza, por tanto, minimizar los falsos positivos, con el coste asociado de generar más falsos negativos. Se trata de una elección de política explícita, que protege a muchas personas de la estigmatización asociada a una etiqueta de "alto riesgo", pero a la vez limita la capacidad del sistema para detectar de forma anticipada todos los casos problemáticos. El denominado "Informe Tiresias", una auditoría especializada de programas y algoritmos computerizados a nivel empresarial y de organizaciones realizada el año 2023 sobre el RisCanvi (Dibria,2023), ha insistido precisamente en esta tensión y ha señalado la necesidad de un debate explícito sobre cuál es el equilibrio adecuado entre seguridad y no-discriminación.

11. Los valores del AUC (*Area Under Curve*) oscilan entre un 0,50 y un 1,00 siendo el primer valor el índice de ausencia de capacidad predictiva y el segundo la perfección predictiva.

El Informe Tiresias también ha puesto de relieve otras limitaciones técnicas. El algoritmo original del RisCanvi, desarrollado en 2009 e implementado directamente en la arquitectura de la base de datos, se basa esencialmente en la aplicación de la regresión logística clásica. Aunque estadísticamente es correcto, este planteamiento dificulta su actualización periódica y resulta menos flexible que las técnicas más actuales, como las de "Machine Learning" y otras, hoy disponibles. Las simulaciones realizadas en esa auditoría y *a posteriori* muestran que modelos como CatBoost o XGBoost podrían mejorar la capacidad discriminativa del RisCanvi en los pronósticos entre un 3 % y un 30 %, según el resultado considerado. Estas mejoras deben ponderarse frente a las exigencias de transparencia y explicabilidad de las predicciones.

Más allá del algoritmo, la implementación cotidiana del RisCanvi plantea también problemas propios. Es importante considerar estos aspectos porque la calidad del resultado depende, en último término, de la calidad de la información y de la forma en que se codifican los factores de riesgo. Entrevistas con profesionales usuarios habituales del RisCanvi han documentado ciertos problemas de ambigüedad en algunos factores —por ejemplo, el "rol delictivo destacado" o las "actitudes procriminales"— que requieren inferencias valorativas complejas y pueden generar diferencias significativas entre evaluadores. La alta carga de trabajo cotidiano dificulta, en ocasiones, la realización de valoraciones verdaderamente colegiadas, y en la práctica determinadas codificaciones se completan por una sola persona. La existencia de valoraciones previas tiende a anclar la percepción del profesional, dificultando el reconocimiento de cambios reales en factores dinámicos. A ello se suman sesgos potenciales ligados a variables como el origen nacional, la etnia o el nivel socioeconómico, que el diseño de la herramienta no puede eliminar por sí solo (Alemán, 2023).

También se ha descrito una percepción de efecto "caja negra" respecto al funcionamiento interno del algoritmo. Muchos usuarios conocen los factores de riesgo y manejan las categorías de riesgo resultantes, pero no pueden explicar con claridad cómo se combinan los factores para generar el resultado global, lo que limita su capacidad crítica a la hora de aceptar o cuestionar la salida del sistema. Esta situación contribuye a una paradoja ya que si bien formalmente el modelo mantiene al profesional en el centro de la decisión, el peso simbólico de la etiqueta de riesgo —sobre todo cuando llega a organismos externos, como juzgados o fiscalías— puede otorgar al resultado una influencia casi determinante. La literatura ha descrito este fenómeno como "automatización encubierta" de la decisión, el algoritmo se presenta como auxiliar, pero en la práctica condiciona de forma decisiva resoluciones que

afectan a derechos fundamentales. La literatura ha descrito este fenómeno mediante el concepto de "automation bias" (Busuioc, 2023), señalando cómo los sistemas algorítmicos que formalmente funcionan como apoyo a la decisión pueden convertirse en la práctica en "decisores primarios" (Završnik, 2021): el algoritmo se presenta como auxiliar, pero condiciona de forma decisiva resoluciones que afectan a derechos fundamentales.

Los retos técnicos y operativos del RisCanvi se proyectan directamente sobre la eficacia del tratamiento y, en particular, sobre el riesgo de efectos iatrogénicos. La evidencia catalana sobre programas de violencia de género ilustra de forma contundente este problema. Los datos del estudio de reincidencia de 2020 ya mencionado muestran que, entre los internos clasificados de bajo riesgo de reincidencia violenta, aquellos que participaron en el programa especializado para agresores de pareja presentaron una tasa de reincidencia penitenciaria del 26,2 %, mientras que los de perfil similar que no lo realizaron reincidieron en un 15,4 % (Capdevila,M. et al.; 2023). Someter a personas de bajo riesgo a intervenciones intensivas no solo no mejora sus resultados, sino que puede empeorarlos, probablemente por mecanismos de etiquetado, exposición a pares de mayor riesgo y consolidación de una identidad delictiva que no estaba firmemente establecida. Este tipo de hallazgos no son excepcionales y la literatura internacional sobre RNR lleva décadas advirtiendo contra la sobreintervención en perfiles de bajo riesgo (Andrews y Dowden, 2006; Bonta y Andrews, 2017).

Desde esta perspectiva, los errores de clasificación del riesgo —derivados tanto de las limitaciones algorítmicas como de la subjetividad en la codificación— tienen un efecto multiplicador en la asignación inadecuada de tratamientos y en la aparición de efectos iatrogénicos, con lo que la prescripción de programas de rehabilitación a los internos condenados queda supeditada a los niveles de riesgo de reincidencia. A su vez, variables ajenas al riesgo, como la disciplina regimental (por ejemplo, la presencia de expedientes disciplinarios recientes), adquieren un peso excesivo a la hora de priorizar el acceso a programas o a permisos, en detrimento de una aplicación rigurosa de los principios del modelo RNR.

Al mismo tiempo, la experiencia acumulada con el RisCanvi ha permitido refinar la comprensión de los perfiles de riesgo. Un estudio reciente de Sánchez de Ribera et al. (2025), basado en análisis de clases latentes con datos del protocolo, ha identificado cuatro grandes clústeres de internos varones con configuraciones de factores diferenciadas. Un primer grupo, de bajo riesgo, se caracteriza por puntuaciones reducidas en la mayoría

de los factores; un segundo grupo, de alto riesgo, concentra puntuaciones elevadas en casi todas las dimensiones, con especial intensidad en los factores clínicos y de personalidad; un tercero muestra riesgo medio o alto en variables clínicas y de personalidad, pero menor carga en los dominios criminal y sociofamiliar; y un cuarto clúster, también de alto riesgo, presenta el patrón inverso, con problemas marcados en lo criminal y sociofamiliar pero una psicopatología más moderada. Esta tipología subraya que la categoría genérica de "alto riesgo" encierra realidades muy distintas: mientras un interno del segundo clúster requerirá previsiblemente un abordaje clínico intensivo, otro del cuarto grupo se beneficiará sobre todo de intervenciones centradas en el empleo, la vivienda y la red social. El valor añadido del RisCanvi reside, precisamente, en facilitar esta individualización, más allá de la simple etiqueta categorial del riesgo.

La gestión del riesgo no se agota en el interior de la prisión. La transición hacia la comunidad —el llamado "reentry"— constituye un momento de vulnerabilidad especialmente elevada, pero también una ventana de oportunidad para la influencia de los programas de intervención postpenitenciario. En los últimos años, Cataluña ha desarrollado dispositivos de apoyo postpenitenciario que se sitúan en la vanguardia europea. Entre ellos destaca CerclesCat, adaptación del modelo internacional Circles of Support and Accountability (CoSA) (Richards, K. 2022), dirigido principalmente a personas condenadas por delitos sexuales de alto riesgo que regresan a la comunidad. El programa organiza, por medio de una combinación de voluntarios y de expertos en tratamiento profesional de internos penitenciarios, un "círculo interno" de entre cuatro y seis voluntarios que, junto con la persona excarcelada, conforman un espacio de apoyo práctico y emocional, y un "círculo externo de profesionales" —delegados de ejecución, psicólogos, fuerzas de seguridad— que supervisan el caso y garantizan la coherencia de la intervención.

Las evaluaciones de CerclesCat muestran resultados muy positivos: las tasas de reincidencia de los participantes se sitúan en torno al 11,8 %, frente a valores de 17,6 % y 25,7 % en grupos de control comparables (Cercles, 2016, 2020). Desde el punto de vista económico y social, se estima que el programa multiplica por once el beneficio de la inversión inicial y evita entre veintiuna y treinta nuevas víctimas. Junto a CerclesCat, otros dispositivos como programas de mentoría, arreglos de gestión interagencial del tipo Multi Agency Public Protection Arrangements (MAPPA) o planes específicos de "retorno seguro" conforman un ecosistema de apoyo que prolonga, en la comunidad, la lógica de la gestión estructurada del riesgo iniciada en prisión.

4. CONCLUSIONES

A la luz de toda esta experiencia, pueden articularse algunas conclusiones de alcance general. En primer lugar, la implantación del RisCanvi ha contribuido de forma decisiva a la profesionalización de la gestión del riesgo en el sistema penitenciario catalán, alineando la práctica diaria con los estándares de la criminología basada en la evidencia. La combinación de un instrumento actuarial validado, una aplicación reiterada y una integración profunda en las decisiones de clasificación y tratamiento ha permitido no solo mejorar la capacidad predictiva, sino también orientar la asignación de recursos hacia quienes más los necesitan.

En segundo lugar, la evidencia disponible sugiere que el protocolo alcanza niveles de validez comparables a los mejores instrumentos internacionales. Sin embargo, su eficacia real depende menos de la potencia estadística del algoritmo que de la calidad de su implementación: la claridad de los factores de riesgo, la suficiencia de los tiempos para la recogida de información, la práctica real del trabajo en equipo, la formación continua y la existencia de mecanismos de supervisión y auditoría son determinantes. Las debilidades recogidas en el Informe Tiresias —sobre todo en materia de actualización técnica, transparencia del código y equilibrio entre falsos positivos y falsos negativos— muestran que la mejora continua no es un lujo, sino una condición de legitimidad.

En tercer lugar, la reducción de la reincidencia observada en Cataluña parece responder a un conjunto de factores convergentes: la expansión de programas de tratamiento coherentes con el modelo RNR, la generalización de las salidas progresivas y la consolidación de un entramado de apoyo comunitario. Los estudios del CEJFE han demostrado que la salida escalonada mediante tercer grado o libertad condicional se asocia a tasas de reincidencia dramáticamente inferiores —en torno al 5,4 % y al 9,4 %, respectivamente— en comparación con las excarcelaciones directas desde segundo o primer grado, incluso cuando se controlan los sesgos de selección mediante técnicas de "*propensity scoring*" (Capdevila et al., 2023). Del mismo modo, la participación en programas especializados de violencia se asocia a reducciones de más de diez puntos porcentuales en las tasas de reingreso (Capdevila et al., 2023).

En cuarto lugar, el análisis crítico de la práctica muestra que las herramientas de riesgo, si se utilizan de forma imperfecta, pueden contribuir a un tratamiento ineficaz e incluso dañino. Los errores de calibración, la subjetividad en la codificación de ítems ambiguos, la ausencia de trabajo en equipo, la

presión de tiempo y la tendencia a sobrerreaccionar defensivamente ante el riesgo mediático se combinan para desviarse de los principios del modelo RNR. El resultado puede ser una sobreintervención sobre personas de bajo riesgo y un uso insuficiente de recursos con perfiles que tendrían más que ganar, generando efectos iatrogénicos y desperdiciando oportunidades de rehabilitación.

Por último, el uso de algoritmos predictivos en decisiones que afectan a la libertad y al estatus jurídico de las personas suscita cuestiones éticas de primer orden. La predicción actuarial se refiere siempre a grupos, no a individuos concretos, y su "verdad" es probabilística, no determinista (Hart, Michie y Cooke, 2007). Basar la denegación de un permiso o una libertad condicional en una estimación que, inevitablemente, cometerá errores, obliga a decidir, como sociedad, qué nivel de riesgo estamos dispuestos a aceptar, qué peso otorgamos a la prevención frente a la reinserción y qué garantías de transparencia y de derecho a la explicación ofrecemos a quienes son evaluados (Angwin et al., 2016; O'Neil, 2016; Chugh, 2021). La potencial reproducción de desigualdades estructurales a través de factores aparentemente neutros —nivel educativo, empleo, barrio de residencia— añade una capa más de complejidad al debate sobre "justicia algorítmica".

El diseño del RisCanvi como sistema de Juicio Profesional Estructurado para valorar el riesgo de los internos penitenciarios a reincidir y comportarse violentamente, que mantiene formalmente al profesional como último responsable de la decisión, ofrece un marco prudente para navegar entre estas tensiones. La investigación sobre sistemas híbridos humano-algorítmicos muestra que la combinación de un "buen algoritmo y un buen profesional" produce decisiones más precisas y consistentes que cualquiera de los dos por separado (Garrett, Crozier y Goodsey, 2020; Portela et al., 2024). El desafío, en adelante, será consolidar esta complementariedad sin caer ni en un "tecnoptimismo" ingenuo ni en una "tecnofobia" paralizante.

En definitiva, el protocolo RisCanvi se ha convertido en una pieza central del modelo penitenciario catalán, articulando la gestión del riesgo y el diseño de intervenciones en todas las fases de la ejecución penal. Su aportación a la reducción de la reincidencia y a la mejora de la seguridad interna en las prisiones es significativa, pero no automática. Depende de la inversión sostenida en modernización técnica, en formación de profesionales, en mecanismos de control de calidad y en programas de "reentry" postpenitenciario. Sobre todo, depende de que el sistema mantenga una convicción de fondo ya que el riesgo es una estimación sobre la probabilidad de determinados comportamientos y no una sentencia sobre la esencia de las personas. La

legitimidad última de un modelo de gestión del riesgo no reside solo en su precisión estadística, sino en su capacidad para conjugar dignidad, proporcionalidad y esperanza razonable en la posibilidad de cambio. En la medida en que el RisCanvi siga subordinado a estos valores y se integre en un ecosistema institucional orientado a la rehabilitación y a la responsabilidad compartida, podrá seguir siendo un referente de criminología aplicada al servicio de un sistema penal más eficaz y humano.

5. REFERENCIAS

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S.,... & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*(3), 341-382.

Alemán Aróstegui, L. (2023). El uso del RisCanvi en la toma de decisiones penitenciarias. *Estudios Penales y Criminológicos*, 44(ext), 1-54.

Andrés-Pueyo, A., & Redondo, S. (2007). Predicción de la violencia: Entre la peligrosidad y la valoración del riesgo de violencia. Papeles del Psicólogo, 28(3), 157-173.

Andrés-Pueyo, A., Arbach-Lucioni, K., & Redondo, S. (2018). *The RisCanvi: A new tool for assessing risk for violence in prison and recidivism*. En J. P. Singh, D. G. Kroner, J. S. Wormith, S. L. Desmarais & Z. Hamilton (Eds.), *Handbook of recidivism risk/needs assessment tools* (Cap. 13, pp. 255-268). Wiley-Blackwell.

Andrews, D. A., & Bonta, J. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior*, 17(1), 19-52.

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct* (5.ª ed.). Routledge.

Andrews, D. A., & Dowden, C. (2006). Risk principle of case classification in correctional treatment: A meta-analytic investigation. *International Journal of Offender Therapy and Comparative Criminology, 50*(1), 88-100.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's *software* used across the country to predict future criminals. And it's biased against black. *ProPublica*. https://www.propublica.org.

Blevins, K. R., Listwan, S. J., Cullen, F. T., & Jonson, C. L. (2010). A general strain theory of prison violence and misconduct: An integrated model of inmate behavior. *Journal of Contemporary Criminal Justice, 26*(2), 148-166.

Bonta, J., & Andrews, D. A. (2017). *The psychology of criminal conduct* (6.ª ed.). Routledge.

Busuioc, M. (2023). "Human-AI Interactions in Public Sector Decision Making: «Automation Bias" and «Selective Adherence" to Algorithmic Advice." *Journal of Public Administration Research and Theory*, 33(1), 153-169.

Capdevila Capdevila, M., Blanch Serentill, M., Ferrer Puig, M., Andrés-Pueyo, A., Framis Ferrer, B., Comas López, N., Garrigós Bou, A., Boldú Pedro, A., Batlle Manonelles, A., & Mora Encinas, J. (2015). *Tasa de reincidencia penitenciaria 2014*. Centre d'Estudis Jurídics i Formació Especialitzada (CEJFE), Departament de Justícia, Generalitat de Catalunya.

Capdevila Capdevila, M., Framis Ferrer, B., Soler Iglesias, C., Andrés-Pueyo, A., Ruiz Sarrión, L., Arrufat Pijuan, A., Díez Lerma, R., Ribas Plano, P., & Blanch Serentill, M. (2023). *Taxa de reincidència penitenciària 2020*. Centre d'Estudis Jurídics i Formació Especialitzada (CEJFE), Departament de Justícia, Generalitat de Catalunya.

Cercles. (2016). *Informe d'avaluació del programa CerclesCat 2013-2015*. Departament de Justícia.

Cercles. (2020). *Avaluació d'impacte de CerclesCat (2015-2019)*. Departament de Justícia.

Chugh, N. (2021). Risk assessment tools on trial: Lessons learned for "Ethical AI" in the criminal justice system. 2021 IEEE International Symposium on Technology and Society (ISTAS), 1-5.

Cooke, D. J. (1991). Violence in prisons: The influence of regime factors. *The Howard Journal of Criminal Justice*, 30(2), 95-109.

Dribia Data & Direcció General d'Afers Penitenciaris. (2024). *Informe Tiresias: Auditoria de l'algorisme RisCanvi* (Versión del 9 de enero de 2024). Departament de Justícia, Drets i Memòria, Generalitat de Catalunya.

DiIulio, J. J. (1987). *Governing prisons: A comparative study of correctional management*. The Free Press.

Dowden, C., & Andrews, D. A. (1999). What works for female offenders: A meta-analytic review. *Crime & Delinquency, 45*(4), 438-452.

Durose, M. R., Cooper, A. D., & Snyder, H. N. (2014). *Recidivism of prisoners released in 30 States in 2005: Patterns from 2005 to 2010* (NCJ 244205). Bureau of Justice Statistics.

Feeley, M., & Simon, J. (1992). The new penology: Notes on the emerging strategy of corrections and its implications. *Criminology, 30*(4), 449-474.

Férez-Mangas, D., & Andrés-Pueyo, A. (2024). *Violència intrainstitucional a les presons de Catalunya*. Centre d'Estudis Jurídics i Formació Especialitzada (CEJFE), Departament de Justícia, Generalitat de Catalunya.

Garland, D. (2001). *The culture of control: Crime and social order in contemporary society*. University of Chicago Press.

Hannah-Moffat, K. (2005). Criminogenic needs and the transformative risk subject: Hybridizations of risk/need in penality. *Punishment & Society, 7*(1), 29-51.

Gendreau, P., Goggin, C., & Law, M. (1997). Predicting prison misconducts. *Criminal Justice and Behavior, 24*(4), 414-431.

Hart, S. D., Douglas, K. S., & Guy, L. S. (2016). The structured professional judgment approach to violence risk assessment: Origins, nature, and advances. In D. P. Boer (Ed.), *The Wiley handbook on the theories, assessment and treatment of sexual offending* (pp. 643-666). Wiley-Blackwell.

Hart, S. D., Michie, C., & Cooke, D. J. (2007). Precision of actuarial risk assessment instruments: Evaluating the "margins of error" of group v. individual predictions of violence. *British Journal of Psychiatry, 190*(Suppl. 49), s60-s65.

Irwin, J., & Cressey, D. R. (1962). Thieves, convicts and the inmate culture. *Social Problems, 10*(2), 142-155.

Landenberger, N. A., & Lipsey, M. W. (2005). The positive effects of cognitive-behavioral programs for offenders: A meta-analysis of factors associated with effective treatment. *Journal of Experimental Criminology, 1*(4), 451-476.

Lipsey, M. W., & Cullen, F. T. (2007). The effectiveness of correctional rehabilitation: A review of systematic reviews. *Annual Review of Law and Social Science, 3*, 297-320.

Martinson, R. (1974). What works? Questions and answers about prison reform. *The Public Interest, 35*, 22-54.

Monahan, J., Metz, A., & Garrett, B. (2018). Judicial appraisals of risk assessment in sentencing. Behavioral Sciences & the Law, 36(5), 565-575.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Portela, M., Pereira, A., Silva, S., & Oliveira, J. (2024). Human-algorithm collaboration in criminal justice decision-making: A systematic review. *Journal of Criminal Justice, 89*, 102050. [Referencia modelo; adaptar a la publicación real si la tienes].

Richards, K. (2022). *Desistance from Sexual Offending: The Role of Circles of Support and Accountability*. Routledge.

Sánchez de Ribera, O., Trajtenberg, N., Férez-Mangas, D., & Andrés-Pueyo, A. (2025). Prison violence: a latent class analysis of adult male offenders using the RisCanvi. *Psychology, Crime & Law*. (Advance online publication.)

Sykes, G. M. (1958). *The society of captives: A study of a maximum security prison*. Princeton University Press.

Wolff, N., Shi, J., & Bachman, R. (2008). Measuring victimization inside prisons: Questioning the questions. *Journal of Interpersonal Violence*, 23(10), 1343-1362.

Yukhnenko, D., Cherukuri, A., Wolf, A., La Ragione, L. L., Jones, P. B., & Fazel, S. (2023). Criminal recidivism rates globally: A 6-year systematic review update. *Journal of Criminal Justice*, 88, 102115.

Završnik, A. (2021). "Algorithmic justice: Algorithms and big data in criminal justice settings." *European Journal of Criminology*, 18(5), 623-642.

Capítulo 6

Evaluación técnica y ética del sistema RisCanvi: potencialidades y riesgos en la predicción algorítmica penitenciaria

Iulen Ibáñez Baños

CEO de Datacy

SUMARIO: 1. INTRODUCCIÓN. 2. PRESUPUESTOS TÉCNICOS Y DE DATOS. 3. POTENCIALIDADES DE EVOLUCIÓN. 4. DISCUSIÓN. 5. MINIMIZACIÓN DE RIESGOS. 6. CONCLUSIONES. 7. REFERENCIAS.

1. INTRODUCCIÓN

Este trabajo examina el sistema RisCanvi desde una doble perspectiva técnica y ética, combinando la información procedente de la auditoría externa realizada por Dribia Data Research (2024) con una reflexión personal sobre los límites de la predicción algorítmica en contextos de decisión pública. RisCanvi, herramienta utilizada en el sistema penitenciario catalán para valorar el riesgo de reincidencia, constituye un caso ilustrativo de los desafíos de transparencia, calidad del dato y gobernanza algorítmica en entornos sensibles. El texto analiza su diseño técnico, los resultados obtenidos y las recomendaciones de mejora, situando el debate en la intersección entre la analítica avanzada y la ética aplicada a la administración de justicia.

En los últimos años, la expansión del uso de algoritmos en la administración pública ha reabierto debates clásicos sobre responsabilidad, objetividad y justicia. Los algoritmos, entendidos en sentido amplio, son secuencias de operaciones lógicas que transforman datos en decisiones o recomendaciones. Sin embargo, no todos los algoritmos son iguales: algunos se limitan a aplicar

fórmulas predefinidas por expertos humanos, mientras que otros aprenden patrones directamente de los datos, modificando sus parámetros mediante entrenamiento estadístico. Esta distinción entre "algoritmos-calculadora" y "algoritmos-inteligentes" resulta esencial para entender el caso RisCanvi.

RisCanvi fue concebido como una herramienta actuarial para estimar el riesgo de reincidencia de personas internas en el sistema penitenciario catalán. Su lógica es fundamentalmente determinística: combina un conjunto de factores de riesgo ponderados para producir una clasificación (riesgo bajo, medio o alto). En consecuencia, su capacidad predictiva depende más de la validez del diseño inicial que de la retroalimentación empírica de los datos. A pesar de sus limitaciones, RisCanvi representa un esfuerzo temprano y relevante por sistematizar la toma de decisiones en el ámbito penitenciario. Y, en este sentido, es una herramienta muy positiva.

2. PRESUPUESTOS TÉCNICOS Y DE DATOS

El informe de auditoría elaborado por Dribia Data Research (2024) ofrece un diagnóstico detallado del modelo y de los datos que lo sustentan. Los resultados muestran que el rendimiento del algoritmo, medido mediante la métrica AUC-ROC, oscila entre 0,68 y 0,75 dependiendo del tipo de reincidencia evaluada, valores comparables a los de sistemas similares como COMPAS o OASys. En términos de equidad, la auditoría concluye que no existen sesgos discriminatorios graves, si bien se observan diferencias leves en grupos protegidos (menores de 30 años o personas extranjeras) donde el modelo tiende a subestimar el riesgo.

El análisis también revela limitaciones estructurales: la plataforma técnica no se integra plenamente con el sistema SIPC (Sistema de Información Penitenciaria Catalán), lo que impide automatizar validaciones y dificulta la trazabilidad de los datos. Además, la ausencia de documentación viva y la dependencia de un único proveedor externo reducen la transparencia y la posibilidad de revisión independiente. Desde una óptica de ingeniería de datos, estas carencias no solo afectan a la robustez técnica del algoritmo, sino también a su legitimidad institucional.

Otro aspecto crítico es la calidad del dato. La auditoría subraya que, aunque el volumen de información disponible ha aumentado desde la creación del sistema, persisten inconsistencias, vacíos y heterogeneidades en las bases de datos de origen. Para un algoritmo que no aprende automáticamente, la fiabilidad de las conclusiones depende directamente de la consistencia de los registros administrativos. Por tanto, la mejora de RisCanvi no puede abordarse

únicamente desde la optimización algorítmica, sino también (y, quizás, sobre todo) desde una política de datos públicos más estructurada y verificable.

3. POTENCIALIDADES DE EVOLUCIÓN

Entre las recomendaciones del informe Tiresias se encuentra la posibilidad de implementar un modelo predictivo más moderno, basado en aprendizaje automático supervisado, concretamente mediante la técnica CatBoost. Este tipo de modelos permite reentrenar los parámetros con cada nueva remesa de datos, adaptándose mejor a los cambios en la población penitenciaria y a los patrones de reincidencia. Asimismo, ofrecen herramientas avanzadas de explicabilidad (como SHAP values) que facilitan entender el peso de cada variable en la predicción individual, mejorando la transparencia y la confianza institucional.

Sin embargo, el paso hacia un modelo de inteligencia artificial no debe abordarse como una mera actualización tecnológica. Implica redefinir los límites entre la decisión humana y la automatización, y exige crear un marco de gobernanza que garantice el control democrático del algoritmo. La potencial ganancia en precisión estadística no puede justificar una pérdida en explicabilidad o en rendición de cuentas. Desde mi punto de vista, la evolución de RisCanvi debería seguir una lógica incremental: reforzar primero la calidad y trazabilidad del dato, y solo después avanzar hacia modelos más complejos.

4. DISCUSIÓN

El caso RisCanvi es especialmente interesante porque refleja las tensiones propias de la era algorítmica en la gestión pública. Su diseño responde a una aspiración legítima: reducir la arbitrariedad en decisiones sensibles —como la concesión de permisos o la valoración del riesgo de reincidencia— mediante criterios objetivos y replicables. No obstante, esta misma aspiración choca con una realidad estructural: los datos administrativos son productos humanos, con sesgos, omisiones y contextos que ningún modelo puede eliminar completamente.

Como analista y docente en analítica avanzada de datos, me parece crucial subrayar que la neutralidad algorítmica es una ilusión. En un «algoritmo-calculadora', el modelo refleja las decisiones de diseño de quienes lo construyen: qué variables incluir, cómo ponderarlas, qué métricas optimizar y qué umbrales de riesgo aceptar. En el caso de RisCanvi, los puntos de corte

fueron definidos para minimizar falsos positivos —es decir, evitar clasificar erróneamente a una persona como de alto riesgo—, lo cual es éticamente defendible pero estadísticamente costoso: aumenta el número de falsos negativos, o sea, personas clasificadas como bajo riesgo que luego reinciden.

En un «algoritmo-inteligente', la no-neutralidad se introduce sesgando los datos con los que se entrena el modelo.

El dilema de fondo no es técnico sino político: ¿preferimos un sistema que «falle por exceso" o que «falle por defecto"? La respuesta dependerá del equilibrio entre derechos individuales y seguridad colectiva que la sociedad considere aceptable. Por eso, más allá de la métrica AUC-ROC o de la precisión global, lo que debería preocuparnos es la capacidad institucional de explicar por qué un algoritmo decide lo que decide. La explicabilidad no es un lujo técnico, sino una condición para la legitimidad democrática de la inteligencia artificial en la gestión pública.

5. MINIMIZACIÓN DE RIESGOS

A partir de la auditoría y del análisis conceptual, pueden formularse tres estrategias clave para minimizar los riesgos de un sistema predictivo de reincidencia. Primero, la creación de una gobernanza algorítmica formal, con un comité mixto que combine perfiles jurídicos, técnicos y criminológicos, responsable de validar versiones y revisar sesgos. Segundo, el desarrollo de una infraestructura de datos robusta y auditada, que documente el origen, las transformaciones y el uso de cada variable. Y tercero, la formación continua de los profesionales penitenciarios, para que comprendan las limitaciones del modelo y puedan interpretarlo con criterio.

Estas tres dimensiones —gobernanza, infraestructura y capacitación— constituyen los pilares de un enfoque responsable de la analítica penitenciaria. Sin ellas, cualquier intento de sofisticar el modelo mediante aprendizaje automático estaría condenado a reproducir los mismos problemas bajo una apariencia más moderna. En otras palabras, el futuro de RisCanvi no depende tanto de la inteligencia del algoritmo como de la inteligencia institucional que lo rodea.

6. CONCLUSIONES

RisCanvi es una herramienta valiosa, no por su sofisticación técnica, sino por lo que simboliza: el esfuerzo de una administración por basar sus deci-

siones en evidencia y no en intuiciones. Su principal mérito reside en haber abierto un debate necesario sobre la transparencia, la calidad del dato y la rendición de cuentas en el uso de algoritmos públicos. La auditoría de 2024 demuestra que, incluso en contextos complejos, es posible evaluar empíricamente los modelos y proponer mejoras desde un enfoque constructivo.

En mi opinión, el futuro del análisis de riesgos penitenciarios no pasa por algoritmos cada vez más complejos, sino por ecosistemas de datos más confiables, abiertos y auditables. La tecnología debe ser un instrumento de racionalización y justicia, no un sustituto del juicio humano. Si algo nos enseña RisCanvi, es que el verdadero desafío no está en predecir el futuro, sino en hacerlo de manera justa, transparente y responsable.

7. REFERENCIAS

American Psychological Association. (2020). *Publication Manual of the American Psychological Association* (7th ed.). https://apastyle.apa.org/.

Dribia Data Research. (2024). *Auditoría del algoritmo RisCanvi: Informe Tiresias*. Centre d'Estudis Jurídics i Formació Especialitzada (CEJFE).

Ibáñez, I. (2025). Retos criminológicos ante los instrumentos de valoración y gestión de riesgos en el ámbito penitenciario: una exploración de su adaptación al ámbito vasco. Datacy.

Capítulo 7

La arquitectura de la toma de decisiones en equipo en el ámbito penitenciario: Retos ético-jurídicos de las herramientas predictivas de violencia y de quebrantamiento de condena

Gema Varona Martínez[12]

Instituto Vasco de Criminología (IVAC-KREI)

Esto supone un cambio fundamental con respecto a las formas de gobernanza anteriores. El Estado industrial moderno buscaba controlar a su población mediante técnicas demográficas y estadísticas, de ahí su obsesión por la clasificación, la legibilidad y la cuantificación. El poder en la modernidad digital no necesita ver a los individuos de la misma manera; en su lugar, detecta patrones de comportamiento, grupos de interacciones y tendencias que pueden ser guiadas en lugar de ordenadas. Lo que pierde en claridad, lo gana en flexibilidad. En lugar de mapear la realidad en un modelo estadístico rígido, responde en tiempo real, ajustándose a los flujos de datos.

Si el fracaso del Estado industrial fue su intento de encajar a las personas y la naturaleza en cuadrículas rígidas, el fracaso del poder digital podría ser su intento de representar el mundo como una serie de patrones dinámicos. Las metáforas de las redes y los grupos son tan parciales e incompletas como la metáfora de la red homogénea que las precedió... El poder digital dejará su huella: remodelará nuestra forma de trabajar, de ser gobernados y de entendernos a nosotros mismos. Toda lógica de gobierno tiene sus puntos ciegos, y todo intento de imponer orden en el mundo crea nuevas formas de desorden... (Törnberg y Uitermark, 2025).

12. Directora del Instituto Vasco de Criminología/Kriminologiaren Euskal Institutua (EHU). Orcid: https://orcid.org/0000-0002-2794-2239

SUMARIO: 1. INTRODUCCIÓN: UN ACERCAMIENTO ÉTICO-JURÍDICO Y CRIMINOLÓGICO A LA ADAPTACIÓN DE LA HERRAMIENTA RISCANVI. 1.1. *Datos básicos sobre RisCanvi*. 1.2. *Algunas ventajas y desventajas de las herramientas de valoración y gestión del riesgo, en particular en el funcionamiento de RisCanvi*. 1.3. *Requisitos tecnológicos y jurídicos básicos para una posible adaptación de RisCanvi.* 1.4. *Objetivo y estructura del capítulo.* 2. MARCO NORMATIVO DEL CONSEJO DE EUROPA EN EL ÁMBITO PENAL SOBRE EL USO DE LA INTELIGENCIA ARTIFICIAL Y TECNOLOGÍAS SIMILARES. 3. MARCO JURÍDICO SOBRE LA INTELIGENCIA ARTIFICIAL EN LA UNIÓN EUROPEA. 4. MARCO JURÍDICO BÁSICO, COMUNITARIO Y ESPAÑOL, SOBRE PROTECCIÓN DE DATOS, EN PARTICULAR EN RELACIÓN CON EL USO DE PERFILACIONES Y PREDICCIONES EN EL ÁMBITO PENITENCIARIO: EL REGLAMENTO GENERAL DE PROTECCIÓN DE DATOS (RGPD) DE 2016 DE LA UE, LA LEY ORGÁNICA 3/2018 DE PROTECCIÓN DE DATOS PERSONALES Y GARANTÍA DE LOS DERECHOS DIGITALES, LA DIRECTIVA 2016/680 Y LA LEY ORGÁNICA 7/2021 SOBRE PROTECCIÓN DE DATOS PERSONALES EN EL ÁMBITO PENAL, Y SU CONEXIÓN CON LA LEY ORGÁNICA GENERAL PENITENCIARIA Y EL REGLAMENTO PENITENCIARIO. 4.1. *Normativa europea y española general sobre protección de datos.* 4.2. *Normativa europea y española de protección de datos en el ámbito penal-penitenciario: la Directiva 2016/680, la Ley Orgánica 7/2021 sobre protección de datos personales en el ámbito penal y el Reglamento penitenciario español.* 5. CONTRASTE DEL MARCO NORMATIVO CON UN ANÁLISIS DE RESOLUCIONES JUDICIALES QUE MENCIONAN LA TABLA DE VALORACIÓN DE RIESGO Y LA HERRAMIENTA RISCANVI. 6. CONTRASTE A TRAVÉS DE UN TRABAJO DE CAMPO REALIZADO EN EUSKADI: OBSERVACIÓN NO PARTICIPANTE DE LA TOMA DE DECISIONES TÉCNICAS EN LAS TRES PRISIONES VASCAS. 6.1. *Metodología del trabajo de campo de observación no participante*. 6.2. *Consideraciones éticas.* 6.3. *Limitaciones*. 6.4. *Resultados del análisis temático con palabras y extractos ilustrativos*. 6.4.1. Complejidad en el manejo y limitaciones de la información disponible. 6.4.2. Interdisciplinariedad en el trabajo en equipo. 6.4.3. Retos y complementariedad a la hora de llegar a consensos desde la diversidad de perspectivas técnicas y humanas, donde se reconocen posibles sesgos. 6.4.4. Manejo diverso de tablas en papel de factores de valoración como criterios de orientación y justificación. 6.4.5. Desconocimiento, suspicacia, interés y ofrecimiento de colaboración ante la experiencia del RisCanvi en Cataluña y su posible adaptación en el País Vasco. 6.4.6. Necesidad de más recursos humanos, materiales y tecnológicos. 6.5. *Consideración final*. 7. CONTRASTE A TRAVÉS DE UN CUESTIONARIO Y

UN GRUPO DE OPINIÓN DESARROLLADO CON ESTUDIANTES UNIVERSITARIOS VASCOS. 7.1. *Resultados: número de respuestas y perfiles sociodemográficos*. 7.2. *Mayor identificación con alguna de las personas que participan en el pódcast sobre RisCanvi (abogados/as, funcionarios/as, técnicos/as que la elaboraron, perfeccionan y/o auditan, condenados/as..) y mayor preocupación en el uso de RisCanvi*. 7.3. *Mayor beneficio que aportaría la utilización de este tipo de herramienta en las prisiones vascas*. 7.4. *Adecuación de la inversión de fondos públicos en este tipo de herramienta y, en su caso, de sistemas de control*. 7.5. *Otros comentarios adicionales*. 7.6. *Grupo de discusión*. 7.7. *Consideraciones generales.* 8. RECAPITULACIÓN: NAVEGANDO ENTRE UN JUICIO CLÍNICO, UN JUICIO PROBABILÍSTICO Y UN SISTEMA MIXTO. 9. REFERENCIAS.

1. INTRODUCCIÓN: UN ACERCAMIENTO ÉTICO-JURÍDICO Y CRIMINOLÓGICO A LA ADAPTACIÓN DE LA HERRAMIENTA RISCANVI

Siguiendo a Cañas (2025), miembro de la Comisión de Derecho Penitenciario del Colegio de la Abogacía de Córdoba*:*

> La inteligencia artificial (IA)[13] llega cada vez a más sectores de la actividad humana y su uso generalizado en el ámbito penitenciario no tardará demasiado en llegar a nuestro centro penitenciario de referencia. Llegados a este punto, ¿sabremos los abogados impugnar una decisión adoptada con base en una herramienta tecnológica de este nivel o acaso vencerán los automatismos a nuestro sistema de individualización científica?

13. Según se recoge en la Estrategia Española de Inteligencia Artificial (2020-2025), del Ministerio para la Transformación Digital y de la Función Pública, en línea con la Estrategia Europea, la Comisión de la UE define la IA como "sistemas de software (y posiblemente también de hardware) diseñados por humanos que, ante un objetivo complejo, actúan en la dimensión física o digital: percibiendo su entorno, a través de la adquisición e interpretación de datos estructurados o no estructurados, razonando sobre el conocimiento, procesando la información derivada de estos datos y decidiendo las mejores acciones para lograr el objetivo dado. Los sistemas de IA pueden usar reglas simbólicas o aprender un modelo numérico, y también pueden adaptar su comportamiento al analizar cómo el medio ambiente se ve afectado por sus acciones previas". Suele diferenciarse entre la inteligencia artificial estrecha (IAE) y la inteligencia artificial general (IAG), que sería capaz de realizar una serie de tareas cognitivas y prácticas, más semejantes a las de un ser humano. Más allá de eso se encuentra la perspectiva teórica de la superinteligencia artificial (ASI) (Yampolskiy, 2025). La técnica de IA más extendida se conoce como aprendizaje automático. Permite identificar patrones en los datos y aplicar este conocimiento a nuevos datos, de modo que la IA puede aprender por sí misma a partir de los datos.

> Lo cierto es que ya en España, en 2023, se intentó utilizar la IA en un proyecto piloto en el Centro Penitenciario Mas d´Enric de Tarragona. La prueba consistía en analizar a través de las imágenes registradas por las cámaras de vigilancia interna las expresiones faciales y el lenguaje corporal de los reclusos para prevenir riesgos y clasificar a los sujetos en función de su susceptibilidad a causar incidentes de seguridad. Tal clasificación podría dar lugar a decisiones que afectasen al tratamiento penitenciario, a la posible obtención de permisos y otros beneficios. Y todo ello sin que el sujeto en cuestión hubiese tenido que incurrir en ninguna conducta reprobable. El proyecto se detuvo tras la adjudicación del contrato, sin llegar siquiera a implantarse. El argumento para la paralización del experimento fue la entonces inminente entrada en vigor del Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (en vigor desde el 1 de agosto de 2024), lo que aconsejaba una revisión antes de poner en marcha el proyecto. A día de hoy su puesta en marcha no se ha producido (pfos. 1º y 2º).

Cañas también señala que, no utilizando en estos momentos la inteligencia artificial[14], RisCanvi (en castellano: riesgo y cambio), la herramienta penitenciaria objeto de este capítulo, sí utiliza algoritmos predictivos, si bien, en principio, de forma razonablemente adecuada:

> Lo cierto es que RisCanvi funciona, o al menos sus errores no han generado más debate que aquel que puedan generar los errores humanos. Según el único estudio público que ha evaluado el algoritmo de RisCanvi, en 2014, se calculó que el 77% de los internos a quienes se identificó como de riesgo medio o alto reincidieron, mientras que el 57% de los que no volvieron a delinquir habían sido etiquetados como de bajo riesgo (pfo. 3º).

1.1. DATOS BÁSICOS SOBRE RISCANVI

La administración penitenciaria catalana afirma encontrarse "en primera línea de los países de Europa", desde la puesta en marcha, en 2009, de la herramienta RisCanvi[15], la cual solo se aplica en el ámbito catalán, con competencia plena en la gestión de las prisiones, como es el caso actual del País

14. Véase, en esta monografía, el capítulo de Iulen Ibáñez.

15. Dentro del Departamento de Justicia y Calidad Democrática de la Generalitat, véase la página web oficial sobre RisCanvi (de donde procede la cita textual y la información general aludida en esta página), en https://justicia.gencat.cat/ca/ambibosco/reinsercio_i_serveis_penitenciaris/serveis_penitenciaris/RisCanvi/. Dicha web recoge el Resumen del Protocolo de evaluación del riesgo (RisCanvi) (2017); el *Manual de aplicación del protocolo de valoración del riesgo* (2019); la *Escala de evaluación RisCanvi screening* (2019); la *Escala de evaluación Riesgo completo* (2009); la *Auditoría del algoritmo. Informe final* (2024) y la *Auditoría del algoritmo. Resumen ejecutivo* (2024).

Vasco. Desde que se concibió, considerando lo que ya se hacía en algunos países[16], se trató de responder a dos necesidades distintas, si bien relacionadas:

a) la preocupación social creciente en relación con la reincidencia delictiva violenta, a partir de casos trágicos y muy mediáticos de violencia sexual[17], y

b) la necesidad de mejorar la tecnificación y cualificación de los profesionales responsables de la ejecución penal.

RisCanvi se define por la administración penitenciaria catalana como una herramienta pionera de evaluación y gestión del riesgo del conjunto de la población penitenciaria, en concreto aquellas personas que se encuentran en centros penitenciarios de régimen ordinario y abierto y también personas en libertad condicional. En la actualidad, a través del RisCanvi se realizan dos evaluaciones, a modo de cálculos probabilísticos: un *screening* y una evaluación completa que considera 43 factores que aluden a variables estáticas y dinámicas tradicionalmente relacionadas con la delincuencia. La información del sistema procede fundamentalmente de los expedientes y de las aportaciones de juristas, psicólogos, educadores y trabajadores sociales y, allí donde se valore su trabajo, criminólogos. Implica también entrevistas a los presos cada seis meses. Se trata de asignar valores a cada variable analizada, algo que se realiza siguiendo la opinión técnica de los profesionales, sin perjuicio de la información que pueda servir de guía. En principio, otros operadores jurídicos, como jueces, fiscales o letrados, y la propia persona interna, solo reciben la puntuación final, sin que puedan tener información sobre los criterios aplicados y su interpretación.

1.2. ALGUNAS VENTAJAS Y DESVENTAJAS DE LAS HERRAMIENTAS DE VALORACIÓN Y GESTIÓN DEL RIESGO, EN PARTICULAR EN EL FUNCIONAMIENTO DE RISCANVI

En realidad, las tablas de valoraciones de riesgo (TVR), para clasificaciones, concesión de permisos[18], etcétera, llevan mucho tiempo utilizándose en papel u

16. Véase el capítulo de Aitor Martínez.
17. En concreto, la salida de prisión de los violadores del Vall d'Hebron y Del Eixample.
18. Según García (2025):
Cuando, en 1993, el Ministerio de Justicia encargó al psicólogo Miguel Clemente que desarrollara una herramienta predictiva sobre presos, era poco habitual escuchar la palabra "algoritmo". A principios de ese año, dos apicultores encontraron en una fosa los cuerpos de Toñi, Miriam y Desiré, las víctimas de los crímenes de Alcàsser,

hojas de cálculo (véanse las observaciones sobre las mismas, en relación con el epígrafe posterior sobre el trabajo de campo realizado para este capítulo). Estas tablas han sido criticadas por su antigüedad y su falta de soporte científico.[19]

Frente a las críticas de falta de apoyo científico y desfase, así como de efectos discriminatorios respecto a mujeres, personas extranjeras o en exlcusión, se ha subrayado que, para la administración penitenciaria, esta tabla constituye una herramienta más y que, en todo caso, existe la posibilidad de recurso. Se subraya que lo decisivo es el estudio global que realiza el equipo técnico del centro penitenciario mediante un estudio complejo y multifactorial. Ahora bien, en línea con la observación de campo realizada para la elaboración de este capítulo, una educadora social en el centro penitenciario de Valdemoro, miembro del grupo de trabajo en salud mental en prisión de la Asociación Española de Neuropsiquiatría, afirma que la herramienta se legitima también para otorgar un barniz un poco científico a una decisión que está tomada de antemano (García, 2025). En todo caso, instrucciones de

---

un suceso que conmocionó al país y en el que uno de sus autores había escapado de prisión el año anterior usando un permiso penitenciario.
El crimen de Alcàsser dio lugar a la creación de la Tabla de Variables de Riesgo (TVR), una fórmula de diez factores que, 32 años después, aún sirve como herramienta para que los profesionales de prisiones decidan si otorgar o no permisos de salida de prisión. Solo en 2024, ha sido citada en más de 200 resoluciones de distintas audiencias provinciales como motivo de peso para dar o denegar estos permisos, en varias de ellas incluso en contra de la decisión de la prisión (…) El estudio concluyó un total de diez factores de riesgo: extranjería, drogodependencia, profesionalidad delictiva, reincidencia, quebrantamientos previos, si está en régimen cerrado, si ha tenido permisos previos, deficiencia convivencial, lejanía del lugar de disfrute del permiso respecto a la prisión y presiones internas -episodios de amenazas o agresiones con otros internos-. Cada uno con una puntuación del 0 a 3 y un peso específico en la ecuación. El resultado es un valor porcentual que parte del riesgo muy bajo (5%), normal (entre el 20 y el 35%) y alto (a partir del 40%), unas valoraciones mucho más restrictivas que las planteadas inicialmente. Según Clemente, el punto de corte estadístico debería ser 50 sobre 100. "Lo que hace la administración es ponerlo en el 35, de modo que evita que más internos salgan a la calle para evitarse problemas, aunque sea a costa de su rehabilitación", señala el psicólogo. "Lo comenté, pero nos dijeron que eso no era cosa nuestra", añade (pfos 2º, 3º y 6º).
Sobre el origen de esta tabla, véase, entre otras, la publicación, de 1993, de Clemente, et al., titulada *Validación y depuración de la Tabla de Variables de Riesgo en el disfrute de permisos penitenciarios de salida*. Cfr. Castro (2021) y Fanega, Fresneda y Beltrán (2025). Sobre las críticas a las herramientas predictivas en relación con el radicalismo en prisión o delitos yihadistas, para el ordenamiento español, vid. Ministerio del Interior (2021) y Teijón (2022). Cfr. el desarrollo de herramientas en relación con delitos de odio, en Trickett y Bryan (2025).

19. Véanse extractos de la entrevista a una técnica de la Secretaría General de Instituciones Penitenciaria recogidos en el capítulo de Aitor Martínez.

Instituciones Penitenciarias (IIPP) relativas a los permisos obligan a motivar los informes que contienen las decisiones de las Juntas de Tratamiento, en particular, si se desvían de los factores incluidos en las tablas. Estos informes llegan al Juzgado de Vigilancia Penitenciaria, órgano que autoriza la propuesta de concesión de permisos de la Junta, y a las Audiencias Provinciales, las últimas en decidir cuando el preso o la Fiscalía recurren la decisión del Juzgado (tal y como se alude en el análisis jurisprudencial realizado en este capítulo sobre las menciones en recursos a RisCanvi). Sin embargo, para muchos letrados, las resoluciones sobre estos recursos conllevan mucho tiempo y esfuerzo que pueden impactar negativamente en la vida del preso.

Estas críticas a la TVR son comunes en la utilización de RisCanvi, pero se añade un hecho diferencial al recopilar y manejar, mediante una herramienta informática predictiva, un conjunto de datos sobre muchos individuos, bajo control penitenciario, lo que permite juicios probabilísticos sobre el comportamiento de individuos y, en un futuro, elaborar patrones y predicciones con uso de la inteligencia artificial.

Es cierto que la herramienta RisCanvi actual no utiliza, en el momento de publicación de este capítulo, la inteligencia artificial, sino que implica cinco algoritmos[20] que estiman para cada individuo considerado, respectivamente, el riesgo de:

a) violencia autodirigida,

b) violencia intrainstitucional (hacia funcionarios/as del sistema penitenciario),

c) quebrantamiento de condena,

d) reincidencia general y

e) reincidencia violenta.

Criticado por posibles sesgos y falta de transparencia, algo ya indicado y que constituye una crítica generalizada en el uso de algoritmos en la toma

20. La palabra *algoritmo* procede del apellido del matemático persa, nacido en el actual Uzbekistán: Mohamed ben Musa Khwarazmi, arabizado como al Khwarizmi y latinizado como Algoritmi. Este matemático y geógrafo es considerado padre del álgebra y es autor de un tratado sobre resolución de ecuaciones. Vivió, aproximadamente, entre los años 780 y 850. En castellano, es conocido como Al-Juarismi. Su nombre también ha dado lugar a la palabra *guarismo.* Al-Juarismi fue astrónomo y jefe de la biblioteca de la Casa de la Sabiduría de Bagdad. Véase más información en https://www.britannica.com/biography/al-Khwarizmi

de decisiones en el ámbito penal, RisCanvi fue auditado en 2024 (Dribia Data y Direcció General d'Afers Penitenciaris, 2024), lo que supuso su primera auditoría externa encomendada por el gobierno catalán, desde su creación en 2009 y sus mejoras en 2011 y 2017. En el punto cinco de dicha auditoría se expresó que no se detectaron riesgos discriminatorios graves. En todo caso, se recomendó no solo "abrir los datos del modelo para análisis científico y potencialmente al público en general (anonimizados adecuadamente)", sino también "hacer público el código y la documentación del algoritmo para transparencia" (p. 8 y ss.). Además, diversos autores citados en este capítulo también han planteado cuestiones más amplias relativas a la ética, la privacidad y la seguridad de la herramienta, cuestiones especialmente importantes en el uso de la inteligencia artificial que, como ya se ha señalado, RisCanvi todavía no emplea, pero es previsible que lo haga.

Asimismo, en una Jornada realizada en la Facultad de Derecho de Valencia, el 1 de octubre de 2024, juristas y criminólogos discutieron el uso de algoritmos e inteligencia artificial en el sistema de justicia penal, indicando sus potencialidades y riesgos (Varona, 2024). En dicha Jornada se presentó el resultado de una auditoría crítica de Amnistía Internacional[21], dirigida por la profesora Lucía Martínez Garay[22], en relación con tres sistemas predictivos utilizados en la justicia penal española: VioGén[23], RisCanvi y Veri-

21. Véase también el informe de esta misma organización sobre vigilancia policial en los Países Bajos en: https://www.es.amnesty.org/en-que-estamos/noticias/noticia/articulo/paises-bajos-pongan-fin-a-los-peligrosos-experimentos-policiales-de-vigilancia-masiva/

22. Véase su capítulo en esta monografía.

23. El Sistema de Seguimiento Integral en los casos de Violencia de Género (Sistema VioGén), de la Secretaría de Estado de Seguridad del Ministerio del Interior, se define como un "modelo metodológico (que) constituye una herramienta esencial en la protección policial a las víctimas de violencia de género, y un referente a nivel internacional por su carácter dinámico, multidimensional y basado en evidencia empírica". Se puso en funcionamiento el 26 de julio del 2007, en cumplimiento de lo establecido en la Ley Orgánica 1/2004, de 28 de diciembre, de Medidas de Protección Integral contra la Violencia de Género. Incorpora una valoración policial inicial y otra de seguimiento del riesgo y se sigue la lógica de los estudios epidemiológicos y los marcadores de riesgo, sin manejar nociones de causalidad, sino de probabilidad, siendo posible elevar el nivel de riesgo por el agente protector, sin perjuicio de ese juicio probabilístico. Desde su implantación en 2007, el protocolo de valoración policial del riesgo en casos de violencia de género ha sido mejorado, en particular con la revisión introducida por la Instrucción 4/2019 de la Secretaría de Estado de Seguridad. Además, en 2025, se publicó su Protocolo 2025 y se indicó que se había procedido a una mejor calibración de los algoritmos y a una gestión avanzada de notificaciones automatizadas. Según se expresa textualmente en la página web del Ministerio del Interior, donde

Pol[24] (Martínez Garay et al., 2024). En concreto, sobre RisCanvi se criticaba que, entre los factores de riesgo analizados, se aludía a la baja capacidad intelectual, su vida sexual, una opinión mala del gobierno o del sistema judicial o si cumplía condena en una prisión a mucha distancia de su casa. Concluía que podía resultar discriminatoria con las personas con pocos recursos económicos, con enfermedades mentales (sin importar de qué tipo) o con drogodependientes, discriminación que, en todo caso, parece ocurrir en toda administración penitenciaria caracterizada por replicar y ampliar la injusticia social existente fuera de prisión y de la que es responsable toda la sociedad (Baratta, 1982; Díez Ripollés y García España, 2020).

Según la publicación del segundo resumen estadístico detallado del sistema RisCanvi, en el segundo semestre de 2024, de una población peni-

---

no se alude a auditoría externa alguna: "VioGén es, necesariamente, un sistema vivo en evolución constante que ha sido objeto de una reciente renovación. Las mejoras implementadas en VioGén 2 incluyen la optimización de los protocolos de actuación y la mejora en el seguimiento de las víctimas a través de una mayor integración de datos en tiempo real, al facilitar la interoperabilidad con otros sistemas del ámbito policial, judicial, penitenciario, y servicios sociales" (https://sistemaviogen.ses.mir.es/publico/viogen). Respecto de VioGén, en 2022, se publicó, señalando su falta de transparencia, una auditoría externa por parte de la Fundación Éticas y la Fundación Ana Bella (https://eticasfoundation.org/wp-content/uploads/2025/02/ETICAS-Auditoria-Externa-del-sistema-VioGen-_-20220308.docx-2.pdf). Otro estudio de la Asociación Española de Profesores/as de Derecho Administrativo (AEPDA) advierte de fallos en relación con el algoritmo y la falta de transparencia en la evaluación del riesgo (Alonso García, 2025), en particular por la previsible integración de IA (Gordo y Rubio-Martín, 2024). Así mismo, sobre discriminación por género, vid. Libano (2025). El País Vasco ha desarrollado su propio sistema: EBA (*Emakumeen eta Etxekoen Babesa*) y, desde noviembre de 2024, está coordinado con VioGén. Vid. https://www.euskadi.eus/gobierno-vasco/-/noticia/2024/los-sistemas-eba-del-departamento-seguridad-y-viogen-2-del-ministerio-interior-ya-estan-interconectados/. No se conoce ninguna auditoría externa de EBA. En la página web oficial del Departamento de Seguridad únicamente se alude a que la empresa AENOR auditó el proceso conforme a la Norma UNE-EN 9001-2000, certificado el 17 de abril de 2008 y que las víctimas que lo deseen pueden completar un formulario sobre la satisfacción con el sistema (https://www.ertzaintza.euskadi.eus/lfr/web/ertzaintza/violencia-de-genero).

24. VeriPol se utilizó por la policía para predecir denuncias falsas en delitos contra el patrimonio (vid. en https://www.policia.es/_es/comunicacion_prensa_detalle.php?ID=4433&idiomaActual=es#). Esta herramienta empleaba inteligencia artificial en cuanto al aprendizaje automatizado. Fue presentada en 2018 como un algoritmo capaz de detectar denuncias falsas de robo con violencia, con una precisión de más del 90%. Sin embargo, en octubre de 2024, dejó de estar operativa, según confirmó, a la organización *Civio*, el Gabinete Técnico de la Dirección General de Policía. Desde el Ministerio de Interior se señaló que carecía de validez en procedimientos judiciales.

tenciaria total de 8.488 personas en Cataluña, 6.805 fueron evaluadas en el período correspondiente (81%). De estas: 1.069 con escala RC-S (*Screening*) y 5.736 con escala RC-C (Completa). Dentro de estas evaluaciones, el 69% de las escalas RC-S indicaron riesgo bajo en todos los criterios. Solo el 12% de las RC-C obtuvieron la misma clasificación. Globalmente, un 23% de la población evaluada muestra un perfil de riesgo bajo en todos los criterios de predicción. En relación con dicho resumen estadístico, Bataller (2025) concluye lo siguiente, en el sentido de enfocar la utilización de la herramienta hacia un trabajo realmente individualizado, de reinserción y preventivo:

> La presencia de factores de riesgo (FR) múltiples y acumulados confirma una población con perfil criminógeno complejo, donde confluyen elementos estructurales (desempleo, pobreza), personales (trastornos psiquiátricos) y delictivos (historia de violencia, reincidencia). El diseño de políticas debe evitar la aplicación homogénea de programas y apostar por la estratificación del riesgo.
>
> El análisis revela variaciones significativas en la distribución de los factores de riesgo según el tipo de centro (cerrado, preventivo, abierto, jóvenes[25]):
>
> CP Joves: alta prevalencia de FR como el consumo de drogas, falta de planes de futuro, problemas de ocupación y autolesiones.
>
> CP Brians 1: destaca por la baja presencia de conflictos y expedientes disciplinarios (como centro de preventivos).
>
> Centros abiertos: menor presencia general de factores de riesgo (excepto en CO Tarragona y COB2, donde destacan reincidencias concretas).
>
> El informe valida la idea de una criminalidad institucionalmente diferenciada, donde el tipo de delito, el perfil psicológico y las condiciones sociales se distribuyen según el tipo de centro. Este fenómeno apoya la modularización del tratamiento y la necesidad de dispositivos especializados (p.ej., para jóvenes, reincidentes, agresores sexuales o trastornos graves).
>
> El informe permite extraer líneas de actuación estratégicas en materia de intervención penitenciaria desde una perspectiva criminológica. Por ejemplo, fomentar la prevención secundaria: focalización en internos con múltiples FR acumulados (drogas, respuesta limitada al tratamiento, sin apoyo familiar). Además, desarrollar programas específicos: atención diferenciada a jóvenes y reincidentes con carencias estructurales (educación, ocupación). También introducir terapias intensivas dirigidas a centros con alta prevalencia de trastornos de personalidad y autolesiones. Apostar por la justicia restaurativa y promoverla en centros con bajo nivel de conflictos, pero alta presencia de reincidencia. Y, finalmente, una revisión de protocolos: adaptar la valoración RisCanvi en función del perfil institucional (cerrado, abierto, preventivo, etc.) (pfos. 15-17º).

25. Cfr. León y Bartolomé (2024).

1.3. REQUISITOS TECNOLÓGICOS Y JURÍDICOS BÁSICOS PARA UNA POSIBLE ADAPTACIÓN DE RISCANVI

Conviene advertir, de cara a un primer paso imprescindible en cualquier adaptación en otras administraciones penitenciarias, que los sistemas algorítmicos predictivos solo funcionan adecuadamente con la recogida sostenida y el análisis riguroso de grandes cantidades de datos pertinentes y de calidad, así como con recursos informáticos con la potencia y capacidades necesarias para procesarlos. Por tanto, la primera pregunta es si las administraciones interesadas en este tipo de herramientas cuentan con este requisito previo. Además, tal y como recuerda la Recomendación del Consejo de Europa sobre esta materia, según se indicará más adelante:

> Es importante destacar que la calidad de los datos no solo depende de su representatividad general, sino también de que correspondan a la existencia de diferentes grupos minoritarios, de modo que, en última instancia, no sean discriminatorios por su falta de representación[26].

Junto a ello, por mandato legal, es imprescindible minimizar los riesgos ético-jurídicos de la predicción algorítmica, como el título de este libro sugiere y se ha indicado en el epígrafe anterior. Desde un plano jurídico, la introducción de sistemas de predicción algorítmica y con IA, como indica Cañas (2025, pfo. 9º), plantea "un dilema de fondo entre eficiencia tecnológica y la preservación de derechos fundamentales" (referidos principalmente a la individualización de la pena, la presunción de inocencia y la igualdad ante la ley), pero también un profundo debate científico y ético-social. Nellis (2025) se pregunta si, "en un mundo en el que la IA se está volviendo omnipresente, la ética de la predictividad en sí misma es cada vez más difícil de cuestionar" (pfo. 4). Por ello, se debe seguir indagando sobre si las personas internas verán reforzados sus derechos al introducirse un sistema donde existen desequilibrios de poder en un contexto, el penitenciario, que ya los acoge y si la sociedad en su conjunto, las víctimas en particular y el propio sistema penitenciario y sus trabajadores se verán beneficiados de todo ello a medio y largo plazo.

1.4. OBJETIVO Y ESTRUCTURA DEL CAPÍTULO

Este capítulo aspira a abordar solo algunas aristas de esos dilemas y debates. A la luz de las potencialidades beneficiosas para la gestión y valoración del riesgo y de los peligros señalados, las siguientes páginas se refieren

26. O, en sentido inverso, pero igualmente discriminatorio, de sobrerrepresentación.

al marco normativo ineludible que debe respetarse a la hora de adaptar la herramienta RisCanvi por cualquier administración penitenciaria. Para ello, se realiza una revisión bibliográfica básica[27] sobre aspectos normativos de este tipo de sistemas informáticos, comenzando con el marco normativo, principalmente de *soft law*, del Consejo de Europa en el ámbito penal para referirnos después al Reglamento europeo de Inteligencia Artificial y repasar, brevemente, los textos vinculantes que constituyen el marco jurídico sobre protección de datos, en particular, en relación con el uso de perfilaciones y predicciones en el ámbito penitenciario: el Reglamento General de Protección de Datos (RGPD) de 2016 de la UE, la Ley Orgánica 3/2018 de Protección de Datos Personales y Garantía de los Derechos Digitales, la Directiva 2016/680 y la Ley Orgánica 7/2021 sobre protección de datos personales en el ámbito penal, y su conexión con la Ley Orgánica General Penitenciaria (LOGP) y el Reglamento Penitenciario (RP). Todo ello se complementa con un análisis jurisprudencial, junto con un trabajo de campo, desarrollado a lo largo de 2025, consistente en la observación de la toma de decisiones en los equipos técnicos de los tres centros penitenciarios del País Vasco. A ello se suma el análisis de un cuestionario y un grupo de discusión con alumnado universitario de dicha Comunidad Autónoma. Finalmente, se trazan unas conclusiones provisionales sobre el impacto del paso de un juicio clínico a otro probabilístico, o a un sistema mixto, en el ámbito de la toma de decisiones relativas a la gestión y valoración de riesgos en el ámbito penitenciario.

2. MARCO NORMATIVO DEL CONSEJO DE EUROPA EN EL ÁMBITO PENAL SOBRE EL USO DE LA INTELIGENCIA ARTIFICIAL Y TECNOLOGÍAS SIMILARES

If I were a nuclear power plant
they would release me
from security detention.
They would take
the remaining risk.
I am, however, a human being.
(Helmut Pammler)[28]

27. El listado de referencias recoge también algunos textos que, aunque no citados finalmente en el texto, han servido de lecturas básicas para la contextualización del tema objeto de estudio.

28. Poema de una persona en prisión en Alemania, traducido al inglés por el Prof. J. Feest.

Si bien estamos ante *soft law* o normas que no son de obligado cumplimiento, aunque sí configuran estándares de buenas prácticas con un valor ético-jurídico, dentro del Consejo de Europa[29], la *Carta Ética Europea sobre el uso de la inteligencia artificial en los sistemas judiciales y su entorno*[30], firmada el 3 de diciembre de 2018, recoge los siguientes principios:

a) el respeto de los derechos fundamentales;

b) la no discriminación, previniendo específicamente el desarrollo o la intensificación de cualquier discriminación entre individuos o grupos de individuos;

c) la calidad y seguridad respecto al procesamiento de datos y decisiones judiciales;

d) la transparencia, imparcialidad y equidad; y

e) el control del usuario para que sean actores informados y en control de sus elecciones.

En el capítulo 7, sobre cuestiones específicas de la justicia penal, como la prevención de delitos y la evaluación de la reincidencia y del nivel de peligro, se señala que el uso de algoritmos e inteligencia artificial en los procesos penales ha mostrado el riesgo de trasladarnos a posiciones deterministas, en lugar de a una individualización de la sanción, como se recoge en los estándares internacionales en la materia (Cañas, 2025).

Casi seis años tras la publicación de dicha Carta por el Consejo de Europa, el 9 de octubre de 2024, el Consejo de Ministros de los Estados Miembros adoptó la Recomendación CM/Rec(2024)5 relativa a los aspectos éticos y organizativos del uso de la inteligencia artificial y tecnologías

29. Por su parte, no existen normas de las Naciones Unidas, en el marco general o penal. En 2024 un nuevo informe de un grupo de expertos de la ONU afirmó que era necesario regular a escala mundial el floreciente campo de la inteligencia artificial (IA), y añadió que el desarrollo y uso de esta tecnología "no puede dejarse únicamente al capricho de los mercados" ya que afecta a los derechos humanos. Naciones Unidas ha creado el Órgano Asesor de Alto Nivel sobre Inteligencia Artificial.

30. Véase también la Recomendación CM/Rec(2021)8 del Comité de Ministros a los Estados miembros sobre la protección de las personas en relación con el tratamiento automatizado de datos personales en el contexto de la elaboración de perfiles que, si bien, no menciona aspectos penales, sí recoge la posibilidad de excepciones en relación con el concepto de seguridad pública. Cfr. el Convenio de 1981 para la protección de las personas con respecto al tratamiento automatizado de datos de carácter personal, de obligado cumplimiento.

similares en las prisiones y servicios de libertad condicional (Nellis, 2024)[31]. Si bien se reconoce la utilidad de la inteligencia artificial para la mejora de los sistemas de gestión, se subraya que las decisiones siempre habrán de ser tomadas por parte de un ser humano, tratando de minimizar los sesgos, sin perjuicio de que la IA permita manejar más cantidad de datos de forma más rápida.

En el Memorando explicativo de esta Recomendación se subraya *que* el uso de la IA[32], las herramientas algorítmicas y las tecnologías digitales es incipiente, sin que se cuente con políticas o legislación sobre su uso por parte de los servicios penitenciarios y de libertad condicional (Puolakka y Van De Steene, 2021). Se indica expresamente que dicho uso debe reforzar, y no debilitar, el papel fundamental del factor humano. Además, debe haber una reflexión previa sobre su necesidad, proporcionalidad e impacto en términos de mejora de cara a los objetivos legítimos a los que coadyuvan. El Tribunal Europeo de Derechos Humanos, en lo que constituye jurisprudencia vinculante, se refiere a la necesidad, proporcionalidad, legitimidad y control, en concreto a la "necesidad social apremiante" y, en particular, a la proporcionalidad al fin legítimo perseguido, si las razones aducidas por las autoridades nacionales para justificarla son "pertinentes y suficientes" (véanse S. y Marper contra el Reino Unido [GC], nos. 30562/04 y 30566/04, 4 de diciembre de 2008, apartado 101; y Glukhin contra Rusia, n.º 11519/20, 4 de julio de 2023, apartado 78, relativo al uso de la tecnología de reconocimiento facial).

En el Memorando de la Recomendación se indica expresamente que, teniendo en cuenta la Recomendación Rec(2006)2-rev del Comité de Ministros a los Estados miembros sobre las Reglas Penitenciarias Europeas y las sentencias del Tribunal Europeo de Derechos Humanos, deben respetarse los derechos humanos de los reclusos y de todas las personas afectadas por los

31. Véase el capítulo de Fernando Miró en esta monografía. También en relación con los estándares metodológicos y éticos de estas herramientas, véase el capítulo de Izaskun Ibabe.

32. Según el Memorando: "En el contexto del uso de la IA y las tecnologías digitales conexas, existen básicamente dos escenarios: 1) la IA sustituye la toma de decisiones humanas, o 2) la IA asiste la toma de decisiones humanas. Ambos casos están cubiertos por esta norma. El personal también puede verse afectado por el uso de la IA, por ejemplo, en el caso de un sistema de reconocimiento facial basado en IA que detecta agresiones o intentos de suicidio, o el abuso de sustancias". Todas las citas del Memorando en este capítulo son traducciones propias de la autora cuyos originales se encuentran en francés y en inglés. El término *offenders* se traduce aquí como personas ofensoras.

servicios penitenciarios, garantizando que las personas privadas de libertad, o bajo la supervisión de aquellos, conserven todos los derechos que no se vean legalmente implicados por la decisión que les condena o mantiene en prisión o bajo control. Estos principios y, en general el respeto de la dignidad humana, deben permanecer intactos cuando se utilicen la IA y las tecnologías relacionadas en las prisiones y en los servicios de libertad condicional, algo que también afecta a las empresas que las desarrollen y apliquen, debiendo verse sometidas a auditorías y, en su caso a responsabilidad civil o penal si causan daños, en particular aquellos relativos a posibles sesgos. Esos sesgos pueden plantear una cuestión de discriminación en virtud del artículo 14 del Convenio Europeo de Derechos Humanos, como ya sucedió en el caso Basu contra Alemania (n.º 215/19, de 18 de octubre de 2022). De hecho, el enfoque de derechos humanos[33] en el desarrollo de estas herramientas debe verse como algo básico y enriquecedor, siempre y cuando esté informada la sociedad y, en particular, los internos, posibilitando, en la medida de lo posible, su participación respecto del proceso de decisión y, en todo caso, en su cuestionamiento.

En relación con esto último, siguiendo con el articulado de la Recomendación, como mínimo, debería haber disposiciones sobre el acceso a un recurso efectivo ante una autoridad competente (incluidas las autoridades judiciales y las autoridades de control de la protección de datos)[34]; el derecho a que las decisiones adoptadas o basadas en la IA y las tecnologías digitales conexas sean revisadas por un ser humano; y la obligación de las autoridades públicas de llevar a cabo una revisión humana adecuada de los procesos basados en la IA y las tecnologías digitales conexas y de proporcionar a las personas físicas o jurídicas pertinentes información significativa sobre el papel de la IA en la adopción de decisiones que les afecten o en la información sobre dichas decisiones, si bien el Memorando prevé posibles excepciones cuando existan motivos legítimos imperativos que excluyan o limiten dicha revisión o divulgación.

El Memorando reconoce que el contacto humano es esencial en el ámbito penitenciario y que, en muchos territorios se cuenta con recursos escasos de personal, un personal que debe ser protagonista también de cualquier incorporación tecnológica nueva, con una formación adecuada y una par-

33. En el marco de las Naciones Unidas, véanse las Normas Mandela y, en el marco del Consejo de Europa, vid. las Reglas Europeas de Conducta para las Prisiones y las Reglas Europeas sobre las sanciones y medidas comunitarias.

34. Véase la referencia en este capítulo a la sentencia del Tribunal Supremo español en el caso BOSCO.

ticipación en el desarrollo y toma de decisiones[35], además de una reflexión sobre el impacto en su profesión y trabajo.

Por todo ello, el Memorando se mueve entre el reconocimiento de las ventajas del uso de herramientas algorítmicas y la IA[36], en concreto, y sus desventajas o riesgos. Entre estos últimos, señala dos propios de los entornos penitenciarios:

a) La instrumentalización o degradación de las relaciones entre el personal y las personas ofensoras si, en lugar de tratarlos de forma genuinamente interpersonal, el contacto se media cada vez más a través de máquinas que recopilan y codifican datos sobre ellos en cada encuentro (o incluso de forma constante, si se les vigila con dispositivos de seguimiento).

b) La supervisión del rendimiento y la productividad de los empleados en los lugares de trabajo puede aumentar enormemente si se utilizan sensores (portátiles o integrados en edificios y equipos) y sistemas de *software* (no necesariamente IA completa) para recopilar, analizar y comparar datos con un grado de detalle sin precedentes.

La idea de progreso de la IA es que el personal podrá dedicar menos tiempo a actividades rutinarias y más a actividades relevantes, pero, como se indica en el Memorando del Consejo de Europa, la cuestión está en la definición de esos adjetivos (rutinarias y relevantes) y, en particular, en el significado profundo de humanidad que puede existir por debajo de algunas actividades calificadas como rutinarias que pueden constituir, precisamente, en momentos de intercambio de conversación o de contacto físico, fundamentales en todo proceso de reinserción.

Ahora bien, como ya se ha indicado, el Tribunal Europeo de Derechos Humanos ha dictaminado que el tratamiento de los datos personales de las

35. Cfr. el Código Europeo de Ética para el Personal Penitenciario y las Normas del Consejo de Europa sobre la libertad vigilada, con especificaciones detalladas sobre la ética profesional y la formación del personal, según señala el propio Memorando del Consejo de Europa.

36. En el Memorando se alude a que: "la IA puede agilizar el ciclo de gestión de las personas ofensoras, hacerlo más rentable y optimizar la compatibilidad de los servicios y las necesidades de las personas ofensoras (…) Finlandia también está desarrollando una nueva herramienta de gestión de personas ofensoras, RISE AI, que será un componente basado en la IA del nuevo sistema de gestión de las personas ofensoras para ayudar a evaluar a las personas ofensoras y orientarlos hacia los servicios y unidades más adecuados durante su condena".

personas internas debe realizarse de conformidad con la ley, que solo puede perseguir objetivos legítimos específicos y que debe ser proporcionado, es decir, necesario para alcanzar dichos objetivos (Nuh Uzun y otros contra Turquía, nºs. 49341/18, 29 de marzo de 2022, apartado 83). Además de tener en cuenta los costes de cualquier sistema de gestión y valoración de riesgos[37], lo que incluye su protección respecto de riesgos a la privacidad y posibles ataques a las bases de datos, en el Memorando se especifica claramente que:

> El tratamiento de datos debe ser leal, lícito y proporcionado en relación con el fin específico y legítimo perseguido y reflejar en todas las fases del tratamiento un equilibrio justo entre todos los intereses en juego, ya sean públicos o privados, y los derechos y libertades en juego.

Considerando que, en el ámbito penitenciario, existe un desequilibrio de poder entre los servicios penitenciarios y de libertad condicional y las personas condenadas, el consentimiento no puede constituir por sí mismo una base jurídica suficiente, sin perjuicio de ponderar la necesidad de protección de la sociedad. Se observa, de nuevo, que, en este sector, se permiten excepciones que legitiman algunas actividades que afectan al derecho a la intimidad, priorizando el interés público o la seguridad pública, siempre respetando el contenido mínimo de los derechos fundamentales y ponderando los principios de legitimidad, necesidad, proporcionalidad y eficacia.

Sin embargo, en este balanceo complejo, el Memorando es claro:

> La IA no debe sustituir a los seres humanos en los procesos de toma de decisiones, sino que debe funcionar como una herramienta para los seres humanos, proporcionando precedentes, recomendaciones u opciones para un curso de acción concreto, dejando que los profesionales y los gestores humanos tomen las decisiones finales basándose en datos e información más precisos, completos y objetivos que los recopilados con los métodos tradicionales.

Además:

> El diseño de un algoritmo para su uso en el contexto penitenciario y de libertad condicional requiere ser preciso sobre lo que se quiere conseguir y comprender los sesgos típicos del pensamiento humano[38] que se supone que la IA debe sustituir (Fry, 2018). En el mejor de los casos, los algoritmos podrían superar los efectos perjudiciales de los sesgos cognitivos (Sunstein, 2018) en lugar de repetirlos[39].

37. En este sentido, la calidad de los datos debe formar parte de un ciclo de evaluación y valoración continuas.
38. Pueden mencionarse sesgos de confirmación, de anclaje, de exceso de confianza, etc.
39. Las citas bibliográficas son las que recoge el propio Memorando.

Conviene recordar que la mayoría de las herramientas de evaluación de riesgos en el sistema penitenciario tienden a seguir el modelo dominante de evaluación de riesgos basado en el riesgo, la necesidad y la capacidad de respuesta (RNR, por sus siglas en inglés) (Andrews, Bonta y Hoge, 1990), el cual también se ha enfrentado con diversas críticas (Larsen et al., 2025). En todo caso, como reconoce el Memorando del Consejo de Europa, toda evaluación de riesgos es un proceso dinámico e inacabado que debe basarse no solo en la evitación de riesgos, sino en el apoyo para el cambio que siempre, por su propia definición en comportamientos individuales y sociales complejos, contendrá incertidumbres. En relación con ello, en el Memorando se reconoce que todas las normas se quedarán obsoletas dado el rápido avance tecnológico de este tipo de herramientas y su impacto, a nivel económico y social, lo que obliga a una revisión periódica de las mismas y, en caso necesario, a su modificación.

Finalmente, dentro del Consejo de Europa, es necesario citar el *Convenio Marco del Consejo de Europa sobre Inteligencia Artificial y Derechos Humanos, Democracia y Estado de Derecho*, de 2024. Una vez entre en vigor, constituirá el primer tratado internacional jurídicamente vinculante en este ámbito, es decir, ya no estaríamos ante *soft law*. Ahora bien, no tiene referencias expresas al ámbito penal o penitenciario, aunque sí se indica que las Partes de la Convención Marco no están obligadas a aplicar las disposiciones del tratado a las actividades relacionadas con la protección de sus intereses de seguridad nacional, siempre con garantía de que dichas actividades respeten el derecho internacional y las instituciones y procesos democráticos. Esta Convención Marco se aplicaría, en todo caso, cuando los sistemas de IA pudieran interferir con los derechos humanos, la democracia o el Estado de Derecho. El objetivo es garantizar que las actividades relacionadas con el ciclo de vida de los sistemas de inteligencia artificial sean plenamente compatibles con aquellos, al tiempo que no se frene el progreso tecnológico y la innovación. Abierto a la firma el 5 de septiembre de 2024, fue firmado por la UE[40].

3. MARCO JURÍDICO SOBRE LA INTELIGENCIA ARTIFICIAL EN LA UNIÓN EUROPEA

Aunque, como se ha reiterado en esta monografía, RisCanvi, en estos momentos, no integra la inteligencia artificial (IA), es muy posible que lo

40. Decisión (EU) 2024/2218 del Consejo, de 28 de agosto de 2024, relativa a la firma, en nombre de la Unión Europea, del Convenio Marco del Consejo de Europa sobre Inteligencia Artificial, Derechos Humanos, Democracia y Estado de Derecho.

haga en un futuro, de ahí que sea necesario considerar el marco jurídico regulador específico. En el ámbito de la Unión Europea, cuatro años más tarde del Libro Blanco sobre la inteligencia artificial, de febrero de 2020, se publicó el Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.º 300/2008 (UE) n.º 167/2013 (UE) n.º 168/2013 (UE) 2018/858 (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE (UE) 2016/797 y (UE) 2020/1828. Este Reglamento se conoce como Reglamento o Ley de Inteligencia Artificial consolida una visión de precaución respecto del uso de la inteligencia artificial[41]. El Reglamento de IA entró en vigor el 1 de agosto de 2024 y, como ya se ha indicado, será plenamente aplicable dos años después, el 2 de agosto de 2026, con algunas excepciones:

- Las prohibiciones y las obligaciones en materia de conocimientos básicos sobre IA, con una entrada en vigor prevista para el 2 de febrero de 2025.
- Las normas de gobernanza y las obligaciones relativas a los modelos GPAI[42], con una entrada en vigor prevista para el 2 de agosto de 2025.
- Las normas para los sistemas de IA de alto riesgo, integrados en productos regulados, tienen un período de transición ampliado hasta el 2 de agosto de 2027.

Esta normativa europea se adaptará en España mediante una futura Ley para el buen uso y la gobernanza de la Inteligencia Artificial. El 11 de marzo de 2025, el Consejo de Ministros aprobó el anteproyecto de dicha Ley[43], la cual se justifica para garantizar que el uso de la inteligencia artificial sea "ético, inclusivo y beneficioso para las personas". El anteproyecto de Ley

41. Vid. el Reglamento, en todos los idiomas de la UE, con documentación complementaria, en la página oficial en https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai. Véase en ella el Código de buenas prácticas de modelos de IA de uso general (GPAI, por sus siglas en inglés). En 2025 la Comisión Europeo recibió presiones, principalmente por parte de los "lobbies" tecnológicos de los EE. UU. en las negociaciones comerciales, para frenar la aplicación de este Reglamento (Barrio, 2025).
42. Un modelo GPAI se define como cualquier modelo entrenado utilizando más de 10^{23} FLOPS (operaciones en coma flotante por segundo) y capaz de generar salidas de lenguaje (texto/audio), de texto a imagen o de texto a vídeo.
43. Véase en https://digital.gob.es/comunicacion/notas-prensa/mtdfp/2025/03/2025-03-11

consta de treinta y siete artículos agrupados en cuatro capítulos, dos disposiciones adicionales, cuatro disposiciones finales y un anexo.

El Reglamento (UE) de Inteligencia Artificial europeo prohíbe, entre otras actividades, la puesta en servicio o la utilización de sistemas de IA para evaluar o clasificar a personas físicas o a colectivos de personas durante un período determinado de tiempo, atendiendo a su comportamiento social o a características personales, o de su personalidad, conocidas, inferidas o predichas, de forma que la puntuación ciudadana resultante provoque una o las dos situaciones siguientes:

(i) un trato perjudicial o desfavorable hacia determinadas personas físicas o colectivos de personas en contextos sociales que no guarden relación con los contextos donde se generaron o recabaron originalmente;

(ii) un trato perjudicial o desfavorable hacia determinadas personas físicas o colectivos de personas que sea injustificado o desproporcionado con respecto a su comportamiento social o la gravedad de éste.

En el Considerando 42 del Reglamento se alude al principio básico de la presunción de inocencia:

> En consonancia con la presunción de inocencia, las personas físicas de la Unión siempre deben ser juzgadas basándose en su comportamiento real. Las personas físicas nunca deben ser juzgadas a partir de comportamientos predichos por una IA basados únicamente en la elaboración de sus perfiles, en los rasgos o características de su personalidad, como la nacionalidad, el lugar de nacimiento, el lugar de residencia, el número de hijos, el nivel de endeudamiento o el tipo de vehículo, sin una valoración humana y sin que exista una sospecha razonable, basada en hechos objetivos comprobables, de que dicha persona está implicada en una actividad delictiva.

El Reglamento realiza una clasificación de cuatro tipos de riesgos (Ebers, 2024), ampliamente aludida en esta monografía, a saber:

- Riesgo mínimo donde simplemente se recomiendan códigos concretos de conducta para las empresas que utilizan herramientas de inteligencia artificial.
- Riesgo específico de transparencia donde se debe informar claramente de la interacción con inteligencia artificial o de la generación de contenidos con la misma.

- Riesgo alto, **como** en sanidad o contratación, donde se deben cumplir con sistemas de reducción del riesgo, con conjuntos de datos de alta calidad e información clara para el usuario o supervisión humana, entre otros requisitos de transparencia y calidad. Los modelos predictivos en el ámbito judicial y penitenciario se encontrarían dentro de usos de inteligencia artificial con riesgo alto.
- Riesgo inadmisible cuando existe una clara amenaza para los derechos fundamentales de las personas y el uso de la inteligencia artificial se prohíbe, como ya se establecía en la normativa de protección de datos aludida en el epígrafe anterior.

En definitiva, debe prestarse atención al riesgo inadmisible y al riesgo alto, simbolizados en las dos imágenes siguientes de la pirámide de escalada del riesgo, debiendo ser excepcionales y, en el caso del riesgo alto, acompañados por medidas ineludibles y cuya falta de cumplimiento es supervisada y sancionada.

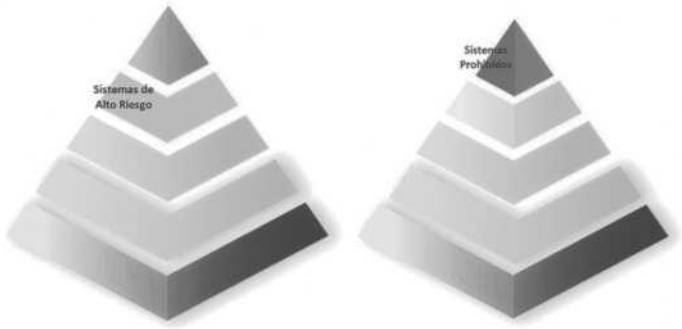


Imagen 1: *Pirámide de riesgo*. Fuente: https://aesia.digital.gob.es/es/actualidad/recursos/ria-sistemas-prohibidos

En concreto, según el artículo 5. 1 d), constituyen prácticas de IA prohibidas (Cotino, 2025):

> La introducción en el mercado, la puesta en servicio para este fin específico o el uso de un sistema de IA para realizar evaluaciones de riesgos de personas físicas con el fin de valorar o predecir el riesgo de que una persona física cometa un delito basándose únicamente en la elaboración del perfil de una persona física o en la evaluación de los rasgos y características de su personalidad; esta prohibición no se aplicará a los sistemas de IA utilizados para apoyar la valoración humana de la implicación de una persona en una actividad delictiva que ya se base en hechos objetivos y verificables directamente relacionados con una actividad delictiva.

En 2025, el *Innovation Lab* de Europol publicó una guía práctica titulada *AI bias in law enforcement*, centrada en la reducción de sesgos algorítmicos en el ámbito policial, ante datos incompletos o prejuiciosos, con base en los principios de la normativa de la Unión Europea.

El informe propone recomendaciones prácticas que, en relación con este capítulo, pueden adoptarse también en el ámbito penitenciario. Destacamos seis de ellas que fueron puestas de relieve en el blog del Departamento de Interior de la Generalitat catalana, en julio de 2025[44]:

- Obligatoriedad de auditorías independientes de los sistemas de inteligencia artificial, antes y durante su uso.
- Supervisión humana constante y capacidad de intervenir en las decisiones automatizadas.
- Análisis crítico de los datos de entrenamiento y especial atención a posibles fuentes de discriminación.
- Asegurar la diversidad y la ética en los equipos de desarrollo y uso de la inteligencia artificial.
- Garantizar la transparencia y asegurar que las decisiones de los sistemas sean comprensibles tanto para los operativos como para la ciudadanía.
- Establecer protocolos de revisión continua para evaluar el impacto de la inteligencia artificial a largo plazo.

En definitiva, la ética, la responsabilidad y los derechos humanos, sin perjuicio de conflictos de derechos o valores, no pueden olvidarse en cualquier política de seguridad pública, incluyendo la política penitenciaria (Solar Calvo y Lacal Cuenca, 2025). Para ello, la Agencia Española de Supervisión de Inteligencia Artificial (AESIA) puede tener un papel fundamental. Creada antes de la entrada en vigor del Reglamento Europeo, según se indica en su página web[45]:

> La Agencia Española de Supervisión de Inteligencia Artificial es el organismo público encargado de garantizar el uso ético y seguro de la inteligencia artificial en España. Su principal misión es asegurar que tanto entidades públicas como privadas cumplan con la normativa vigente, protegiendo la privacidad, igualdad de trato y derechos fundamentales.

44. Véase en https://notesdeseguretat.blog.gencat.cat/2025/07/21/como-podemos-evitar-los-sesgos-de-la-inteligencia-artificial-en-la-policia-una-guia-practica-de-la-europol/
45. Véase en https://aesia.digital.gob.es/es

4. MARCO JURÍDICO BÁSICO, COMUNITARIO Y ESPAÑOL, SOBRE PROTECCIÓN DE DATOS, EN PARTICULAR EN RELACIÓN CON EL USO DE PERFILACIONES Y PREDICCIONES EN EL ÁMBITO PENITENCIARIO: EL REGLAMENTO GENERAL DE PROTECCIÓN DE DATOS (RGPD) DE 2016 DE LA UE, LA LEY ORGÁNICA 3/2018 DE PROTECCIÓN DE DATOS PERSONALES Y GARANTÍA DE LOS DERECHOS DIGITALES, LA DIRECTIVA 2016/680 Y LA LEY ORGÁNICA 7/2021 SOBRE PROTECCIÓN DE DATOS PERSONALES EN EL ÁMBITO PENAL, Y SU CONEXIÓN CON LA LEY ORGÁNICA GENERAL PENITENCIARIA Y EL REGLAMENTO PENITENCIARIO

Aunque suele limitarse a la noción de policía predictiva, RisCanvi también puede relacionarse, al menos en un sentido extenso, con la idea de la perfilación delictiva en tanto en cuanto implica contrastar determinadas variables individuales con una predicción de riesgo de diversos comportamientos violentos (o quebrantamientos de condena), con base en la información suministrada de una recopilación de datos longitudinales del conjunto de personas internas en prisión. Por lo tanto, lo primero que debe considerarse es el marco regulador de la recogida y tratamiento de dichos datos, con pleno respeto de la normativa de la Unión Europea y en la normativa estatal.

4.1. NORMATIVA EUROPEA Y ESPAÑOLA GENERAL SOBRE PROTECCIÓN DE DATOS

Tanto el Reglamento General de Protección de Datos (RGPD) de la UE (art. 4. 4)[46] como la LO 7/2021 (art. 5 d) se refieren a la perfilación como cualquier tratamiento automatizado de datos personales para evaluar o predecir aspectos personales (Cueto, 2025; Pereira, 2024; Pérez Salazar, 2024; Sachoulidou,

46. Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento General de Protección de Datos). Según se expresa textualmente en su artículo 4, la "elaboración de perfiles" se refiere a: 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.

2023; Chelioudakis, 2020; Zeng Ustun y Rudin, 2017; Zouave y Marquenie, 2017; Bosco et al., 2014). En relación con ello, en línea con los estándares internacionales citados, ambas normas reconocen el derecho a no ser objeto de decisiones que produzcan efectos jurídicos, o una afectación significativa, que estén basadas únicamente en el tratamiento automatizado. Se trata de una prohibición contenida en el artículo 22 del RGPD[47] y en el artículo 14 de la LO 7/2021[48], los cuales también contemplan excepciones específicas.

Asimismo, según el artículo 23 de la Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación[49]:

> *1. En el marco de la Estrategia Nacional de Inteligencia Artificial, de la Carta de Derechos Digitales y de las iniciativas europeas en torno a la Inteligencia Artificial, las administraciones públicas favorecerán la puesta en marcha de mecanismos para que los algoritmos involucrados en la toma de decisiones que se utilicen en las administraciones públicas tengan en cuenta criterios de minimización de sesgos, transparencia y rendición de cuentas, siempre que sea factible técnicamente. En estos mecanismos se incluirán su diseño y datos de entrenamiento, y abordarán su potencial impacto discriminatorio. Para lograr este fin, se promoverá la realización de evaluaciones de impacto que determinen el posible sesgo discriminatorio. 2. Las administraciones públicas, en el marco de sus competencias en el ámbito de los algoritmos involucrados en procesos de toma de decisiones, priorizarán la transparencia en el diseño y la implementación y la capacidad de interpretación de las decisiones adoptadas por los mismos. 3. Las administraciones públicas y las empresas promoverán el uso de una Inteligencia Artificial ética, confiable y respetuosa con los derechos fundamentales, siguiendo especialmente las recomendaciones de la Unión Europea en este sentido. 4. Se promoverá un sello de calidad de los algoritmos.*

Según indica la Agencia Española de Protección de Datos, con relación al concepto "decisión basada únicamente en el tratamiento automatizado", este se encuentra desarrollado en el apartado iv.a de las Directrices WP 251 sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679, revisadas por última vez y adoptadas el 6 de febrero de 2018, por el Grupo de protección de las personas en lo que

47. Art. 22 RGPD: "todo interesado tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar".

48. Según el art. 14, se prohíben: "decisiones basadas únicamente en un tratamiento automatizado, incluida la elaboración de perfiles, que produzcan efectos jurídicos negativos para el interesado o que le afecten significativamente".

49. Este artículo debe leerse en relación con la normativa de la UE. En relación con la discriminación por diversidad funcional en prisión, en una comparativa entre el garantismo de las normas europeas y estadounidenses, cfr. Perlin, Ellis Cucolo y Kulesza (2025).

respecta al tratamiento de datos personales, creado por la Directiva 95/46/ce del Parlamento Europeo y del Consejo, de 24 de octubre de 1995:

> *El artículo 22, apartado 1, se refiere a las decisiones "basadas únicamente" en el tratamiento automatizado. Esto quiere decir que no hay participación humana en el proceso de decisión.*
>
> *El responsable del tratamiento no puede obviar las disposiciones del artículo 22 inventándose una participación humana. Por ejemplo, si alguien aplica de forma rutinaria perfiles generados automáticamente a personas sin que ello tenga influencia real alguna en el resultado, esto seguiría siendo una decisión basada únicamente en el tratamiento automatizado.*
>
> *Para ser considerada como participación humana, el responsable del tratamiento debe garantizar que cualquier supervisión de la decisión sea significativa, en vez de ser únicamente un gesto simbólico. Debe llevarse a cabo por parte de una persona autorizada y competente para modificar la decisión. Como parte del análisis, debe tener en cuenta todos los datos pertinentes.*
>
> *El responsable del tratamiento debe identificar y registrar el grado de participación humana en el proceso de toma de decisiones y en qué punto se produce esta.*

Muchas decisiones automatizadas en realidad involucran cierto grado de intervención humana, sin embargo, para considerarse como tal, tiene que ser activa y no solo un gesto simbólico, es decir, tiene que tener un grado determinado de relevancia y capacidad. De nuevo:

> *Si un ser humano revisa y tiene en cuenta otros factores para tomar la decisión final, dicha decisión no estará "basada únicamente" en el tratamiento automatizado.*

En el caso que nos ocupa, desde la Generalitat de Catalunya se afirma que la herramienta RisCanvi no supone tomar decisiones basadas únicamente en un tratamiento automatizado. Sin embargo, no queda claro, en cada caso concreto, el peso de dicho tratamiento en la toma de esas decisiones.

Recientemente, no tanto en lo que respecta a la toma de decisiones, sino a su transparencia, cobra especial relevancia una sentencia del Tribunal Supremo español (n.º 1119/2025, de 11 de septiembre) que obliga a facilitar el código fuente de la aplicación informática BOSCO, utilizada por el Ministerio para la Transición Ecológica y el Reto demográfico para que las empresas comercializadoras de energía eléctrica puedan comprobar si las personas solicitantes cumplen con los requisitos legales para ser beneficiarias del bono social. La resolución estima el recurso de casación interpuesto por la Fundación Ciudadana *Civio* y sienta doctrina sobre el derecho a conocer los algoritmos y programas que condicionan decisiones administrativas de gran impacto social y que afectan a la vida de miles de personas.

En 2018, *Civio* solicitó al MITECO el acceso al código fuente, especificaciones técnicas y pruebas de BOSCO. El ministerio no contestó la solicitud de acceso a la información. Interpuesta reclamación, el Consejo de Transparencia y Buen Gobierno (CTBG), en su resolución R/0701/2018, de 18 de febrero de 2019, concedió parte de la información —documentación técnica y pruebas— pero negó el código fuente, alegando que debía protegerse la propiedad intelectual en virtud del artículo 14.1.j) de la Ley 19/2013, de 9 de diciembre, de Transparencia, Acceso a la Información Pública y Buen Gobierno (LTAIBG). El Juzgado Central de lo Contencioso-Administrativo n.º 8 y la Audiencia Nacional confirmaron esa interpretación. Sin embargo, el Tribunal Supremo ha entendido el derecho de acceso a la información pública como un derecho constitucional ejercitable por la entidad. Evidentemente, la sentencia tiene implicaciones que van más allá del bono social eléctrico ya que puede afectar al futuro del uso de algoritmos, inteligencia artificial y *software* automatizado en la gestión pública, reforzando la exigencia de transparencia y control de la ciudadanía (Pedro, 2025). Por ende, podría afectar al uso de RisCanvi o cualquier herramienta similar en un futuro, en particular en lo que se refiere al deber de transparencia cuando afecta a derechos fundamentales de un número importante de personas internas o bajo control penitenciario que puede considerarse como colectivo vulnerable.

4.2. NORMATIVA EUROPEA Y ESPAÑOLA DE PROTECCIÓN DE DATOS EN EL ÁMBITO PENAL-PENITENCIARIO: LA DIRECTIVA 2016/680, LA LEY ORGÁNICA 7/2021 SOBRE PROTECCIÓN DE DATOS PERSONALES EN EL ÁMBITO PENAL Y EL REGLAMENTO PENITENCIARIO ESPAÑOL

La Directiva (UE) 2016/680 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales por parte de las autoridades competentes para fines de prevención, investigación, detección o enjuiciamiento de infracciones penales o de ejecución de sanciones penales, y a la libre circulación de dichos datos (Montoro Sánchez, 2023), derogó la Decisión Marco 2008/977/JAI del Consejo y fue adaptada en nuestro ordenamiento mediante la Ley Orgánica 7/2021 sobre protección de datos personales en el ámbito penal[50]. Esta Ley, mediante su disposición final primera, modificó

50. En el Preámbulo de dicha Ley se indica expresamente que: "Resulta pertinente poner de relieve que la citada directiva que transpone esta Ley Orgánica se aprobó como respuesta a las crecientes amenazas para la seguridad en el contexto nacional e internacional, que tienen, en numerosos casos, un componente transfronterizo. Por esta

la Ley Orgánica 1/1979, de 26 de septiembre, la Ley Orgánica 1/1979, de 26 de septiembre, General Penitenciaria, con la introducción de un nuevo artículo 15 bis sobre tratamientos de datos de carácter personal, el cual establece lo siguiente:

1. Admitido en el establecimiento un interno, se procederá a verificar su identidad personal, efectuando la reseña alfabética, dactilar y fotográfica, así como a la inscripción en el libro de ingresos y a la apertura de un expediente personal relativo a su situación procesal y penitenciaria, respecto del que se reconoce el derecho de acceso. Este derecho solo se verá limitado de forma individualizada y fundamentada en concretas razones de seguridad o tratamiento.

2. El tratamiento de los datos personales de los internos se regirá por lo previsto en la Ley Orgánica de protección de datos personales tratados para fines de prevención, detección, investigación y enjuiciamiento de infracciones penales y de ejecución de sanciones penales. Los datos personales de categorías especiales que no figuren en el apartado anterior se podrán tratar con el consentimiento del interesado. Sólo se prescindirá de dicho consentimiento cuando sea estrictamente necesario y se efectúe con las garantías adecuadas para proteger el derecho a la protección de datos de los interesados, atendiendo al tipo de datos que se traten y a las finalidades de los distintos tratamientos dirigidos a la ejecución de la pena.

Asimismo, siguiendo la normativa europea y española, citada anteriormente, según el artículo 6 del Reglamento Penitenciario español (Real Decreto 190/1996, de 9 de febrero, por el que se aprueba el Reglamento Penitenciario)[51], sobre limitación del uso de la informática penitenciaria (Simón, 2022):

1. Ninguna decisión de la Administración penitenciaria que implique la apreciación del comportamiento humano de los reclusos podrá funda-

---

razón, la cooperación internacional y la transmisión de información de carácter personal entre los servicios policiales y judiciales de los países implicados se convierten en un objetivo ineludible. En efecto, los atentados terroristas de Nueva York en 2001 supusieron un punto de inflexión en la necesidad de reforzar la cooperación judicial y policial en la lucha contra el terrorismo, como volvería a ponerse de manifiesto con ocasión de los atentados de Bruselas y Niza en 2016".

51. El apartado 2 de este artículo fue modificado por el artículo único 1 del Real Decreto 419/2011, de 25 de marzo, cuando también se añadió el 4. Por su especial relevancia para este estudio, reproducimos a continuación los artículos 7 a 9 del Reglamento penitenciario.
Artículo 7 Recogida y cesión de datos de carácter personal de los internos, redactado por el número dos del artículo segundo del R.D. 268/2022, de 12 de abril, por el que

mentarse, exclusivamente, en un tratamiento automatizado de datos o informaciones que ofrezcan una definición del perfil o de la personalidad del interno.

---

se modifica el R.D. 190/1996, de 9 de febrero, por el que se aprueba el Reglamento Penitenciario («B.O.E.» 13 abril).

1. Cuando los datos de carácter personal de los reclusos se recojan para el ejercicio de las funciones propias de la Administración penitenciaria no será preciso el consentimiento del interno afectado, salvo en los supuestos recogidos en el artículo 15 bis de la Ley Orgánica 1/1979, de 26 de septiembre.
2. Tampoco será preciso el consentimiento del recluso afectado para ceder a otras Administraciones públicas, en el ámbito de sus respectivas competencias, los datos de carácter personal contenidos en los ficheros informáticos penitenciarios que resulten necesarios para que éstas pueden ejercer sus funciones respecto de los internos en materia de servicios sociales, Seguridad Social, custodia de menores u otras análogas.
3. También se podrán ceder datos de carácter personal contenidos en los ficheros informáticos penitenciarios sin previo consentimiento del afectado cuando la cesión tenga por destinatarios al Defensor del Pueblo o institución análoga de las Comunidades Autónomas que ejerzan competencias ejecutivas en materia penitenciaria, al Ministerio Fiscal o a los Jueces o Tribunales, en el ejercicio de las funciones que tienen atribuidas, así como cuando se trate de cesión de datos de carácter personal relativos a la salud de los reclusos por motivos de urgencia o para realizar estudios epidemiológicos.
4. La cesión de datos a los que se refieren los apartados 2 y 3 de este artículo se hará siempre de acuerdo con lo previsto en la Ley Orgánica 7/2021, de 26 de mayo, de protección de datos personales tratados para fines de prevención, detección, investigación y enjuiciamiento de infracciones penales y de ejecución de sanciones penales.
5. Las transferencias internacionales de datos de carácter personal contenidos en los ficheros informáticos penitenciarios se efectuarán en los supuestos de prestación de auxilio judicial, de acuerdo con lo establecido en los tratados o convenios en los que sea parte España o en la normativa de la Unión Europea que sea de aplicación.

Artículo 8 Datos penitenciarios especialmente protegidos

1. No obstante lo dispuesto en el artículo anterior, los datos de carácter personal de los reclusos relativos a opiniones políticas, a convicciones religiosas o filosóficas, al origen racial y étnico, a la salud o a la vida sexual, que hayan sido recabados para formular los modelos individualizados de ejecución o los programas de tratamiento penitenciarios, sólo podrán ser cedidos o difundidos a otras personas con el consentimiento expreso y por escrito del recluso afectado o cuando por razones de interés general así lo disponga una Ley. (Cfr. el artículo 6 del Convenio de 1981, del Consejo de Europa, para la protección de las personas con respecto al tratamiento automatizado de datos de carácter personal (Convenio 108), el cual establece lo siguiente: «Los datos personales que revelen el origen racial, las opiniones políticas o las creencias religiosas u otras creencias, así como los datos personales relativos a la salud o la vida sexual, no podrán ser objeto de tratamiento

2. La recogida, tratamiento automatizado y cesión de los datos de carácter personal de los reclusos contenidos en los ficheros se efectuará de acuerdo con lo establecido en la legislación sobre protección de datos de carácter personal y sus normas de desarrollo.

3. Las autoridades penitenciarias responsables de los ficheros informáticos penitenciarios adoptarán las medidas de índole técnica y organizativa necesarias para garantizar la seguridad de los datos de carácter personal en ellos contenidos, así como para evitar su alteración, pérdida, tratamiento o acceso no autorizado, y estarán obligadas, junto con quienes intervengan en cualquier fase del tratamiento automatizado de este tipo de datos, a guardar secreto profesional sobre los mismos, incluso después de que haya finalizado su relación con la Administración penitenciaria.

4. La Administración penitenciaria podrá establecer ficheros de internos que tengan como finalidad garantizar la seguridad y el buen orden del establecimiento, así como la integridad de los internos. En ningún caso la inclusión en dicho fichero determinará por sí misma un régimen de vida distinto de aquél que reglamentariamente corresponda.

---

automatizado, salvo que la legislación nacional prevea garantías adecuadas. Lo mismo se aplicará a los datos personales relativos a condenas penales»).

2. Cuando se soliciten de la Administración Penitenciaria este tipo de datos especialmente protegidos por medio de representante del recluso, deberá exigirse, en todo caso, poder especial y bastante otorgado por el mismo en el que conste expresamente su consentimiento para que su representante pueda tener acceso a dichos datos personales del recluso.

Artículo 9 Rectificación y conservación de los datos

1. Los reclusos podrán solicitar de la Administración penitenciaria la rectificación de sus datos de carácter personal contenidos en los ficheros informáticos penitenciarios que resulten inexactos o incompletos. De la rectificación efectuada se informará al interesado en el plazo máximo de dos meses desde su solicitud, así como al cesionario o cesionarios, en el supuesto de que los datos incorrectos hubiesen sido objeto de cesión previa.
2. Los datos de carácter personal de los reclusos contenidos en los ficheros informáticos penitenciarios no serán cancelados cuando, ponderados los intereses en presencia, concurran razones de interés público, de seguridad y de protección de los derechos y libertades de terceros, así como cuando posean un valor intrínseco de carácter histórico y estadístico a efectos de investigación.

Estos artículos deben leerse en relación con el art. 15 bis. 1 LOGP y la STC 161/2021, en relación con el derecho al acceso al expediente penitenciario que solo puede limitarse de forma justificada.

5. CONTRASTE DEL MARCO NORMATIVO CON UN ANÁLISIS DE RESOLUCIONES JUDICIALES QUE MENCIONAN LA TABLA DE VALORACIÓN DE RIESGO Y LA HERRAMIENTA RISCANVI

Según García (2025), a quien ya hemos hecho alusión en una nota a pie del apartado introductorio, en su análisis de la base de datos del CENDOJ, referida a 2024, en cuanto a la alusión a la Tabla de Variables de Riesgo (TVR)[52], para la concesión de permisos, se recogen:

> al menos 250 resoluciones de 19 audiencias provinciales (…) que citan la TVR, 102 de ellas de la de Madrid. En su gran mayoría —206— se basan en el resultado del algoritmo para decidir sobre el permiso de salida de prisión del preso, en 37 ocasiones en contra de la decisión de la Junta de Tratamiento. Casi siempre —169— es para denegar el permiso porque el resultado va de "riesgo elevado" hasta "máximo", pero hay resoluciones que incluso se basan en la TVR, aunque haya dado como resultado riesgo "normal", para denegar el permiso en contra de la opinión de la Junta de Tratamiento, al considerar que es un nivel de riesgo "no asumible", rebajando aún más el umbral del riesgo aceptable" (pfo. 22º).

En nuestra observación de campo en el País Vasco no hemos advertido el uso plantillas automatizadas que utilicen directamente algoritmos, aunque sí parecen utilizarse, al menos en ciertas valoraciones de riesgo, en Instituciones Penitenciarias, como alude García (2025) y se muestra en la imagen siguiente.

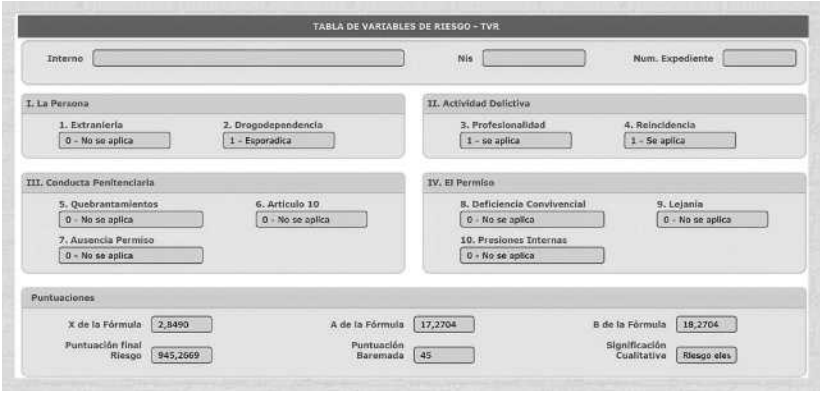


Imagen 2: *Tabla de Valoración de Riesgo*. Fuente: https://www.proyectoprisiones.es/riesgoquebrantamiento-permisos/

52. Completada con la Tabla de Concurrencia de Circunstancias Peculiares.

Otra cuestión es la mención expresa en las resoluciones de órganos jurisdiccionales penales catalanes de la herramienta RisCanvi (Alemán, 2023)[53]. En la búsqueda realizada el 6 de agosto de 2025, en la base CENDOJ, encontramos, desde noviembre de 2024 hasta la fecha indicada, 6.595 resultados[54], lo que muestra su plena integración en el sistema penitenciario catalán. Las resoluciones se refieren, fundamentalmente, a la concesión de permisos, pero también a libertades condicionales y a clasificaciones de grado. Proceden en su casi totalidad de la Audiencia Provincial de Barcelona, aunque se han encontrado algunas resoluciones de las Audiencias de Girona y Tarragona. La muestra inicial indicada, para valorar posibles evoluciones recientes jurisprudenciales, se completó con una alerta de búsqueda desde el 6 de agosto hasta el 31 de octubre de 2025. En esta última fecha, en la base de CENDOJ se contabilizaban 7.442 resoluciones que mencionan el término RisCanvi (algunas nuevas con referencia a 2016, 2017 y 2018 que se van incorporando de forma tardía a la base de datos referenciada). En la última búsqueda realizada, el 13 de noviembre de 2025, las 7.455 resoluciones encontradas, todas ellas, autos, se distribuían por años de la siguiente manera:

- 2025 (712)
- 2024 (1550)
- 2023 (1047)
- 2022 (821)
- 2021 (1321)
- 2020 (708)

53. Véase también su contribución en esta monografía.

54. 3.574 resultados con indicación en el texto de la resolución judicial de "riesgo alto". Del total de 6.595 resoluciones en menos de dos años que mencionan el término RisCanvi, según el buscador del CENDOJ, 1.777 (1.941 a 1 de noviembre) se refieren a resoluciones que desestiman el recurso interpuesto por el Ministerio Fiscal (búsqueda realizada en la misma fecha indicada por los descriptores: "RisCanvi Y "Desestima(r) el recurso de apelación interpuesto por el Ministerio Fiscal"". Aparecen 579 resultados que incluyen "Desestima(r) el recurso de apelación interpuesto por la defensa" (623 a 1 de noviembre). En 498 del total de resoluciones aparece "estimar el recurso de apelación interpuesto por la defensa" (en su caso contra el auto del JVP desestimando el recurso de alzada del interno contra la denegación de un permiso) (530 a 1 de noviembre de 2025). En todo caso, debe considerarse que la justificación de la estimación o desestimación de los recursos puede no centrarse, total o parcialmente, en la valoración de RisCanvi, por lo que la consideración de estas cifras solo tiene un valor meramente exploratorio, a la espera de investigaciones más pormenorizadas.

- 2019 (726)
- 2018 (399)
- 2017 (94)
- 2016 (32)
- 2015 (15)
- 2013 (1)
- 2012 (15)
- 2011 (13)
- 2010 (1)

Del análisis de contenido realizado, con todas sus limitaciones, no puede concluirse que haya una utilización restrictiva de RisCanvi en el sentido de menor concesión de permisos, si bien, sin perjuicio, como ya ha sido indicado, del coste que supone a los internos la tramitación y resolución de un recurso.

A modo de ejemplo de la alusión a RisCanvi en los autos judiciales analizados, siempre respecto de los razonamientos jurídicos, la resolución judicial de la AP de Barcelona 294/2025, de 20 de febrero de 2020, resolviendo negativamente el recurso interpuesto por el Ministerio Fiscal contra el auto del Juzgado de Vigilancia Penitenciaria concediendo el permiso de salida, con informe favorable de la Junta, indica:

> Por lo que aunque se detecte un riesgo alto en la variable de violencia intrainstitucional y medio en violencia autodirigida y reincidencia violenta, según la escala RisCanvi, consta en el informe que el interno ha trabajado en los programas su reacción o conducta violenta y desajustada, contando actualmente con elementos y capacidades para detectar las señales propias que precipitaban su conducta violenta, por lo que dichos riesgos podrían ser gestionados adecuadamente mediante los factores protectores concurrentes al haber adquirido el interno estrategias de gestión y control de impulsos, por lo que a pesar del pronóstico de la escala RisCanvi, se encuentra en condiciones de gestionar de forma adecuada ese riesgo y es merecedor de que se le otorgue la confianza a fin de que pueda hacer un buen uso del permiso, sirviéndole el permiso como estímulo para seguir con una adecuada trayectoria penitenciaria, preparándose para su vida en libertad, y, en definitiva, para la consecución del fin primordial de su reeducación y reinserción social (artículos 25.2 CE, 1 LGP y 2 RP).

> En consecuencia, consideramos que procede desestimar el recurso de apelación interpuesto por el Ministerio Fiscal y confirmar la decisión del juez de vigilancia penitenciaria al considerar ajustada a derecho la resolución recurrida.

También cabe mencionar, de forma ilustrativa, la resolución 305/2025 de esa misma Audiencia, con esa misma fecha, que desestima el recurso de apelación interpuesto por el Ministerio Fiscal contra el auto que autorizó el programa concreto de atención especializada propuesto por la Junta de Tratamiento del centro penitenciario, indicando dicha resolución lo siguiente:

> Si bien constan altos los riesgos de la escala RisCanvi en violencia intrainstitucional, reincidencia general y quebrantamiento de condena, este hecho está motivado por factores estáticos y dinámicos, pero, a la vista de los informes de los técnicos, podemos concluir que esos riesgos pueden ser gestionados adecuadamente al haber adquirido el interno estrategias para detectar y superar situaciones de riesgo y para afrontar situaciones complejas, siendo importante destacar que ha disfrutado ya de dos permisos de 24 horas y uno judicial de tres días sin incidencias y con valoración positiva en su tratamiento. Por lo que, a la vista de todo ello, consideramos al interno merecedor de que se le otorgue la confianza a fin de que pueda hacer un buen uso del permiso, por considerar que dicho permiso puede incidir positivamente en su tratamiento, sirviéndole como estímulo para seguir con una adecuada trayectoria penitenciaria y preparándose para su puesta en libertad, por lo que, concurriendo los presupuestos establecidos en los artículos 47.2 de la Ley Orgánica General Penitenciaria y 154 del Reglamento Penitenciario, resulta procedente aprobar la propuesta adoptada por la Junta de Tratamiento favorable a la concesión del permiso de salida, con las condiciones contenidas en la propuesta.

En otros casos, como en el auto de la AP de Barcelona 1183/2018, que desestimó el recurso de apelación de la defensa sobre la denegación de un permiso de salida, se indica en sus razonamientos jurídicos la posición garantista del tribunal:

> TERCERO. Por tanto, la concesión de un permiso implica realizar un pronóstico, una predicción para el futuro acerca de la existencia de un riesgo cierto de reincidencia, violenta o no, toda vez que es imposible saber con certeza *a priori* qué uso va a hacer el interno del permiso una vez fuera de prisión. Para realizar esta predicción, existe un instrumento de evaluación y gestión del riesgo, RisCanvi, cuyo objeto es mejorar las evaluaciones de pronóstico y por tanto gestionar eficazmente los permisos para conseguir la finalidad primordial de reinserción social. Para ello, los miembros de los equipos multidisciplinares de los centros penitenciarios (psicólogos, juristas, criminólogos, trabajadores sociales, educadores y pedagogos) obtienen información del interno (entrevistas con él y con las personas de su entorno,

> expedientes, observación directa, coordinación con otros profesionales...), y conforme a unos ítems se valoran todos factores de riesgo (entre estos los ya aludidos por el TC). Complementado con los informes individualizados de los diferentes profesionales que trabajan día a día con el interno y que por tanto están en las mejores condiciones para valorar su situación.
>
> CUARTO. Nada es infalible y más cuando se trata de comportamientos humanos. No obstante, a través del protocolo de evaluación y gestión RisCanvi, así como de la valoración efectuada individualmente y de forma individualizada por todos los profesionales que están en contacto diario con el interno en el centro penitenciario, así como con su entorno, valorando todos los factores concurrentes, se hacen predicciones rigurosas. De ahí que cuando el resultado, tanto del RisCanvi como de los miembros del equipo de tratamiento, sea de riesgo bajo, no existen motivos razonables para denegar el permiso, con independencia de la naturaleza del delito y gravedad de la pena, ya que se han valorado todos los factores de riesgo, incluyendo especialmente la clase y gravedad del delito, y se ha concluido que el riesgo de reiteración delictiva o de quebrantamiento de condena es bajo, y por tanto existe base fundada para considerar que el interno hará un buen uso del permiso, siempre acompañado de medidas de control externas. Como ya se ha dicho, es imposible asegurar este extremo ya que nos movemos en el campo de las predicciones de comportamiento a futuro. De ahí que los resultados del RisCanvi en la gestión del riesgo solamente sean, bajo, medio o alto. La única forma que habría para asegurar que el preso no será fuente de riesgo para la comunidad es mediante su inocuización, es decir, su retención en prisión el máximo tiempo posible. Cierto que así la seguridad es total. No obstante, también lo es que a costa de abandonar el fin de reinserción social y conseguir la desocialización del interno que, no olvidemos, en algún momento saldrá del centro penitenciario, y a la que desde luego no está orientada nuestra legislación penitenciaria que se encamina obviamente a cumplir lo dispuesto en el art. 25 CE.

En muchas ocasiones, la mención a RisCanvi es meramente descriptiva, en cuanto a su utilización en el sistema penitenciario, pero, en otras ocasiones, las audiencias determinan que RisCanvi, si bien debe tenerse en cuenta, no puede utilizarse de forma determinista, algo que es alegado tanto por los letrados de la persona interna, al recurrir denegaciones de sus solicitudes, como por la propia Junta que concede el permiso, indicando otros aspectos que deben tenerse en cuenta y que atemperan el riesgo. Ahora bien, cuando todos los indicadores son altos, resulta mucho más difícil contrarrestar ese pronóstico realizado por RisCanvi. Ejemplo de lo indicado anteriormente se encuentran en el auto de la AP de Barcelona 506/2025, en que se recoge lo siguiente:

> Es cierto que la escala RisCanvi arroja resultados de riesgos medios, pero se trata de una herramienta que no debe suplir el criterio humano.

En el auto de la AP de Girona 223/2025, en respuesta a un recurso de apelación del Ministerio Fiscal contra un permiso, recurso desestimado, se expresa lo siguiente:

> Esta Sala no está de acuerdo con esta pretensión del Ministerio Fiscal, máxime cuando en dicho informe se hacen constar las razones por las que reputa conveniente la concesión del permiso solicitado, pues como se ha señalado en otras resoluciones "el instrumento de valoración de evaluación de riesgo conocido como RisCanvi empleado para realizar valoraciones de riesgo a partir del tratamiento estadístico de toda una serie de variables personales del interno no puede utilizarse como una especie de oráculo infalible, pues son de sobra conocidos los riesgos de "falsos positivos" y "falsos negativos", todo ello sin perjuicio de su evidente utilidad como elemento indicativo o valorativo" (véanse AAP Girona Sección 3ª de 29 de noviembre de 2024 y de 15 de mayo de 2024).

También se desestima el recurso de apelación del Ministerio Fiscal contra el permiso concedido, en el auto de la AP de Barcelona 811/2025, donde se argumenta lo siguiente:

> máxime cuando, como se razona por la Magistrada de Vigilancia, el equipo técnico considera que el interno, a pesar del pronóstico de la escala RisCanvi, se encuentra en condiciones de gestionar de forma adecuada esos riesgos, por lo que es merecedor de que se le otorgue la confianza a fin de que pueda hacer un buen uso del permiso, y que dicho permiso puede incidir positivamente en el mismo, sirviéndole como estímulo para seguir con una adecuada trayectoria penitenciaría y preparándola para su puesta en libertad, por lo que, concurriendo los presupuestos establecidos en los artículos 47 de la Ley Orgánica General Penitenciaria y 154 del Reglamento Penitenciaria resulta procedente aprobar la propuesta, adoptada por la Junta de Tratamiento favorable a la concesión del permiso de salida, con las condiciones contenidas la propuesta. por lo que la autorización del permiso debe mantenerse (…) Se valora asimismo que pese que en la escala RisCanvi presente riesgo medio en reincidencia violenta y general y alto en quebrantamiento de condena, en parte son debidos a una recaída en consumo de tóxicos y las regresiones de grado, teniendo, por el contrario como factores protectores un entorno familiar adecuado, el mantenerse abstinente en tóxicos, cuenta con hábitos laborales y tiene voluntad de cambio, habiendo realizado una salida programada estando el permiso condicionado a continuar con la vinculación con el CAS. En estas condiciones se ha de mantener el criterio conforme el interno se hace merecedor de que se le otorgue la confianza a fin de que pueda hacer un buen uso del permiso, y que dicho permiso pueda incidir positivamente en el mismo, sirviéndole como estímulo para seguir con una adecuada trayectoria penitenciaría y preparándolo para su puesta en libertad, por lo que, concurriendo los presupuestos establecidos en los artículos 47.2 de la Ley Orgánica General Penitenciaria y 154 del Reglamento Penitenciario,

> resultando procedente aprobar la propuesta, adoptada por la Junta de Tratamiento favorable a la concesión del permiso de salida, con las condiciones contenidas en aquella.

En ocasiones también se alude en las resoluciones a otras herramientas específicas de valoración del riesgo en agresores sexuales SVR-20 (*vid.*, por ejemplo, los autos de la AP de Barcelona 2280/2024 y 926/2025). Asimismo, se cita la herramienta SARA (específica para delitos cometidos en el ámbito de la violencia de género) en el Auto AP de Barcelona 418/2025, donde se indica:

> presenta riesgos bajos en todos los parámetros de la escala RisCanvi así como en algunos de los criterios de la escala SVR-20 (*Sexual Violence Risk*) y SARA, siendo pues, que las circunstancias penitenciarias expuestas evidencian que los resultados moderados en alguno de los parámetros de estas dos últimas escalas específicas carecen de relevancia al objeto de impedir el disfrute del permiso reclamado, especialmente cuando el interno está en un estado avanzado del cumplimiento de una larga condena en tanto que ha cumplido más de 8 años en régimen cerrado. Así, por todo lo expuesto la Sala ha de estimar el recurso del interno a la denegación de permiso al no resultar la resolución recurrida suficientemente expresiva de un juicio racional lógico que determine la necesidad de denegar el permiso propuesto pues no puede sustentarse la denegación de un permiso en la simple afirmación de los resultados en cualquier escala de riesgo desconectados con circunstancias penitenciarias que los sustenten, ya sean riesgo moderados, medios o altos.

En otras resoluciones, de forma similar a la anterior, además de las circunstancias penitenciarias, parece considerarse también el paso del tiempo. Así, en el auto de la AP Barcelona 403/2025, se alude a ese factor, concluyendo que:

> (…) las óptimas consideraciones criminalísticas emitidas por los técnicos tienen fiel reflejo en el resultado de la evaluación de riesgos de la escala RisCanvi con un resultado bajo en todos los parámetros, de lo que se infiere ausencia de peligrosidad criminal.

También en el auto de la AP Barcelona 697/2025 se estima el recurso del interno contra la resolución que revoca el cumplimiento en régimen general:

> La defensa del interno alega que se han tenido en cuenta los riesgos de la escala RisCanvi sin que se haga mención a la fecha en que se obtuvieron los mismos, refiriendo que datan de diciembre de 2023 sin haberse valorado por tanto todos los avances tratamentales llevados a cabo durante el 2024, y añadiendo que es poco realista que esos indicios no se hayan movido pese a realizar los programas con optima adherencia, interesando que se pida el Ris-

> Canvi a fecha actual. Por otro lado, no comparte la conclusión de la juzgadora referida a que los factores de riesgo superan a los de protección, considerando que pueden compensarse unos con otros. Añade también que no puede dejarse de lado que el interno tiene hábitos laborales y apoyo familiar. Y por último hace mención a que en el plan de trabajo se indica la necesidad de continuar trabajando los factores de riesgo, pero a través de un grupo externo, entendiendo por ello procedente la aplicación del 78.2 con fines terapéuticos y de reinserción. En las alegaciones realizadas tras la desestimación del recurso de reforma añade la falta de motivación de la resolución recurrida, reiterando la necesidad de reclamar el RisCanvi actualizado. Respecto a esta última manifestación, señalar que la juez de vigilancia recabó del centro penitenciario antes de resolver la reforma el RisCanvi actualizado del interno recibiéndose fechado en 13 de febrero de 2025 y constando los mismos parámetros que ya obraban en la causa excepto la reincidencia general que pasa de riesgo medio a bajo. Constan asimismo dos manuscritos del interno de 12 y 18 de febrero de 2025, unidos también a la causa con carácter previo a la resolución de la reforma. El interno está condenado a pena de 56 años y 4 meses de prisión por cuatro delitos de agresión sexual, un delito de robo con violencia e intimidación y un delito de robo con violencia e intimidación en casa habitada y con instrumento peligroso, todos ellos en virtud de una única ejecutoria.

En sentido contrario, el Auto AP 235/2025 indica que, en la escala RisCanvi, todos los riesgos son altos, por lo que:

> si bien es cierto, y así lo tiene reiterado la Sala, que los niveles de riesgo en cualquier escala no determinan el riesgo real, sí son relevantes cuando se corresponden, coinciden o apuntan en la misma dirección que las circunstancias penitenciarias que resultan de los informes de los técnicos de instituciones penitenciarias.

Sin embargo, también estima el recurso del interno la AP Barcelona 192/2025:

> Dejando a un lado los riesgos del RisCanvi a los que después volveremos, los otros extremos son fijos e inmodificables en la vida del interno, y pretender, en base a ellos, denegar el permiso comportaría no poder concedérselo nunca. La vida delincuencial que ha llevado, su problemática toxicológica y la reincidencia, así como el regresar de un permiso un día tarde, han sido abordados tratamentalmente en el centro penitenciario con resultado favorable. De hecho, no puede obviarse que retomó la cadena permisiva en 2023, disfrutando de 2 permisos de 48 horas, uno de 3 días y tres de 6 días, todos ellos sin incidencias, siendo el motivo de la paralización de esas salidas que al cuarto permiso de 6 días regresó un día tarde. Ello evidencia que el desarrollo de los permisos, a nivel de reincidencia, de consumo o de consecución de objetivos previstos, tuvo lugar sin incidencia alguna, siendo el único punto a cuestionar que regresó un día tarde en uno de los permisos

> de 6 días. Este retraso ha sido trabajado con el interno, y la Junta considera que debe continuar con la cadena que inició en su día, sin que los riesgos del RisCanvi sean impedimento para ello. Así, no solo se ha constatado que esos riesgos se han minimizado por la forma en que se han desarrollado las salidas, sino que, además, y teniendo en cuenta su cadena delictiva y penitenciaria, algunos de los ítems no van a variar nunca. No puede tampoco compartirse la conclusión de la juzgadora referida a la lejanía del licenciamiento, ya que, si bien es cierto que está previsto para 2028, no puede obviarse que se inició en 1997, habiendo superado con creces las tres cuartas partes de la condena. Por lo dicho, siendo la evolución buena, la situación estable en este momento, tanto económica como familiar, y habiéndose iniciado la cadena permisiva sin más incidente que el retraso en un reingreso que ya ha sido abordado y superado tratamentalmente, no hay motivo que justifique la no concesión del permiso, debiendo, en consecuencia, estimarse el recurso. Vistos los preceptos legales citados y los demás de aplicación, la Sala resuelve estimar el recurso de apelación interpuesto por la defensa.

En este sentido, puede verse una referencia a la evolución de la jurisprudencia para reconsiderar los factores estáticos[55], así en el auto de la AP Barcelona 340/2025 se recoge lo siguiente:

> los riesgos bajos se corresponden con las circunstancias penitenciarias del interno acreditadas en autos, tales como la realización y superación del PIT (Plan de Intervención Tratamental), su buena conducta, primariedad delictiva, su observación... siendo que el riesgo alto se evidencia que obedece a factores estáticos, y ahí está el error de la motivación del órgano *a quo* en tanto que no solo se evidencia por los hechos, circunstancias y naturaleza del delito (extranjero no residente legal que trata de introducir droga en España y es detenido en el aeropuerto, careciendo de todo vínculo en España al encontrarse en tránsito) sino porque expresamente se recoge así en los informes cuando, como hemos recogido en la presente resolución, se expresa que el riesgo en "Trencament de condemna: ALT [...] es relaciona amb l'absència de vinculacio i arrelament al país ja que va ser interceptat a l'aeroport del Prat". Por otro lado, si bien es cierto que el informe del psicólogo se refiere a la realización de un permiso de 24 horas para valorar la gestión del riesgo de quebrantamiento no es menos cierto que el objeto de tal permiso no es otro que posibilitar que realice "gestions relacionades amb l'enviament dels diners que ha estat estalviant al Centre per poder ajudar económicament a la familia que es troba al país d'origen", Perú, por lo que si además tenemos en

55. Es cierto, que la valoración, conforme al RisCanvi, se califica como "media" en algunas de las áreas objeto de análisis (reincidencias y quebrantamiento de condena), si bien en su mayoría se corresponden en gran medida con elementos estáticos derivados de su carrera delictiva, pero entienden los técnicos que estos riesgos pueden ser gestionados adecuadamente mediante las herramientas que ha adquirido con sus programas de tratamiento y se encuentran por ello compensados.

> cuenta que dispone de recursos procedentes de su trabajo en el centro y que el permiso se disfrutará en un recurso residencial a lo que se suma el sentido de la propuesta, favorable, la Sala evidencia que las circunstancias expuestas no exigen ni justifican la exigencia de un requisito no previsto legalmente como es el previo disfrute de salidas o permisos de otra naturaleza antes de disfrutar del permiso judicial propuesto y reclamado. Así, la Sala estima desproporcionadas y no justificadas las razones expuestas por el órgano *a quo* y el Ministerio fiscal a fin de oponerse a la autorización del permiso reclamado por el interno y propuesto por la Junta de Tratamiento, por lo que ha de estimarse el recurso, como se apuntó, y autorizar el permiso. Vistos los preceptos legales citados y los demás de aplicación. La Sala resuelve estimar el recurso de apelación interpuesto por la letrada (…) y autorizando el permiso propuesto y reclamado en los términos fijados por la Junta de Tratamiento.

También desestima el recurso de apelación de Fiscalía, de concesión de permiso, el auto de la AP de Barcelona 703/2025:

> si bien la escala RisCanvi arroja un riesgo medio de reincidencia violenta el interno ha desarrollado una adecuada capacidad de autocrítica y concienciación de sus factores de riesgo y protección que le permiten gestionar el riesgo observado. Y, en todo caso, la citada escala no puede desvincularse del criterio de los profesionales que trabajan con el interno. Ha completado todos los programas psicoeducativos "Para i Pensa" y "Preparant la vida en Comunitat" y el Programa de Intervención en drogodependencias, así como otros itinerarios de violencia (DEVI), todos ellos con valoración favorable. Ha asumido plenamente la responsabilidad de sus actos encontrándose en una senda rehabilitadora y reparadora, apartada de cualquier conducta violenta, sin que existan factores de riesgo que justifiquen la denegación del permiso. Además, ha realizado múltiples salidas programas para acudir al centro hospitalario donde recibe tratamiento y ha disfrutado de tres permisos gubernativos, sin incidencias. Cuenta con domicilio estable en el que convive con su madre y su pareja que le dan soporte emocional y es padre de tres hijos menores. Concurren, pues, todos los requisitos necesarios para la concesión del permiso.

También se valoran cambios y se estima recurso de apelación del interno, en el auto de la AP de Barcelona 612/2025 (en este sentido favorable al interno, véase también el auto 504/2025):

> si bien el riesgo en la escala RISCANVI de reincidencia violenta era entonces alto, este posteriormente fue rebajado a riesgo medio, lo que confirma la capacidad de gestión positiva de este riesgo por parte del interno de acuerdo con los aprendizajes adquiridos en el curso del tratamiento.

Otro ejemplo favorable a la persona interna en prisión se encuentra en el auto de la AP de Barcelona 937/2025, en relación con un interno con

trastorno de pedofilia. Por resultar de especial interés, reproducimos un extracto extenso a continuación:

> Hemos puesto de manifiesto en anteriores resoluciones de la Sala que los resultados medios en el RisCanvi medios (en este caso son altos) no constituyen necesariamente un impedimento para disfrutar de permisos y que las valoraciones medias en algunos de los riesgos pueden ser compensadas con la realización de permisos gubernativos de corta duración o con la gestión mediante otras herramientas. (…) (aludiendo a resoluciones anteriores en este caso, se alegaba entonces que) "en este caso ni se ha producido una profundización en los programas realizados, sino al contrario, y no se ha realizado este refuerzo que los permisos gubernativos comportan, lo que conduce a la necesidad de mostrarse prudentes y estar a la espera de que se inicie el acercamiento al exterior de forma progresiva a través de salidas programadas y permisos gubernativos de horas. Ello permitiría que se lleve a cabo un período de consolidación de los objetivos aprendidos a través del programa de tratamiento y la constatación de dichos aprendizajes, mediante una supervisión paulatina. En el presente caso, la propuesta favorable de la Junta de Tratamiento se refiere a la sesión que tuvo lugar el 27 de noviembre de 2024, habiendo transcurrido cuatro meses desde la anterior sesión de fecha 24 de julio de 2024. A la vista de los informes técnicos aportados actualizados, las circunstancias del interno son prácticamente idénticas, habiendo disfrutado únicamente de dos salidas programadas más, sin haber iniciado los permisos gubernativos de horas. Por ello, no se observan razones objetivas y relevantes que permitan un cambio del criterio que expusimos en nuestra resolución anterior." Actualmente el interno ha avanzado en el cumplimiento de la condena, ha disfrutado de 14 salidas programadas —nueve de intervención terapéutica y cinco lúdicas— y además ha disfrutado de un permiso gubernativo de doce horas el 23 de enero de 2025, ya que, si bien se solicitó uno de 48 horas, la DG lo redujo a 12 horas sin pernocta. Según la valoración psicológica a pesar de haberse trabajado con el interno su desviación sexual —pedofilia— y haber aceptado la derivación a un recurso específico, sigue presentado algunos mecanismos de defensa en relación con el primer abuso sexual producido cuando tenía 14 años y en relación a su hermana de 9 años y si bien ha mejorado en la empatía y ha sido capaz de identificar a las víctimas principales y secundarias, conecta con el mal causado, pero poniendo más peso en su persona, por lo que se ha de continuar trabajando, al igual que se ha de seguir trabajando paraque el reconocimiento de hechos sea pleno y sin ningún tipo de justificación, lo que unido a que el riesgo de reincidencia delictiva sexual según el SVR20 es medio y a la necesidad de ser derivado a un recurso específico para hacer terapia y gestionar los riesgos de recaída, estableciéndose en su propio plan de trabajo que ha de hacer una salida muy progresiva hacia el exterior con una derivación a un recurso específico para tratar su desviación sexual, que está actualmente realizando el programa de preparación de la vida en comunidad y que no lo ha terminado, así como que no ha disfrutado de ningún permiso gubernativo con

> pernocta, es por lo que procede confirmar la resolución recurrida. Por tanto, y sin desconocer las circunstancias positivas puestas de manifiesto en los informes, es cierto que el acercamiento progresivo al exterior constituye una mayor garantía de éxito y de hecho forma parte del programa de violencia sexual, atendiendo al delito cometido, a que es reincidente y a la necesidad de comprobar la consolidación de la intervención tratamental, de ahí que todos los profesionales del centro penitenciario coincidan en la necesidad de la aproximación gradual muy progresiva y a que es evidente que existe en este tipo de delitos un riesgo derivado del trastorno de pedofilia, que aconsejan ese acercamiento progresivo para minimizar el riesgo de reincidencia, es por lo que se considera que se ha de esperar a la consolidación de los objetivos de su PIT, a la terminación del programa de preparación para la vida en libertad y al disfrute de permisos gubernativos por horas y pernocta de menor duración a fin de comprobar cómo el interno se relaciona con el medio externo y si dispone de mecanismos o recursos para gestionar los riesgos existentes, lo que permitiría, más adelante, iniciar con regularidad la cadena permisiva judicial.". En esta última resolución ya constatamos la evolución favorable del interno a nivel tratamental y de evolución personal, pero consideramos que teniendo en cuenta que no había finalizado el último programa encaminado a preparar la vida en comunidad y que no había disfrutado de permisos con pernocta, conceder un permiso de tres días era prematuro. En este momento las circunstancias han variado, ha finalizado los programas con buen resultado, y ha iniciado la cadena permisiva con varias salidas programadas, un permiso de 12 horas y dos permisos de 48 horas antes de la Junta y un tercero después de la misma, todos sin incidencias. Se ha vinculado al recurso realizando seguimiento en el mismo. La necesidad de disfrutar de más permisos de horas que alega tanto el Ministerio Fiscal como la Juez de Vigilancia no vienen fundamentados, por cuanto no solo está avalado el permiso judicial por el informe favorable de la Junta, sino que además un total de tres permisos de 48 horas son más que suficientes para ampliar el disfrute a un día más. Por tanto, y siendo congruentes con nuestra última decisión, debemos estimar el recurso y aprobar el permiso (…) en la forma que ha dispuesto la Junta de Tratamiento.

Ahora bien, en el Auto de la AP de Barcelona 633/2025, en relación con posibles riesgos en delitos sexuales, con una indicación de RisCanvi de riesgo moderado a alto, se observan las dificultades que terminan con la desestimación del recurso del interno, aunque la defensa alegue que la valoración de las herramientas algorítmicas "no debe ser automática y debe ser por el contrario integral". También, en un recurso resuelto en contra de un interno, con trastorno de pedofilia, puede citarse el Auto AP de Barcelona 868/2025:

> (…) presenta un riesgo medio en quebrantamiento de condena según el protocolo RisCanvi y moderado en el protocolo SRV-20 (*Sexual Violence Risk 20*) que es el indicador de riesgo específico para la tipología delictiva que nos ocupa,

> es por lo que se considera necesario que realice un acercamiento progresivo al exterior ya que solo ha disfrutado solo de un permiso de 48 horas, del 12 al 14 de septiembre de 2024. Por tanto, y sin desconocer las circunstancias positivas puestas de manifiesto en los informes, es cierto que el acercamiento progresivo al exterior constituye una mayor garantía de éxito y de hecho forma parte del programa VISEX, atendiendo al delito cometido y a la necesidad de comprobar la consolidación de la intervención tratamental, de ahí que se considere necesario la aproximación gradual dado que es evidente que existe en este tipo de delitos el riesgo derivado del trastorno de pedofilia, que aconsejan ese acercamiento progresivo para minimizar el riesgo de reincidencia, lo que permitirá mientras tanto que se consoliden los objetivos de su PIT a fin de comprobar cómo el interno se relaciona con el medio externo y si dispone de mecanismos o recursos para gestionar los riesgos existentes, lo que permitiría, más adelante, iniciar con regularidad la cadena permisiva. Por ello, y entendiendo que no existen razones para discrepar del parecer de la Magistrada de instancia, procede la desestimación del recurso de apelación, con la consiguiente confirmación del auto denegando el permiso de salida propuesto.

También en el Auto de la AP de Barcelona 1153/2025, aunque el letrado del interno alegue que la valoración de RisCanvi debe matizarse, el tribunal considera lo siguiente:

> (… ya) que su defendido ha cumplido la mitad de la condena, observa buena conducta actualmente, es respetuoso con la institución y los profesionales, ha participado correctamente y ha finalizado los programas de tratamiento, ha disfrutado de cinco salidas programadas, cuenta con apoyo familiar y núcleo acogedor y las puntuaciones de la valoración RisCanvi han sido matizadas en el informe psicológico, sin que la existencia de una sanción pendiente de cancelar constituya un motivo para la denegación de un permiso. En base a ello, interesa la estimación del recurso de apelación, concediendo al interno el disfrute del permiso de salida propuesto por la Junta de Tratamiento. Evacuado el oportuno traslado, el Ministerio Fiscal interesó la desestimación del recurso de apelación y la íntegra confirmación de la resolución impugnada, al entender que, en el momento actual, no concurren garantías suficientes que garanticen que el interno haga un uso correcto del permiso como preparación de la vida en libertad, alegando que tiene prevista la libertad definitiva para el día 15 de agosto de 2039, presenta un comportamiento irregular en prisión, un expediente disciplinario sin cancelar, una insuficiente evolución en su programa de tratamiento, y riesgo alto en todos los parámetros del RisCanvi, por lo que considera que es necesario un acercamiento al exterior de forma muy progresiva con salidas programadas y permisos gubernativos.

Resulta de especial interés, en relación con el marco normativo desarrollado en los epígrafes anteriores de este capítulo, el auto de la AP de Barcelona 1131/2025 que, de nuevo, estima el recurso del interno:

> Aunque se detecte un riesgo medio en la variable de reincidencia violenta de la RisCanvi, sin embargo, del conjunto de la información se desprende que dicho riesgo podría ser gestionado adecuadamente mediante los factores protectores concurrentes, al haber adquirido el interno estrategias de gestión y expectativas de recuperar su estabilidad. No debe olvidarse que el riesgo es un concepto extremadamente indeterminado y subjetivo. Determinar perfiles de riesgo en aras de diseñar mayores cuotas de rigor en atención a este tipo de perfiles, en ocasiones menoscaba el principio de proporcionalidad, porque el infractor es perjudicado por una presunción de riesgo proyectada hacia el futuro y el RisCanvi se estructura en dos partes en cuanto a la aplicación de su contenido, a saber: una que hace referencia a la evaluación y valoración del riesgo y la otra, a la gestión del riesgo evaluado. En esta línea y a la vista de todo lo expuesto, el equipo técnico considera que el interno, a pesar del pronóstico de la RisCanvi, se encuentra en condiciones de gestionar de forma adecuada ese riesgo, por lo que es merecedor de que se le otorgue la confianza a fin de que pueda hacer un buen uso del permiso. Abundando en lo expuesto, es importante destacar que los permisos son funcionales para la preparación de la definitiva libertad, por tanto, es aconsejable que antes de alcanzar esa libertad definitiva comience a disfrutar, de manera progresiva, de permisos judiciales. Por tanto, y sobre la base de que al parecer de los técnicos la evolución del interno le hace merecedor del permiso propuesto, atendiendo a la evolución del tratamiento penitenciario, es por lo que procede desestimar el recurso de apelación interpuesto, evidenciándose que la fundamentación de la Junta de Tratamiento y de la resolución recurrida es expresiva de un juicio del todo punto racional y lógico que determina la necesidad de autorizar el permiso reclamado sin que merezca la censura interesada por el Ministerio Fiscal.

Otro ejemplo en que se desestima el recurso de la Fiscalía es el auto 1650/2018 de la AP Barcelona, donde se alude a la normativa vigente y a las limitaciones de RisCanvi, donde se indica lo siguiente:

> El art.25.2 CE dispone que "las penas privativas de libertad y las medidas de seguridad estarán orientadas hacia la reeducación y la reinserción social", con lo que, con independencia de que puedan desempeñar otros cometidos, se destaca que la función de reinserción constituye la meta primordial de las penas privativas de libertad. Y ello resulta obligado, toda vez que el hecho de cometer un delito no priva a su autor de su condición de ciudadano, por lo que compete al Estado, además de su custodia y retención, la adopción de todas las medidas necesarias para asegurar su plena incorporación a la comunidad. Y esto no se consigue mediante la inocuización, aun temporal, del penado, que solo provoca su desocialización. La resocialización exige un tratamiento durante la estancia en prisión individualizado en virtud del cual se persigue que el preso modifique su comportamiento y lo adecue para la vida en común sin apartarse de la norma; en definitiva, se procura su recuperación social. Por ello, forma parte del tratamiento orientado a la

resocialización la potenciación de los contactos del interno con el mundo exterior. No se pueden considerar los permisos de salida como un derecho o una recompensa, sino como un instrumento resocializador. Ya lo dijo el TC en su importante Sentencia 112/96 (Ponente: Vives Antón): "Todos los permisos cooperan potencialmente a la preparación de la vida en libertad del interno, pueden fortalecer los vínculos familiares, reducen las tensiones propias del internamiento y las consecuencias de la vida continuada en prisión que siempre conlleva el subsiguiente alejamiento de la realidad diaria. Constituyen un estímulo a la buena conducta, a la creación de un sentido de responsabilidad del interno, y con ello al desarrollo de su personalidad. Le proporcionan información sobre el medio social en el que ha de integrarse e indican cuál es la evolución del penado". Y ello, con independencia de la naturaleza del delito que haya cometido, pues ni la Constitución, ni la jurisprudencia del Tribunal Constitucional, ni las Leyes, introducen distinciones por tal razón. Lógicamente, la concesión de un permiso requiere la concurrencia de determinados requisitos, orientados en cierto modo a proteger a la sociedad, como es que no exista riesgo de reiteración delictiva o de quebrantamiento de condena. (…) En el caso que nos ocupa, la Junta de Tratamiento, en su sesión de fecha 21 de junio de 2018 propuso por unanimidad la concesión de un permiso de 3 días para el interno Sr. Miguel. De su expediente resulta que el Sr. Miguel, de 25 años de edad, cumple una condena de 10 años y 27 meses de prisión por 4 delitos de robo con violencia e intimidación (todos por el procedimiento del tirón), cometidos entre los años 2011 y 2012. Su libertad definitiva la tiene prevista para el año 2024. Ingresa en el año 2012. El recurso debe ser desestimado. La etiología delictiva viene motivada por su precariedad económica y la falta de recursos y es cierto que desde los 18 años está institucionalizado. Sin embargo, a juicio de los profesionales que trabajan diariamente con el interno, su evolución es muy buena, está motivado al cambio y entusiasmado por el hecho de poder rehacer su vida en libertad. Ha realizado los programas de tratamiento —educación motivacional y de *conducta delicti*va— con resultados satisfactorios, implicándose de forma activa y asumiendo los contenidos. Ha trabajado en los talleres productivos adquiriendo hábitos laborales. El permiso lo disfrutará en un recurso residencial, aunque sí tiene un referente familiar, su prima Fidela, que, aunque no puede acogerlo en su casa, sí le apoya emocionalmente y tiene capacidad de contención con un discurso prosocial. Es cierto que hay un riesgo medio en reiteración delictiva según RisCanvi, pero los profesionales que trabajan diariamente con el interno consideran que Miguel, internamente, está preparado para iniciar el contacto con el exterior dado que no tiene problemática toxicológica, la pena ha cumplido su efecto intimidatorio, tiene hábitos laborales. Y además consideramos que es un interno joven, 25 años, que ha reflexionado sobre su estilo de vida, estando motivado para cambiar y además habiendo hecho esfuerzos para ello, como asistir a la escuela en el centro, trabajar desde que ingresó en los talleres productivos. Su evolución es muy positiva, la pena ha producido efecto, quiere cambiar de vida y está

> ilusionado con iniciar los contactos con el exterior, siendo ahora el momento óptimo para reforzar su motivación y lo adquirido con su PIT mediante los permisos de salida. En virtud de lo expuesto procede la desestimación del recurso interpuesto.

En definitiva, de este estudio jurisprudencial, centrado principalmente en los últimos doce meses de autos disponibles en la base del CENDOJ, parece apreciarse, incluso a lo largo del tiempo, con mayor conocimiento en los últimos años de la herramienta y sus actualizaciones, un uso ponderado de la misma por parte de los tribunales, sin perjuicio de que puedan existir resoluciones dispares en cuanto a la interpretación de factores internos y externos o estáticos y dinámicos, según un empleo diverso de la terminología en cada resolución y recursos y alegaciones que la motivaron.

Así, en resoluciones más antiguas puede observarse una cierta falta de información o imprecisiones a la hora de describir el funcionamiento o la terminología de RisCanvi. Así, en el Auto de la AP de Barcelona 1776/2017, en recurso de apelación del interno contra la resolución del JVP, se expresa lo siguiente:

> En cuanto a los factores críticos o de riesgo externo, en concreto su núcleo familiar, dicho riesgo, así como su abstinencia al alcohol, serán controlados y supervisados a través del recurso residencial, por lo que se minimizan los riesgos de reincidencia y de quebrantamiento de condena que resultan del RisCanvi y que, no obstante, no coinciden con la valoración efectuada por los profesionales que trabajan diariamente con el interno, quienes consideran que los riesgos están controlados y por eso proponen el permiso al considerarlo necesario para su proceso de reinserción social, debiendo primar esta valoración sobre la efectuada por el RisCanvi, que tiene en cuenta factores estáticos y no tanto dinámicos como la evolución tratamental del interno. En virtud de lo expuesto, procede la estimación del recurso interpuesto.

Conviene, finalmente, preguntarse por cómo hacer compatible la evaluación de los riesgos en violencia de género con las notificaciones a la víctima previstas en la ley para que se reduzca la victimización secundaria[56]. En la actualidad no parece existir la preocupación por una comunicación o explicación adecuadas a la víctima que considere dicha evaluación, incluyendo la del RisCanvi y la clínica, así en el Auto de la AP de Barcelona 1749/2018:

> En este sentido, además de no consumir, ha iniciado el pago de la responsabilidad civil con su trabajo en los talleres. Los resultados de valoración del ries-

56. En ocasiones se estima el recurso de la persona interna, aunque no haya pago de la responsabilidad civil, aunque, junto con otros aspectos, se valora una actitud empática hacia la víctima. Vid. el Auto de la AP de Barcelona 25/2025.

> go según RisCanvi son bajos en todas las variables salvo en quebrantamiento de condena que es alto. Sin embargo, este riesgo aparece minimizado ya que tiene medio externo acogedor, su madre Cándida, que le apoya y le ofrece una estabilidad. Por otra parte, el resultado de riesgo alto es consecuencia de haberse retrasado en reingresar al centro en su último permiso en octubre de 2016. Pero los profesionales que trabajan diariamente con el interno consideran que actualmente su conducta es adaptada y de hecho en la revisión de su PIT se propuso la previsión de permisos en el segundo trimestre del año 2018. Teniendo en cuenta por tanto la concurrencia de factores favorables de cambio del interno y que el permiso va acompañado de una serie de condiciones para garantizar su disfrute en el exterior, procede conceder el permiso propuesto al ser complemento necesario de su proceso de reinserción (…) Conforme dispone el artículo 7.1-e) de la Ley 4/2015, de 27 de abril, del Estatuto de la Víctima del delito, deberán notificarse a toda víctima que así lo hayan solicitado las resoluciones o decisiones de cualquier autoridad judicial o penitenciaria que afecten a sujetos condenados por delitos cometidos con violencia o intimidación y que supongan un riesgo para la seguridad de la víctima. Delitos de los comprendidos en el elenco enumerado en el artículo 13.1.a) de dicho texto legal, homicidio, aborto del artículo 144 CP, lesiones, contra la libertad, tortura y contra la integridad moral, contra la libertad e indemnidad sexual, robo cometido con violencia e intimidación, terrorismo y trata de seres humanos, o análogos. En estos casos y a estos efectos, la Administración penitenciaria comunicará inmediatamente a la autoridad judicial la resolución adoptada para su notificación a la víctima afectada. Dado que el caso analizado está comprendido en los preceptos mencionados, procede comunicar esta resolución al Juzgado de Vigilancia Penitenciaria que corresponda a los efectos previstos legalmente.

6. CONTRASTE A TRAVÉS DE UN TRABAJO DE CAMPO REALIZADO EN EUSKADI: OBSERVACIÓN NO PARTICIPANTE DE LA TOMA DE DECISIONES TÉCNICAS EN LAS TRES PRISIONES VASCAS

Tras las solicitudes y autorizaciones pertinentes y bajo el principio de confidencialidad, anonimato y consentimiento informado, respecto de los datos personales, gracias a la colaboración con el Gobierno Vasco, entre julio y octubre de 2025, la autora de este capítulo realizó aproximadamente unas veinte horas de observación de las reuniones de los equipos técnicos. En los equipos se reunieron trabajadores sociales, educadores sociales, psicólogos, personal administrativo y juristas.

En cada reunión observada debatieron entre seis a dieciséis personas respecto de decisiones, preparatorias para la celebración de las Juntas, sobre

ingresos, clasificaciones, progresiones, salidas a sección abierta y fin de semana, traslados o cambios de destino, analíticas, evaluación y seguimiento de acompañamiento en celda, seguimiento de autolesiones, evaluación y seguimiento de planes específicos, instancias de internos (solicitando aspectos relacionados con vacaciones o participación en círculos restaurativos, por ejemplo), cambios de horarios, permisos ordinarios y extraordinarios, libertades condicionales (iniciación de expediente, aprobación del plan individual, informe de pronóstico inicial, elevación de expediente a libertad condicional, revisión del plan individual de intervención, aceptación de tutela en otro centro de seguimiento), redenciones, actividades o valoración del tercer grado a efectos de expulsión (art. 89 CP)[57]. Estos equipos técnicos preparan el trabajo de la Junta para el estudio de propuestas, incluyendo modelos de intervención y, en su caso, un plan individualizado de intervención de internos de reciente ingreso.

Estos equipos trabajan con documentos con tablas de varias páginas que señalan, por ejemplo, grupos de factores de reinserción o inadaptación, como los siguientes[58]:

a) factores delictivos (primariedad y antigüedad de hechos delictivos; escasa relevancia penal y social; escasa peligrosidad del delito; inicio tardío en la vida del acto delictivo; no existencia de carrera delictiva; delito ocasional…);

b) factores penales (largo período de libertad provisional, sin que consten nuevos delitos; eximentes incompletas o atenuantes muy cualificadas; solicitud de indulto; ingreso voluntario en prisión; cumplimiento de nuevo delito anterior a la libertad provisional…).

c) disminución de la peligrosidad por edad o por avance de enfermedad;

d) referidos a la condena (escasa cuantía; cumplimiento de la mitad de la condena; cercanía de cumplimiento de dos tercios o tres cuartas partes; próximo licenciamiento definitivo; compromiso de cumplimiento de la libertad condicional en su país…);

e) penitenciarios, etcétera, con relevancia del pago o intención de pago, total o parcial, de la responsabilidad civil.

57. Se indicó que Marruecos, por ejemplo, es más favorable a la recepción de personas expulsadas, no así Argelia.

58. Se ha realizado una reordenación propia respecto del análisis documental de la tabla.

En dos centros, se proyectó un Excel[59] durante la reunión, donde se fueron consignando, en un ordenador, las decisiones tomadas, si bien ese era el único ordenador en la sala ya que la mayoría de los expedientes se trajeron en papel por parte de cada persona técnica.

Finalmente, cabe indicar que en reuniones largas de más de dos horas puede apreciarse el cansancio ante la toma de decisiones y la carga de trabajo restante.

6.1. METODOLOGÍA DEL TRABAJO DE CAMPO DE OBSERVACIÓN NO PARTICIPANTE

Sobre el trabajo de observación[60], este se realizó en concreto los días[61] 15 y 16 de julio de 2025 en el centro penitenciario de Martutene (Gipuzkoa)[62]; el 12 de agosto en el centro penitenciario de Zaballa (Araba), viendo varios módulos, uno de ellos de respeto, con más posibles progresiones en grado a través de los artículos 100.2 ó 117 RP, lo que suele implicar más debate, si bien en el mes de agosto, en principio, solo se estudió la clasificación; y el 7 de octubre en el centro penitenciario de Basauri (Bizkaia), con decisiones relativas a personas en régimen abierto.

---

59. Empezaron a utilizarlo durante la pandemia ante la necesidad de contacto online.
60. Quiero expresar mi más sincero agradecimiento al Gobierno Vasco y, en concreto, a José María Bastos, David Garrido, Carmelo Anta Félez, Juan Carlos Calzada y Jesús Guerrero Guecho, así como a todas los profesionales que tuvieron la amabilidad y confianza de acogerme en cada reunión de los equipos técnicos observados. Cabe destacar el carácter extraordinario de esta observación. Con diferente objetivo y ámbito territorial, solo hemos podido encontrar un trabajo similar de observación, respecto de las Juntas catalanas, en el trabajo publicado por Larrauri (2019), donde se indica que, en un 30,43% de los casos analizados, la Fiscalía informaba negativamente respecto de la concesión del permiso porque la clasificación de riesgo era alta según RisCanvi; y en el trabajo de Rovira, Larrauri y Alarcón (2018), donde solo se encuentra una mención a RisCanvi, sin relevancia para el objetivo de este capítulo. Con una metodología de entrevistas a profesionales de las Juntas, que incluyen cuatro menciones a RisCanvi, mayormente desde una perspectiva crítica, se encuentra también el trabajo de Pedrosa (2025).
61. En estos meses se produjeron bajas y permisos vacacionales. Si se considera que, en algunos centros, las personas que votan deben hacer los informes, lo que exige conocer personalmente a la persona interna y su problemática, así como hacerse responsable de lo que se afirma por escrito, algunas decisiones tuvieron que aplazarse a decisiones de una reunión posterior.
62. El primer día con el equipo de medio abierto y el segundo con el equipo de régimen ordinario.

El análisis temático realizado consiste en un método cualitativo para identificar y analizar temas comunes dentro de los datos obtenidos (Braun y Clarke, 2022). En este caso, estos datos proceden de un trabajo etnográfico, mediante la observación externa no participante de los equipos técnicos en las tres prisiones vascas con el objetivo de completar y matizar el análisis realizado de la revisión bibliográfica especializada y el estudio jurisprudencial.

Como ya se ha indicado, dicho trabajo de campo, con una duración total de 20 horas aproximadamente, se realizó con equipos técnicos, con evaluación de módulos diferentes, en los tres centros penitenciarios de Euskadi[63] entre julio y octubre de 2025. Con participación de seis a ocho personas en dos reuniones (llegando a diez y dieciséis en otras dos), con mayoría de mujeres y con un carácter interdisciplinar e intergeneracional.

El estudio temático llevado a cabo es reflexivo ya que, aunque se consideran los objetivos generales y específicos de la presente investigación, no puede prescindirse de la subjetividad de la investigadora en la codificación, el análisis de los datos y la generación de temas[64].

Los temas identificados permiten agrupar significados encontrados en el conjunto de datos basados en la observación de la toma de decisiones en los equipos y en el análisis de la documentación proporcionada relacionada, como las tablas mencionadas. En este trabajo, para identificar los temas, se ha optado por un enfoque mixto o híbrido, es decir, simultáneamente ascendente (o inductivo), que parte de las notas escritas de la observadora, y deductivo (o descendente o teórico) que parte de la bibliografía existente analizada.

6.2. CONSIDERACIONES ÉTICAS

A lo largo de los primeros meses de 2025, se realizaron las gestiones pertinentes de autorización dentro de la Dirección de Justicia del Gobierno

63. Muy diferentes en número de internos y número de funcionarios por interno. Con listados de personas muy extensos, en todos los casos. En ocasiones alguna de las personas que debe estar en la reunión del equipo no puede acudir dada la carga de trabajo, por ejemplo, en relación con atención a la salud mental de las personas internas, sin perjuicio de participar después en la Junta.

64. Como indican Zempi et al. (2025), las seis fases del proceso analítico temático reflexivo son las siguientes: familiarización con los datos; codificación de los datos; generación de temas iniciales a partir de los códigos y los datos codificados; revisión y desarrollo de los temas; definición, denominación y perfeccionamiento de los temas; y redacción del informe (Braun y Clarke, 2022).

Vasco y se trasladó el compromiso ético en la investigación, incluyendo todos los aspectos relativos a la confidencialidad. No se permitió la grabación ni la recogida de ningún dato identificativo o trazable. Por todo ello, no se aportan extractos literales trazables, pero sí la mención de algunas palabras ilustrativas destacadas durante las reuniones observadas.

6.3. LIMITACIONES

Además de remitirnos a las limitaciones generales de este tipo de metodología cualitativa, debe considerarse el impacto potencial de la observadora en las reuniones[65]. Adicionalmente, cabe señalarse que no se han visto, en todos los centros, los diferentes tipos de reuniones de los equipos, con distintos módulos. Asimismo, en algunos casos algunos técnicos no pudieron acudir a la reunión y, en todo caso, hay decisiones que se reservan a la Junta[66].

6.4. RESULTADOS DEL ANÁLISIS TEMÁTICO CON PALABRAS Y EXTRACTOS ILUSTRATIVOS

Se han identificado seis temas emergentes del análisis de las notas de campo de la observación:

(1) Complejidad en el manejo y limitaciones de la información disponible.

(2) Interdisciplinariedad en el trabajo en equipo.

(3) Retos y complementariedad a la hora de llegar a consensos desde la diversidad de perspectivas técnicas y humanas, donde se reconocen posibles sesgos.

(4) Manejo diverso de tablas en papel de factores de valoración como criterios de orientación y justificación.

(5) Desconocimiento, suspicacia, interés y ofrecimiento de colaboración ante la experiencia del RisCanvi en Cataluña y su posible adaptación en el País Vasco.

65. Tras la presentación del proyecto de investigación, se me invitó a sentarme en la mesa con las personas técnicas o, en alguno de los casos ante la escasez de sitio, en un segundo plano.

66. Si bien, lo que se decide en equipo, se adopta en Junta, salvo votos en contra, falta de información, dudas o ausencias.

(6) Necesidad de mayores recursos humanos, materiales y tecnológicos.

Recogemos algunas palabras y frases ilustrativas pero que no pueden interpretarse de forma descontextualizada por lo que su valor es meramente ilustrativo de las complejidades que deben afrontar los equipos de los que se destila una gran profesionalidad e implicación y los que debemos un gran agradecimiento por su trabajo cotidiano y colaboración, un trabajo sumamente complejo, de alta responsabilidad, y poco reconocido socialmente. Las palabras precedidas de asteriscos se refieren a aspectos tratados en general y las palabras entrecomilladas son elaboraciones de la autora a partir de citas textuales recogidas a mano.

6.4.1. Complejidad en el manejo y limitaciones de la información disponible

* Alguna persona de la lista ya está en otra prisión.

* Falta de información, aunque tratan de coordinarse: "no sé si han cumplido los trabajos en beneficio de la comunidad, pedí la información, pero no la tengo todavía".

* Colaboración clave con entidades del tercer sector para poder garantizar salidas, permisos, etc.

* Muy pocas valoraciones de mujeres, en consonancia con su representación porcentual en prisión. No se advierten, a primera vista, diferencias en toma de decisiones sobre hombres y mujeres.

* En algún equipo se repasan también las personas dentro del programa de prevención de suicidios.

* Los propios profesionales comentan los problemas para la estimación o valoración de la voluntariedad y motivación de la persona interna de cara al seguimiento de programas y tratamiento.

* En ocasiones, se señalan confusiones administrativas dada la cantidad de información y los diferentes profesionales, dentro y fuera de prisión, concernidos.

* Alternancia de lenguaje técnico con más cotidiano.

* "No sé si sabe que se le va a expulsar, aunque tiene arraigo".

* No hay problema para expresar desacuerdos o votos en contra.

* Buena atmósfera, sentido del humor.

6.4.2. Interdisciplinariedad en el trabajo en equipo

* Colaboración con SAV (en especial en violencia de género), jueza de vigilancia penitenciaria, entidades sociales, Gobierno Vasco…

* Trabajo muy técnico, abreviaturas.

* Importancia de tener trabajo como elemento de confianza para regresar al centro tras un permiso.

* Hacer tratamiento y trabajar es complicado por horarios, etc.

– "Por otra parte, la posibilidad de trabajo podría relacionarse con más consumo o entrada de droga en prisión".

– Importancia de la necesidad de control de impulsos.

* Se habla de la soledad de algunas personas internas.

* Importancia del trabajo y la familia.

* Necesidad de cursos de digitalización (algunas personas no saben enviar un email).

* La percepción de la gravedad: "Es un delito gravísimo, cometido de una forma brutal, y me alegro de que siga un programa y le vaya bien, pero no voy a informar positivamente de cara al permiso" (TS), "pero ha asumido lo realizado" (Psi).

6.4.3. Retos y complementariedad a la hora de llegar a consensos desde la diversidad de perspectivas técnicas y humanas, donde se reconocen posibles sesgos

– Conocimiento individual de los internos (con nombres, temas familiares y de pareja, adicciones, enfermedades, etcétera), incluso de un número relevante de presos extranjeros, algunos con la problemática de ser adultos que fueron menores extranjeros no acompañados que vienen de pisos tutelados.

* En prisiones con más internos el conocimiento individualizado es comprensiblemente menor, particularmente si hay falta de funcionariado.

* Matices muy diversos por parte de las distintas personas que se escuchan y reflexionan en conjunto.

– Todos los agresores sexuales son muy normales, pero… A este le haremos seguimiento.

– Alguno no quiere salida controlada porque le falta un mes para terminar la condena y no quiere llevar la pulsera y que le vean.

– Yo trabajo para la sociedad, no para el interno, para eso está su abogado.

– Yo pienso aquí más como la víctima, no lo veo, aunque puede pensarse en salidas programadas y potenciar que siga trabajando.

– A veces haces una propuesta de tratamiento o permiso y el interno no quiere en esas condiciones.

– El control de las analíticas a veces es superficial.

– También casos de españoles que vienen a cumplir la condena, en su caso, el último tramo, en España, por ejemplo, por tráfico de drogas en Francia, pero se carece de información, normalmente, sobre el cumplimiento en ese país.

– "No es conflictivo".

– "Nos la ha jugado (no ha vuelto de un permiso), pero tiene un problema y está solo… es un fiestero, un sinvergüenza, tiene miedo a la libertad, conmigo ha acabado, no puedo más, tiene muchos problemas médicos, como indica el informe, viene quebrantando con muy malas formas la confianza de la Junta desde hace años. Si se le traslada, ¿qué pasa con mi relación terapéutica?".

* Búsqueda de alternativas o de progresión de medidas favorables cuando se deniegan permisos, etc.

– "Creo que ha tenido mala suerte, me he leído los delitos y no parecen muy graves, conducta intachable en prisión, le veo muy bien si tiene apoyo".

– "Le podemos decir que nos presente algo a la Junta, una oferta de trabajo, una propuesta razonable…".

* Peligrosidad: "es agresivo… va a decir que estoy conspirando contra él… no ha hecho el tratamiento de drogas", "eres favorable, pero en libertad puede ser muy agresivo… la puede piciar".

– "Está superdesajustado… pero la actividad delictiva no es grave… la peligrosidad es la clave (delito de conducción temeraria), si hay, para dentro… hay que medicarle con el psiquiatra".

– Mañana en Junta ya tomaremos la decisión, intentaremos recabar algo más de información.

– En valoración de informe sobre pronóstico, si está haciendo todo bien y no hay incidencias, debe ser favorable.

– Yo ya estoy apuntando en ficha, al final, es colaboración de todos para que no se nos pase este tema.

– Tiene que hacer programa de violencia de género y no damos otro paso hasta que lo haga.

– Aunque tenemos que hacer una revisión del plan de ejecución cada tres meses, no siempre se hace, no es necesario si no hay cambios, aunque el interno lo pide.

– La empresa está conforme en que haga la formación, pero fuera del horario de trabajo y eso es ilegal.

– Este señor nos torea.

– Este chico tan majo va muy bien, participa en deporte, le ayuda y él ayuda a otra gente, pero hay que cuadrar los horarios.

– Yo creo que le va a favorecer más estar fuera que dentro… trabajar, mantener la relación familiar… se le podría poner la pulsera.

– ¿Lo tenéis claro?… Yo tengo dudas.

– ¿Nos tiramos a la piscina?

– Me fío de ti, ¿le ves voluntad de cambio?

– Las veces que he estado con él, sí.

– Incógnita total, sé que no puedo abstenerme, pero lo haría.

– ¿Qué es lo que le va a venir mejor al interno?

– Yo por una parte me lo quitaría, tenemos un exceso de internos.

– ¿Qué hacemos? Por una parte, no tiene vínculos, pero, por otra parte, tiene proyectos.

– Le he cogido hasta cariño.

– Porque no le has conocido.

– Intenta dar pena.

– Yo le vi mal.

– Entonces tenemos que hablar con más personas.

– ¡Qué horror! (referido a un delito de trata con motivo de explotación sexual).

– Es un mentiroso (referido a un caso de agresión sexual).

– Es buen interno, no creo que dará problemas (referido a un delito de tráfico de drogas).

– ¿Y el tema agresividad?

– Es un segundo grado como la copa de un pino…va a estar aquí mucho… es un delito muy grave (tentativa de asesinato) y mira que creo en la reinserción (indican que la víctima está en coma y que la Diputación dio una ayuda económica).

– Se pide realizar la analítica en otro pueblo para evitar el estigma, aunque pueden surgir otros problemas.

– (caso de violencia de género y sexual)… parece tan correcto y tan majo, pero este es oscuro de narices.

– Le pondría un tercero, pero va a buscar trabajo y seguir programa y si incumple, regresa en grado.

– Yo soy desfavorable… Yo no.

– A este centro vienen internos que ellos mismos se sorprenden de que en otro les den el tercer grado, ya que estaban en observación en el de procedencia.

6.4.4. Manejo diverso de tablas en papel de factores de valoración como criterios de orientación y justificación

* Entre otras, se maneja la tabla 3, titulada "Razones de denegación de permiso por Junta de Tratamiento", con catorce grupos de razones[67] donde se incluye, por ejemplo, "extranjero ilegal" en cuanto la razón de vinculación.

67. Sobre requisitos (grado inadecuado, preventivos, no cuarta parte y mala conducta); actividad delictiva; condenas; conducta penitenciaria; uso de permisos y libertad condicional; drogodependencias, vinculación; actitud reinserción; personalidad-conducta; prematuro (por insuficiente información, etc.); tercer grado finalizad: art. 104. 4 y 197 RP); generales (riesgo de reincidencia, etc.); tramitación (situación tránsito, etc.); y permisos extraordinarios.

* Dan su opinión y diversa información y luego tratan de encajar la decisión con los factores de las tablas.

* Se toma la decisión y luego se justifica con la tabla, un poco el mundo al revés. Sin perjuicio de contar con mayores explicaciones en el informe.

* Tablas que tienen conceptos un tanto indeterminados, por ejemplo: "no presenta un proyecto de vida realista", "pronóstico elevado de reincidencia", "no cuenta con vinculación o arraigo suficientes", "motivación extrínseca para la realización de programa específico de tratamiento". También se maneja una tabla con carencias (ausencia de apoyo social, drogas, ausencia de formación profesional, inadaptación…) que se corresponde con actividades prioritarias (laboral, terapéutica, formativa, ocupacional, deportiva o educativa). Se subraya la multirreincidencia, en su caso.

También existe una plantilla sobre concesión de permisos, "a pesar de la naturaleza delictiva", sobre motivos que "hacen predecir un buen uso" de un permiso (en ella se mencionan, por ejemplo, la primariedad delictiva, la antigüedad de los hechos, la escasa relevancia penal y social de los hechos, la escasa peligrosidad, el inicio tardío en la actividad delictiva, no carrera delictiva, delito ocasional, libertad provisional sin delitos, eximente incompleta o atenuante cualificada, disminución de la peligrosidad por enfermedad o edad o ingreso voluntario en prisión).

* La responsabilidad civil sale en la toma de decisiones ya que se encuentra en la legislación y se recoge en las tablas que se manejan. Se indica que algunos abogados de las víctimas se han quejado de que la notificación del pago muy pequeño mensualmente, se vive mal por parte de algunas víctimas. También parece apreciarse diferencias por tipo de delito. Se afirma que *Aukerak*, organismo del Gobierno Vasco[68], ha conseguido aumentar claramente[69], y por comparación con otras administraciones penitenciarias, el número de personas internas trabajando, pero, como en ocasiones no cobran mucho, tampoco pagan siempre la responsabilidad civil, aunque, en principio, quien trabaja debe pagarla (en una ocasión no constaba pago, aunque estaba trabajando[70]). El Juzgado también puede estimar que la cantidad a pagar es insuficiente y si no se paga la responsabilidad civil se puede incluso perder el tercer grado".

68. Vid. en https://aukerak-gea.euskadi.eus/inicio/
69. En concreto, en el centro de Zaballa, en el momento de realización de la visita trabajaban, aproximadamente, la mitad de los internos. De hecho, algunos internos piden ingreso en Zaballa porque dicen que tiene fama de dar más terceros grados y trabajo (fuente: conversación con técnicos del CP).
70. "Hay gente que no paga y ellos tienen que decir cuánto van a pagar al mes".

En algunas tablas se recogen cuatro aspectos respecto de la responsabilidad civil: voluntad manifiesta de abono de la responsabilidad civil impuesta; compromiso ante el tribunal de su abono; se encuentra abonando la responsabilidad civil; o responsabilidad civil abonada en su totalidad.

* En ocasiones, la responsabilidad civil se paga por la familia.

– Es verdad que cobra muy poco y que vive en un mal barrio, pero no sabemos si la víctima lo necesita más o menos… Bueno, no la conocemos.

* Referencia a los presos de ETA.

* Aritmética penal para cálculo de salidas, con referencia al artículo pertinente de la normativa aplicable (trabajo técnico).

* Cambios longitudinales en informes.

* Caleidoscopio de informaciones, conocimientos, relaciones y perspectivas.

* Papel coordinador de subdirectores o personas que les sustituyen.

* Exclusión social y drogadicción, "tiene que hacer algo con el tema de los consumos (TS)… pero no se le puede obligar y hay que motivarle", "¡qué mal pronóstico tiene ese chico, pronóstico bueno a pesar de cuestiones culturales, idiomáticas, religiosas, tóxicos…!".

– Yo también quiero confiar en él, pero…

– Es una persona camaleónica que se mueve entre estar bien o estar mal con internos y funcionarios.

– ¿Por qué nos quedamos a gente cuando iban destinados a otro centro y estamos con mucha gente? Sigue un tratamiento aquí, pero no era su destino inicial.

* Se cita el Excel que tienen los educadores en el *sharepoint*.

* Trabajo a lo largo del tiempo.

– Mostraba arrepentimiento, pero hoy está sobrado, aunque conmigo bien.

– No tiene dónde dormir.

– De tener una vida normal a verse tirado, y no va a dejar de consumir.

* Cuestionamiento de prohibición de móviles en prisión.

* Consumo de tóxicos con falsa percepción de control.

– Actitud narcisista y borde siempre, pero, bueno, hay que valorar más cosas.

* Discrepancias a la hora de valorar el peso de una incidencia en prisión a lo largo del tiempo (relacionado con desistimiento).

– Me voy a mantener en voto en contra porque no tiene arraigo y tiene expulsión… aunque al final no tenga otra condena pendiente.

– No le voy a premiar con el permiso.

– Dejó embarazada a la novia tras el quebrantamiento… tiene otro niño con una enfermedad mental grave.

– Tiene capacidad 0 de esfuerzo, va a consumir, va a discutir, va a agredir… Esa pareja, los dos, se han criado en un centro de menores, no tienen nada.

– Es reincidente, hay mucho riesgo, y además tiene otra citación.

– Sí, pero como testigo.

– Él no se espera salir, sabe que tiene cosas pendientes.

– Poner lo que queráis como motivación, yo creo que es el 12 el que procede.

– En vez del 10, el 9 mejor, ¿no?

– ¿Cuál era el 10?

– "Estamos hablando de alguien que ha aprendido la lección y este otro no la ha aprendido".

– "Tenéis que hablar con él, tiene trabajo, e intentar que no lo pierda".

– "Nos ha enseñado la foto de sus preciosos cuatro hijos".

– "Un interno me dijo que le obligamos a un curso que no le sirvió para salir".

– "Estamos revisando condenas más graves en algunos tipos de delitos y les estamos dando permisos".

– "No es que sea más grave, es otro contexto".

– "La gravedad es subjetiva"

– "Lo más importante es que consiga un contrato fuera".

– "A mí igual me convences, pero a otra compañera, no".

– "No me la voy a jugar en un delito tan grave, lo lleva bien, pero es pronto".

– ¿Lo dejamos pendiente para Junta?

– Sí, porque los votos pueden cambiar.

– Mala higiene, malas contestaciones… me parece darle un regalo.

– No tiene una conducta ejemplar.

– Me cayó una lágrima y luego la recogí.

– El día que eches lágrimas, llámame para que lo vea.

– ¿Entonces qué es, 86? 4?

* Tema: voluntariedad.

– Entonces, 3, 8, 10, 11 y 24.

– Parece la loto.

6.4.5. Desconocimiento, suspicacia, interés y ofrecimiento de colaboración ante la experiencia del RisCanvi en Cataluña y su posible adaptación en el País Vasco

– "A los internos no les molestaría que escribieras en el ordenador mientras les entrevistas … incluso se podría cruzar con información de psicólogos, etc. de otros centros penitenciarios".

– "La informática está para romper barreras y facilitar la información de dentro y fuera del centro penitenciario o de un territorio u otro … Cada equipo podría decidir si visible pero no editable o capado totalmente".

* Alusión a programas específicos de tratamiento (violencia de género, delitos violentos…).

* Tiene un pronóstico muy malo, con motivos intrínsecos y extrínsecos.

* Insistencia en la importancia de informatizar, pero carencia de medios.

– "Nos han quitado el despacho que tenía el ordenador; tienen que habilitarnos otro".

"La normativa no permite, en sección abierta, acceder a Internet, pero el Reglamento lo permite. ¿Sin vigilancia?".

* Interés por el estudio y los podcasts sobre RisCanvi. De hecho, alguna persona había hablado previamente con compañeros catalanes y le habían manifestado preocupación por la aplicación diversa de la herramienta, según cada centro, y la necesidad de alimentar los datos, en particular en relación con los factores dinámicos.

6.4.6. Necesidad de más recursos humanos, materiales y tecnológicos

– "No damos abasto. Hay mucha movilidad en las bolsas que no son específicas para trabajar en prisión. Algunas personas que intervienen lo hacen como primer día de trabajo. Otras personas están de baja o vacaciones, pero nos falta personal".

* Contradicción entre régimen y tratamiento.

* Sobrepasados por tema de la demanda social en tema de inmigración y delito y con falta de funcionarios.

* Carencias (en la misma reunión, se avisa que un técnico ya no trabajará en el centro): "No puede ser, no hay oposiciones, solo bolsas, para cuando les llamas, ha pasado mucho tiempo, tienen que pasar un curso y para cuando entran… Se les forma y se van. Esto es el día a día, les formamos y se van".

* Problemas, en ocasiones, con la visualización de datos en el ordenador (generalmente el Sistema de Información Penitenciaria propio y del Ministerio del Interior), no siempre se consigue visualizar todo lo solicitado o interesado.

* Hay bastantes decisiones que se aplazan a la Junta, a veces porque falta información o una persona.

6.5. CONSIDERACIÓN FINAL

Del análisis de la observación realizada puede apreciarse la necesidad de contar con instrumentos objetivos de valoración[71], pero también con

71. Con cuestionamiento respecto de la calidad de los datos con los que se trabaja respecto de su representatividad, complementariedad y veracidad, por ejemplo.

el juicio clínico o profesional de cada experto cuya matización por otros saberes resulta en un interesante[72], y complicado equilibrio de perspectivas, de cara a una decisión sobre derechos de las personas bajo control penitenciario que, además, se enmarca en un contexto normativo y una realidad social concreta, donde los recursos disponibles para el sistema penitenciario y social y para las personas, familias y comunidades concernidas es desigual. Una herramienta complementaria digital facilitaría la recogida y gestión de la información, sin perjuicio de la necesidad de decisiones humanas responsables.

7. CONTRASTE A TRAVÉS DE UN CUESTIONARIO Y UN GRUPO DE OPINIÓN DESARROLLADO CON ESTUDIANTES UNIVERSITARIOS VASCOS

Dentro del desarrollo de este trabajo se decidió realizar un breve cuestionario online, utilizando Microsoft y con garantía de total anonimato sin recogerse datos identificativos o trazables. Se invitó a los estudiantes a rellenarlo, de forma totalmente voluntaria, a lo largo de los meses de septiembre y octubre, de forma voluntaria, tras escuchar dos episodios de un pódcast sobre RisCanvi[73]. Se invitó a participar a unos treinta estudiantes de segundo de Criminología, así como a algunos alumnos de Derecho de la Facultad de Derecho de la EHU. El cuestionario podía completarse en euskera, castellano o inglés[74]. Asimismo, se invitó a todos estos estudiantes, el día 13 de octubre de 2025, al seminario abierto sobre RisCanvi, presencial y online, en el campus de Gipuzkoa[75], donde también hubo posibilidad de intercambiar impresiones con las personas expertas invitadas, la mayoría coautoras de esta monografía, además de representantes del servicio de prisiones del Gobierno Vasco.

---

72. A nivel individual cabe preguntarse en qué se basa exactamente el juicio clínico en la formación, experiencia, intuición, relación con la otra persona, etc.
73. Se trata de los dos episodios del pódcast RisCanvi en "Algoritmos y gobiernos", producidos en febrero de 2022 y con una duración aproximada de una hora, con entrevistas a expertos y profesionales afectados por su puesta en marcha, incluyendo miembros de la Abogacía, con los que la autora del capítulo tuvo la oportunidad de coincidir en una formación en Cataluña, y funcionarios de prisiones. Puede accederse al pódcast en https://open.spotify.com/show/7lLPt2yGwdwVHlfXxq2dcV
74. Para dar entrada a los estudiantes Erasmus, además de al grupo en inglés del grado de Criminología de la Facultad de Derecho de la Universidad del País Vasco.
75. Véase la convocatoria en https://www.ehu.eus/es/web/ivac/-/retos-criminologicos-ante-los-instrumentos-de-valoracion-y-gestion-de-riesgos-en-el-ambito-penitenciario

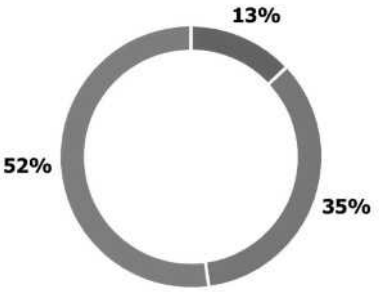
7.1. RESULTADOS: NÚMERO DE RESPUESTAS Y PERFILES SOCIODEMOGRÁFICOS

Se obtuvieron 23 respuestas a un cuestionario de tres preguntas abiertas, con posibilidad de añadir algo más al final, y con tres preguntas relativas a datos sociodemográficos.

Respecto del perfil demográfico de las personas que respondieron, pueden verse los siguientes gráficos.

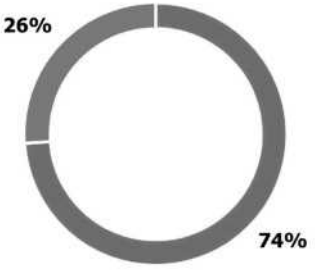
– ¿Cuál es tu nivel de dominio de la tecnología en tu día a día y en tus estudios?

- Principiante/Hasiberria/Beginner 3
- Intermedio/Tartekoa/Intermediate 8
- Avanzado/Aurreratua/Advance 12
- Experto/Aditua/expert 0



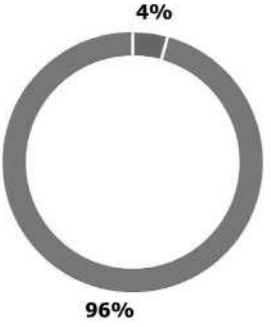
– ¿Cómo te identificas?

- Mujer/emakumea/woman 17
- Hombre/gizona/man 6
- No binario/ez-bitarra/non-binary 0
- Prefiero no responder/Nahiago dut ez erantzun/I prefer not to answer



– ¿Qué estudias?

- Derecho/Zuzenbidea/Law 1
- Criminología/Kriminologia/Criminology 22
- Doble Grado (Derecho + ADE)/Gradu bikoitza (Zuzenbidea + enpresa)/Double degree (Law +… 0
- Otras 0



A continuación, dado su interés, como muestra, no representativa, pero sí meramente ilustrativa de la opinión de los jóvenes sobre RisCanvi, pasamos a reproducir extractos literales ilustrativos de cada respuesta, identificada con una alternancia de colores.

7.2. MAYOR IDENTIFICACIÓN CON ALGUNA DE LAS PERSONAS QUE PARTICIPAN EN EL PÓDCAST SOBRE RISCANVI (ABOGADOS/AS, FUNCIONARIOS/AS, TÉCNICOS/AS QUE LA ELABORARON, PERFECCIONAN Y/O AUDITAN, CONDENADOS/AS...) Y MAYOR PREOCUPACIÓN EN EL USO DE RISCANVI

Me identifico mayormente con los técnicos. Me parece una posición muy interesante poder analizar y ofrecer soluciones a un problema tan complejo. Como hemos escuchado en el podcast, existen herramientas predictivas como esta en muchos países, especialmente en los de habla inglesa, pero Cataluña no contaba con esta capacidad. Por ello, se pidió a los técnicos que ofrecieran una herramienta eficaz para estudiar los marcadores y prevenir la reincidencia en delitos violentos y comportamientos violentos dentro de las prisiones. Dado que el objetivo es evitar que las personas reincidan, diría que mi mayor preocupación es si la herramienta realmente cumple su función o no.
Lo cierto es que no sabría quedarme con uno solo, por un lado, están los presos, cuyo futuro se pone en manos de una herramienta algorítmica de la cual no conocen ni su funcionamiento y les puede hacer pensar que la decisión está fuera del alcance de su mano. Siguiendo con los presos, los abogados que los defienden se tienen que enfrentar a una resolución judicial que se da como científica, dificultando su defensa. Por último, por el lado de las víctimas, es normal que tengan miedo de la liberación del preso, sobre todo si se ha dejado en manos en una herramienta algorítmica que, al igual que los presos, no entienden.
Personalmente, me siento más identificado con los funcionarios que deben aplicar la herramienta, ya que mi principal preocupación es que el algoritmo sea muy rígido en sus criterios, obviando la subjetividad y la experiencia humana y dando lugar a que personas que lo merezcan se les deniegue un tercer grado o un permiso.
Considero que me identifico más con los técnicos que se encargan de mejorar la herramienta dependiendo de las necesidades del momento. Es muy difícil crear una herramienta que englobe los mismos ítems con el transcurso de los años, es por ello que actualizarlos es necesario para ser más objetivos y concretos. Otra de las preocupaciones que se han mencionado en el podcast, y considero que es muy importante tener en cuenta, es la validez científica, es decir, tratar de evitar falsos positivos o negativos, ya que de esa manera se conseguirá llegar a tener más precisión. En relación a esto, considero que es esencial crear algoritmos

que no reproduzcan desigualdades como pasó con COMPAS con las personas negras, inmigrantes, etc. Es importante destacar también la transparencia necesaria en el uso de esta herramienta, haciendo ver que no es un algoritmo impuesto, sino que se trata de una herramienta útil para todas las personas relacionadas con los internos, ya sean funcionarios, jueces, abogados etc. Por último, los derechos y la privacidad de los internos es un aspecto a tener en cuenta, y que puede causar controversias. Creo necesario que los presos sean conocedores de que sus datos serán utilizados en esta evaluación y sean informados de la utilidad de la misma, de manera que se evite esa percepción de control político y social. A efectos prácticos en la rehabilitación, como se ha mencionado, considero importante detectar las áreas de intervención y no solo el tipo de riesgo (alto, medio o bajo).

Escuchadas las diferentes posiciones, me identifico más con la de los abogados, respecto de la utilización de la herramienta. Comparto su visión crítica, en particular, la posibilidad de sesgos y carga ideológica que pueda acentuar la exclusión social. Mi principal preocupación es que se fortalezca el estigma social, en lugar de apoyar una evaluación individualizada y justa de cada persona.

He estado bastante de acuerdo con los distintos puntos de vista expuestos en ambos episodios del podcast, pero, a la hora de valorar el uso de esta herramienta, me identifico más con los presos. Al fin y al cabo, se trata de una herramienta que, a pesar de ser realmente útil, no ha sido apenas objeto de estudio. Muchos beneficios del reo quedan en manos de una herramienta que no saben en qué criterios se basa para dar un resultado u otro, no saben si puede estar sesgado como otras herramientas internacionales, no saben en qué momento están siendo analizados por este instrumento, etc. Por ello, lo que más me preocupa es la falta de transparencia y los sesgos que puedan estar presentes en su *software*.

Me identifico más tanto con los abogados como con los científicos/psicólogos. Creo que esta herramienta predictiva puede ser útil siempre y cuando su algoritmo sea de conocimiento general, puesto que, si no sabes cómo funciona, difícilmente te puedes asegurar de cómo de fiable resulta.

Me identifico principalmente con los funcionarios mayormente, y la preocupación es más general, aunque encuentro muy complicado el implementar con eficacia un programa como RisCanvi.

Me identifico principalmente con los abogados que afirman que el uso de herramientas algorítmicas o la IA supone un cambio de paradigma, ya que se introduce un sistema predictivo diferente en el ámbito jurídico. Creo que es normal sentir incertidumbre ante estas tecnologías. En mi opinión, es fundamental ser escéptico con respecto a RisCanvi, no considerarlo una realidad, sino más bien una herramienta que puede cuestionarse y que puede equivocarse. Debe haber un punto medio entre utilizar los métodos tradicionales y ceder todo el poder decisorio a herramientas algorítmicas o a la IA.

Desde mi punto de vista, creo que el argumento de los abogados fue el más relevante: aunque la aplicación del algoritmo RisCanvi pueda ser útil, si no es transparente, constituye una violación de los derechos de los presos. En el artículo 24.1 de la Constitución española se reconoce el derecho a no quedar en situación de indefensión, lo que de hecho ha ocurrido en ocasiones cuando un preso solicita un permiso de salida y el abogado no ha podido refutar la decisión del algoritmo, dado que es imposible saber cuál fue el proceso de razonamiento y cuáles fueron las razones que lo motivaron. De todos modos, como se explica en el podcast, si la herramienta RisCanvi se utiliza mientras personas reales revisan su trabajo y toman la decisión final, puede ayudarles a decidir.
Me identifico más con los técnicos que elaboran y perfeccionan la herramienta, porque deben equilibrar la objetividad de los datos y equilibrarlo con la complejidad humana de cada caso. La mayor preocupación es que la herramienta termine sustituyendo el criterio profesional y se confíe demasiado en sus resultados sin un análisis crítico.
Después de escuchar los dos episodios, me siento más identificada con los abogados y con los funcionarios, ya que para ambos grupos me parece que es una herramienta muy efectiva para entender personalmente la situación de cada preso. Pero lo que más me preocupa también sería la clasificación que se le hace a cada preso, ya que se puede clasificar mal a las personas o reforzar prejuicios, por ejemplo, como se hace con los presos extranjeros. Además, me da algo de miedo que se confíe demasiado en lo que dice la herramienta, que siempre se basa en la probabilidad y puede que no siempre sea así y se deje de lado la valoración humana y personal de cada caso.
Me identifico más con los abogados, ya que el uso de herramientas de este tipo podría suponer una incapacidad de actuación de los abogados a la hora de representar y defender.
Me identifico más con los técnicos y profesionales que trabajan en prisión porque conocen tanto los límites técnicos como las consecuencias reales del uso de la herramienta. Mi principal preocupación es que se tomen decisiones automáticas que afecten derechos, sin transparencia ni posibilidad de revisión, y que se reproduzcan sesgos existentes.
Me identifico más con las personas presas, ya que son quienes sufren directamente las consecuencias del uso de RisCanvi. Aunque la herramienta se presenta como un medio objetivo para evaluar el riesgo de reincidencia, me preocupa que pueda reforzar prejuicios o estigmas, especialmente si se basa en datos que reflejan desigualdades sociales o comportamientos interpretados fuera de contexto.
Me identifico más con los presos y abogados, porque me preocupa que el RisCanvi se use como una verdad absoluta y no como una ayuda. Mi mayor preocupación es que la herramienta no sea transparente y pueda reproducir prejuicios, afectando decisiones importantes sobre la libertad de las personas.

Me identifico más con los técnicos que tratan de desarrollar una herramienta mejor para trabajar en prisión, aunque mi mayor preocupación es cuán humana y precisa es dicha herramienta.
Después de escuchar ambos episodios, me identifico más con las personas presas porque al final son quienes sufren directamente las consecuencias de lo que decida la herramienta. Mi mayor preocupación es que se le dé demasiado poder a un sistema que no es totalmente transparente ni perfecto: puede tener sesgos, basarse en datos injustos del pasado y acabar perpetuando desigualdades. Además, si se confía ciegamente en lo que "predice", se corre el riesgo de deshumanizar a las personas y dejar de valorar sus circunstancias reales. También me preocupa la falta de claridad sobre qué datos se usan, cómo se protegen y quién responde si hay errores.
Con los abogados. La mayor preocupación a la hora de utilizar esta herramienta es que la decisión que se tome respecto al sujeto, aunque sea una herramienta de soporte y no vinculante puede coartar los derechos y libertades del preso en cuestión.
No puedo evitar pensar en los presos, me pongo en su lugar y de mis mayores preocupaciones sería que, por decisiones de terceras personas (padres, tíos u otros familiares) que cometieron delitos, acabe perjudicada yo.
Me identifico más con los abogados. Mi mayor preocupación es que se deja el futuro del preso en manos de una herramienta y no en una persona.
Me identifico más con los funcionarios y las personas técnicas que explicaban cómo funciona la herramienta. Entiendo que puede ser útil para organizar mejor el trabajo en prisión, pero me preocupa que el algoritmo no sea del todo justo o que se equivoque por los datos que usa. También me da algo de miedo que se tomen decisiones importantes solo por lo que diga una máquina.
Me identifico más con la postura de los técnicos y técnicas que elaboran, perfeccionan y auditan el RisCanvi. Creo que pueden llegar a ser los más adecuados a la hora de elaborar una herramienta objetiva y estandarizada para evaluar el riesgo de reincidencia. La mayor preocupación que puedo llegar a ver es que se convierta en un instrumento de cierta manera estigmatizador, que reduzca a la persona a una puntuación y perjudique a ciertas personas en base a su contexto.

7.3. MAYOR BENEFICIO QUE APORTARÍA LA UTILIZACIÓN DE ESTE TIPO DE HERRAMIENTA EN LAS PRISIONES VASCAS

En mi opinión, el mayor beneficio de utilizar este tipo de herramienta, no solo en las prisiones vascas, sino en las prisiones de todo el mundo, sería la reducción de las reincidencias. Sin embargo, debemos tener en cuenta los detalles de estas prisiones vascas en relación con la sociedad vasca, ya que cada sociedad tiene una composición y un alcance diferentes en cuanto a los delitos que se cometen. Por lo tanto, debemos tener en cuenta el número y los porcentajes de los delitos cometidos por los delincuentes, su edad, su origen socioeconómico, su estado de salud mental o, como se menciona en el podcast, su raza. Esto podría influir en la precisión, la equidad y la flexibilidad de esta herramienta predictiva, que se ajustaría interpretando cuidadosamente los datos específicos de esa sociedad para garantizar que no discrimine ni minimice los resultados al centrarse únicamente en una sola característica.

Principalmente la eficacia de esta, ya que provoca que no sea necesaria tanta mano de obra humana, la cual es limitada. Sin embargo, considero que ese tiempo ahorrado deberá usarse para los casos más complejos, permitiendo dar un mejor resultado. Por lo tanto, no debe usarse únicamente para quitar trabajo, sino para permitir dar un trato más elaborado en los casos que más lo requieran.

Sería un buen modo de ahorrar tiempo y recursos, agilizando el proceso de reinserción, aminorando riesgos y permitiendo destinar los costes económicos y humanos a otras cuestiones de calado.

Considero que es una oportunidad clave para que las tomas de decisiones sean más objetivas, es decir, al conocer que en las prisiones del País Vasco hay una gran cantidad de personas del extranjero, serviría como método para eliminar la subjetividad individual de cada funcionario o juez a cargo del caso. Sin una herramienta objetiva como esta, es más fácil que las percepciones individuales afecten más en la toma de decisiones. Si se implanta esta herramienta, todos los internos estarían evaluados bajo los mismos criterios. Otros beneficios a tener en cuenta, sería en torno a la detección temprana de riesgos individuales para poder identificar perfiles con mayor probabilidad de reincidir o de tener conductas violentas dentro de prisión, además de tener la posibilidad de diseñar programas de intervención específicos, así como adicciones, salud mental etc.

En mi opinión, el supuesto beneficio de utilizar en las cárceles vascas una herramienta como RisCanvi sería, en teoría, la posibilidad de disponer de datos más sistematizados para orientar las decisiones. Sin embargo, creo que este beneficio es muy relativo, ya que el instrumento no parte de criterios neutros. Más que mejorar la gestión penitenciaria, su uso podría reforzar dinámicas de control y exclusión, legitimando bajo la apariencia de objetividad técnica a los diferenciales preexistentes.

El mayor beneficio que aportaría es vinculado a la reinserción, puesto que se trata de una herramienta que identifica los factores de riesgo que aumentan las probabilidades de que un preso reincida. Mediante la identificación de dichos factores, los funcionarios pueden tener en consideración qué aspectos del individuo requieren una mayor atención e intervención, para poder adoptar las medidas necesarias que ayuden a su reinserción.
Este tipo de herramienta podría ayudar con asuntos de sobrepoblación penitenciaria, puesto que, quieras o no, es necesario liberar a gente si queremos que entren otras personas, por lo que un sistema así resultaría bastante útil para lidiar con problemáticas así. Por otro lado, también es una herramienta bastante funcional para la reinserción, puesto que es mejor poner un mayor esfuerzo en quienes tienen menos probabilidades de reincidir.
Si fuese posible, la eficacia en la predicción del riesgo de reincidencia, aplicado mayormente y con más importancia a delitos graves (homicidios, delitos sexuales...).
El uso de estas tecnologías debe mejorarse antes de implantarlas realmente en nuestro sistema judicial. Es cierto que los elementos que utiliza para realizar una valoración de un preso pueden ser políticos, y eso debe convertirse en algo libre de ideología, convirtiéndolo en una herramienta objetiva en el ámbito en el que puede funcionar correctamente. Esto significa que debe haber un control humano de las decisiones que toma RisCanvi; por ejemplo, si afirma que alguien está en peligro de reincidir, algún profesional también debe realizar las pruebas manualmente para asegurarse de que las decisiones se toman de forma adecuada.
Se sabe que las cárceles vascas han estado últimamente en una situación de sobrepoblación, por lo que podría ser muy útil utilizar este tipo de herramienta también aquí. Esto se debe a que, al implementar esta herramienta, los trabajadores responsables de decidir si conceden o no los permisos podrían trabajar más rápido, ya que tendrían todos los factores de riesgo a mano; lo que conduce a más resoluciones de solicitudes de permisos y, por lo tanto, a más presos con la posibilidad de salir y menos que tengan que permanecer en las instituciones abarrotadas.
El mayor beneficio sería una evaluación más objetiva del riesgo, que ayude a tomar decisiones más justas y fundamentadas sobre permisos, reinserción y seguimiento de los presos, reduciendo la influencia de prejuicios personales.
Esta herramienta traería muchos beneficios a las prisiones vascas ya que se valoraría mejor el riesgo y las necesidades de cada uno de los presos, además de que ayudaría a planificar intervenciones y así mejorar la seguridad ciudadana y evitar la reinserción. También, esta herramienta ayudaría a tomar decisiones sobre estos presos de una manera más segura y coherente.

Creo que podría ser útil para la prevención especial, pero a lo que se refiere la prevención general... No creo que sea una herramienta que se deba usar, independientemente del contexto.
El mayor beneficio sería poder personalizar mejor las intervenciones y usar los recursos de forma más eficiente, ayudando a prevenir la reincidencia y mejorar la reinserción, siempre como apoyo a la decisión humana y no como sustituto.
El mayor beneficio sería poder anticipar riesgos y adaptar mejor las intervenciones a cada persona. Esto permitiría una gestión más preventiva y objetiva, ayudando a mejorar la seguridad y facilitar la reinserción, siempre que la herramienta se use con transparencia y supervisión.
El mayor beneficio sería que ayudaría a planificar mejor la reinserción, permitiendo detectar antes los riesgos y necesidades de cada persona presa. Así, se podrían adaptar los programas y recursos de forma más individualizada y eficaz.
Ahorraría tiempo a la hora de gestionar el riesgo.
En mi opinión, el mayor beneficio de usar una herramienta predictiva en las prisiones vascas sería que podría ayudar a tomar decisiones más rápidas y objetivas, especialmente en temas como clasificaciones, progresiones de grado o programas de reinserción. Si está bien diseñada y supervisada, puede detectar patrones de riesgo o de mejora que a veces se pasan por alto, y así optimizar recursos y ofrecer intervenciones más ajustadas a cada persona. En resumen, puede servir como apoyo técnico para hacer el sistema más eficiente y justo, siempre que no sustituya el criterio humano ni se use de forma ciega.
Tiene una alta capacidad para detectar a presos con riesgo bajo de reincidencia sin equivocarse.
La eficiencia que podría traer.
El mayor beneficio sería la celeridad que aportaría a la hora de evaluar el riesgo de los presos.
Podría ayudar a conocer mejor el riesgo de reincidencia de cada persona y a adaptar los programas o tratamientos a cada caso. Si se usa bien, serviría para mejorar la reinserción y aprovechar mejor los recursos.
Siempre y cuando hubiera cierto control, creo que agilizaría todo el proceso de la concesión de permisos, lo cual podría ayudar tanto a internos como trabajadores.



Imagen 4: Nube de palabras sobre posibles beneficios y necesidad de controles.

7.4. ADECUACIÓN DE LA INVERSIÓN DE FONDOS PÚBLICOS EN ESTE TIPO DE HERRAMIENTA Y, EN SU CASO, DE SISTEMAS DE CONTROL

Sí. Es adecuado que se inviertan fondos públicos en este tipo de herramientas, si realmente ayudan a prevenir las recaídas y contribuyen a la seguridad pública. Creo que debería haber algún tipo de control o límite para garantizar que la herramienta funcione de manera eficaz y que realmente ayude a la sociedad a lograr el resultado deseado, como evaluaciones periódicas de la precisión predictiva o informes periódicos sobre los resultados.
Si que considero adecuado que se inviertan fondos en este tipo de herramientas, ya que como he mencionado en la segunda pregunta, herramientas como estas aumentan la eficacia y permiten utilizar ese tiempo en otros apartados que lo necesiten, aunque siempre debe ser como algo complementario y nunca eliminar del todo la participación humana. El único control que pondría es demostrar su eficacia, para asegurarse de que no se ha usado los fondos con otros motivos.
Me parece adecuado invertir en esta tecnología pues es el futuro del sistema penitenciario. Sin embargo, creo que debe estar regulada y sus diagnósticos deben ser ratificados por profesionales, al menos hasta que se perfeccione su índice de aciertos.
Como ciudadano me parece bien que se inviertan los fondos públicos en este tipo de herramientas, ya que tienen su utilidad y favorecer a mantener el orden público. Es un beneficio social con gran potencial, que puede ayudar *a priori*zar recursos, detectar factores de riesgo etc. Sin embargo, debería ser aplicada conociendo que la herramienta vaya acompañada de garantías estrictas, supervisión, controles técnicos y legales. Estas herramientas pueden aportar objetividad, pero sin salvaguardas adecuadas puede reproducir discriminaciones, vulneraciones de derechos que afectan de forma grave a las personas. Los controles que llevaría a cabo estarían relacionados con el marco legal, la evaluación previa y continua, la clasificación de los riesgos, intervención etc.

Como ciudadano, la inversión de fondos públicos en un instrumento como RisCanvi sería adecuada si se solicitan y financian todos los apoyos necesarios y, además, se demuestra mediante evaluaciones independientes que realmente mejoran resultados significativos, como la rehabilitación o la reducción de riesgos reales, sin que aumente la exclusión social. De lo contrario, creo que los recursos públicos deberían destinarse a medidas humanas y sociales, como programas de reinserción, educación o salud mental, para abordar las razones profundas de la delincuencia y promover una justicia realmente transformadora, en lugar de invertir en soluciones tecnológicas que puedan ocultar decisiones políticas con una apariencia de objetividad.
Sí. Considero que se trata de una herramienta útil, puesto que contaría con una base de datos completa que pueda establecer patrones y tendencias relacionados con los factores de riesgo, pudiendo establecer la probabilidad que tiene un preso de reincidir basándose en datos objetivos. Sin embargo, me preocuparía más el hecho de que esta herramienta pasara de ser "orientativa" a una herramienta cuyos resultados apenas se discutan. Además, considero que también es muy importante la transparencia de su *software*. Por ejemplo, si un preso obtiene un resultado de riesgo alto, que pueda tener conocimiento de por qué ha obtenido dicho resultado, a qué factores se les ha dado una mayor importancia, la intervención del funcionario al obtener los resultados, etc.
Me parece adecuado, porque al fin y al cabo es una herramienta de uso para entidades públicas como pueden ser los funcionarios de las prisiones. Aunque evidentemente, emplearía esta herramienta para casos de dudosa claridad, por ejemplo, no emplearía esta herramienta con personas detenidas por violencia de género, pero sí con aquellas detenidas por homicidio imprudente. Esa sería una buena forma de asegurarse de que se emplea responsablemente y no habría que hacer el sobreesfuerzo de emplearla con todos los presos que entren a prisión.
Sí, pero estrictos; creo que se deberían invertir en otros posibles aspectos de rehabilitación del preso para evitar su reincidencia, pero creo que a la vez este como complemento podría ser interesante de postular
Creo que no es lógico intentar evadir estas mejoras en las tecnologías relacionadas con cada área de trabajo. Sin embargo, debe controlarse estrictamente, ya que aún se encuentra en fase de desarrollo y mejora. Lo que ocurre con la IA es que se transforma y evoluciona a medida que se utiliza, y si eso es cierto, el uso de RisCanvi debe hacerse de manera que se mejore a sí mismo. Sin embargo, esto debe hacerse con un control cuidadoso y respetando los derechos de los presos y de aquellos que tienen algo que ver con el sistema judicial.
En mi opinión, destinar una parte de los fondos públicos al desarrollo de este tipo de herramientas sería útil, pero solo si los fondos invertidos también sirvieran para pagar al equipo que revisaría su rendimiento y lo mejoraría cuando fuera necesario. En otras palabras, para controlar este tipo de herramienta debería haber un grupo de trabajadores que se asegurara de que el algoritmo no tomara

ninguna decisión irracional o que la razón detrás de una decisión no se basara en prejuicios. Además, dicho algoritmo debería revisarse periódicamente, teniendo en cuenta lo que predijo y contrastándolo con lo que realmente ocurrió, de modo que sus fallos puedan subsanarse en las predicciones futuras.
Sí, pero siempre teniendo varios controles en mente como transparencia en su funcionamiento, auditorías independientes y supervisión humana en las decisiones para evitar errores o sesgos.
Me parece adecuado que se inviertan fondos públicos en este tipo de herramienta ya que me parece muy efectiva para mejorar la gestión y la seguridad en las prisiones, pero siempre y cuando venga acompañada de unos controles y garantías. Primero que todo, es necesaria que se haga una evaluación previa antes de utilizarla, que haya una transparencia de los tipos de datos que se utilizarán con los presos y abogados, que estos segundos tengan la opción de apelar si lo creen necesario para que haya una valoración humana. Además, considero que no tiene que ser la única herramienta y que después de cada valoración haya también una valoración más profesional, humana y técnica. También considero, que como he mencionado antes, puede que haya algún tipo de discriminación o incluso una evolución en los presos, por ello, una vez cada X tiempo sería necesaria una revisión de la plataforma y si son necesarios cambios, una actualización. Así que, me parece adecuado invertir en este tipo de herramientas siempre y cuando se hagan de forma ética, justa, transparente y con supervisión constante, porque solo así se puede garantizar que realmente sirvan para mejorar y no para perjudicar a nadie.
No. Considero que habría que invertir en solucionar lo que lleva a que la gente delinca en vez de invertir en herramientas para abordar o tratar de solucionar un «problema" que a veces puede incluso no existir aún.
Sí, me parece adecuado invertir fondos públicos si hay controles estrictos: transparencia sobre cómo funciona, auditorías externas e independientes, uso responsable de los datos y garantía de que las decisiones finales siempre las tome una persona, no el algoritmo.
Sí, me parece adecuado invertir fondos públicos en esta herramienta si se usa con control y transparencia. Creo que debería someterse a auditorías independientes, garantizar el derecho a revisión humana y evaluar posibles sesgos. Además, sería importante que se explicara claramente cómo funciona y que los resultados no se usen como criterio único en decisiones sobre los presos.
Sí, me parece adecuado invertir en ella si se controla bien. Debe haber transparencia, auditorías independientes, supervisión humana y posibilidad de revisar decisiones. Así se evita el mal uso y se garantiza que sirva realmente para la reinserción.
Me parece que es apropiado si beneficia a la ciudadanía al menos a largo plazo.

Sí, me parece adecuado invertir fondos públicos en este tipo de herramienta, siempre que haya controles muy claros. Antes de aplicarla debería hacerse una evaluación de impacto, probarla en piloto y garantizar transparencia sobre cómo funciona. Además, tendría que someterse a auditorías externas periódicas, permitir que las decisiones se revisen por personas y proteger al máximo los datos personales. Así se asegura que sirva como apoyo útil al sistema, sin vulnerar derechos ni generar desigualdades.
Sería interesante que esta herramienta pudiera ayudar a cambiar el sistema penal, con el fin de ayudar a presos en libertad, sobre todo a aquellos con riesgo de reincidencia, para ver en cada caso las medidas determinadas que serían precisas proporcionar.
Me parece que cualquier herramienta que pueda ser útil en este ámbito es adecuada, pero al igual que el fondo para estas herramientas viene de nosotros, deberíamos nosotros ser conocedores de toda la información posible de dicha herramienta. Las razones de los ítems, los códigos de programación...
Me parece adecuado, ya que con esta herramienta se está reforzando la seguridad pública, dado que de este modo se evalúan correctamente a los presos, no dejando libres a quienes tienen un alto riesgo de reincidencia.
Sí, me parece bien invertir si realmente mejora el sistema, pero tendría que haber controles claros. Por ejemplo, que se revise cada cierto tiempo por gente independiente, que se expliquen los resultados de forma transparente y que siempre haya una persona que revise las decisiones del algoritmo antes de aplicarlas.
Creo que, al buscar el objetivo de la resocialización de los presos, veo bien que haya cierta inversión para poder lograr ese objetivo siempre y cuando las herramientas que vayan a aplicarse cuenten con las garantías necesarias que demuestren que son eficaces.



Imagen 4: Nube de palabras sobre adecuación en inversión en una herramienta como RisCanvi

7.5. OTROS COMENTARIOS ADICIONALES

Aunque debe destacarse que todas las personas respondieron a todas las preguntas, solo una persona añadió un comentario adicional al final del

cuestionario que representa de forma adecuada la opinión mayoritaria de las personas que respondieron, sin perjuicio de matices interesantes individuales:

> La herramienta presentada me parece estupenda, pero me encuentro con sentimientos encontrados ya que considero que podría llegar a causar más daños que beneficios.

7.6. GRUPO DE DISCUSIÓN

A finales de octubre de 2025 también se realizó un grupo de discusión, alrededor del pódcast mencionado en el cuestionario, con ayuda el profesor de Deusto, Iulen Ibáñez, coautor en esta monografía, en su clase con alumnado de Derecho y Administración de Empresas de la Universidad de Deusto en Donostia-San Sebastián. Según el resumen proporcionado, resultó un grupo muy participativo, con gran debate, donde, de nuevo, se pudieron apreciar dos líneas claras de posicionamiento, sin poder distinguirse claramente una línea hegemónica: unas personas estaban claramente a favor de RisCanvi porque objetiviza la toma de decisiones y otro grupo en contra porque parecía ser más sensible a los riesgos dentro de un contexto penitenciario.

7.7. CONSIDERACIONES GENERALES

Aunque se advierte cierto desconocimiento sobre lo que constituye la inteligencia artificial y una herramienta meramente algorítmica, el conjunto de las respuestas obtenidas evidencia que existe un interés por RisCanvi y una ambivalencia entre la apreciación de las ventajas de la herramienta y los riesgos inherentes propios de su aplicación en el sistema penal. En particular, destaca el ponerse en lugar de las personas presas por parte de los estudiantes de Criminología, quienes fueron la mayoría en responder, mayoritariamente mujeres, como lo son en las aulas de este grado. En relación con estudiantes de Derecho y Administración de Empresas, algunas de las respuestas eran más de corte jurídico, suspicaces frente a la posible vulneración de garantías, aunque otras apreciaban claramente la innovación tecnológica que puede suponer, y el planteamiento de la existencia de RisCanvi y su evolución generó mucho interés y debate.

8. RECAPITULACIÓN: NAVEGANDO ENTRE UN JUICIO CLÍNICO, UN JUICIO PROBABILÍSTICO Y UN SISTEMA MIXTO

> *"(…) la inteligencia humana emerge de una relación práctica y encarnada con el mundo, una habilidad para navegar contextos complejos que los algoritmos no*

> *poseen (…) carecen de la comprensión experiencial y situacional que fundamenta la consciencia y la inteligencia humana (…). A nivel epistemológico, confundimos apariencia con realidad. A nivel social, podríamos legitimar decisiones automatizadas que simulan empatía sin tenerla y ello podría debilitar nuestra comprensión de lo humano al confundir simulación con experiencia. A nivel legal podría incluso conducir a otorgar derechos y responsabilidades a sistemas que no pueden experimentarlos"* (López de Mántaras, 2025).

Como afirman Rodrigues y Fidalgo (2024), la política criminal y penitenciaria de un Estado social de Derecho debe asegurar que las medidas de control impuestas a las personas se equilibren con un uso ético de las tecnologías digitales que contribuya a los fines constitucionales de la reinserción y la seguridad humana. Por ello, las autoras citadas se muestran críticas con la posibilidad de una gobernanza algorítmica que no está basada en un enfoque preventivo clásico, sino de gestión de riesgos a través de la extensión del control preventivo predictivo que contiene un potencial deshumanizador si no se considera la importancia de la evaluación clínica y el contacto con las personas concernidas (Nellis, 2022; Puolakka y Van De Steene, 2021). Para otros autores, la entrada de lo digital también supone una extensión del control opresivo (Jiménez, 2025) porque no permite cuestionar la racionalidad penal de esta primera mitad de siglo, heredera de prácticas anteriores.

Debemos preguntarnos cómo influyen las herramientas algorítmicas y de inteligencia artificial en nuestras decisiones como individuos y como profesionales, así como en el conjunto de la sociedad y, en concreto en el sistema penal (Amoore, 2020; Airoldi, 2022; Fisher, 2022).

Por el momento, sabemos que dichas herramientas no escapan a los prejuicios y sesgos humanos, tan presentes en la justicia penal, como también lo es la falta de información pública, y resultan especialmente atractivos para generar decisiones binarias mediante procesos computacionales que ofrecen cientificidad ante un campo, la criminalidad y la victimización, repleto de incertidumbres que suelen ser irreducibles e, incluso, aconsejables frente a la imposibilidad del control total, pero ello debe ir acompañado de una pedagogía social (Baxter y Dowling, 2025). La filósofa Martha Nussbaum (1986) escribió:

> Ser un buen ser humano es tener una especie de apertura al mundo, una capacidad de confiar en cosas inciertas que escapan a tu propio control, que pueden llevarte a destrozarte en circunstancias muy extremas de las que no tuviste la culpa. Esto dice algo muy importante sobre la condición de la vida ética: que se basa en la confianza en lo incierto y en la voluntad de exponerse; se basa en parecerse más a una planta que a una joya, algo bastante frágil, pero cuya particular belleza es inseparable de esa fragilidad.

Según el criminólogo Daniel Varona (2024), algo que puede relacionarse también con su recordatorio de los principios penales de presunción de inocencia, culpabilidad y proporcionalidad, centrales en los sistemas de justicia penal:

> Cuando las autoridades penitenciarias empezaron a utilizar esta herramienta, ni Jueces de Vigilancia Penitenciaria, ni Fiscalía, ni los abogados, recibieron información detallada sobre el potencial de esta herramienta. La desnuda información sobre si un preso presenta un determinado nivel de riesgo (bajo, medio o alto) sirve de bien poco si el operador no conoce las limitaciones del algoritmo. Así, normalmente bastaba tener un riesgo medio o alto en RisCanvi para que el fiscal impugnase cualquier permiso de salida concedido por la Junta de Tratamiento y el Juez de Vigilancia Penitenciaria. Sin embargo, solo cuando uno es consciente de la alta tasa de falsos positivos que presenta esta herramienta, puede cuestionar su utilización para negar un permiso o una progresión en grado. A pesar de ello, la información sobre el grado de la fiabilidad de la herramienta no acompaña a la propia valoración. Sin una buena formación en las limitaciones de la herramienta, el riesgo de des-responsabilización ("si la máquina dice riesgo alto será por algo") es relevante, máxime en un ámbito en el que están en juego derechos fundamentales (p. xv).

Por todo ello, concluye Daniel Varona (2024):

> ¿y si resulta que es una variable relevante para explicar la delincuencia futura el abandono temprano de la escuela, el nacimiento en un barrio degradado de la ciudad, la falta de empleo o la realización de un trabajo mal remunerado? En estos casos, ¿la herramienta señala a un individuo o a una sociedad que distribuye de forma desigualitaria las oportunidades individuales?
>
> En definitiva, la visión de la delincuencia que en últimas transmite la utilización de estos instrumentos es que estamos ante un problema individual que se podría resolver "científicamente", identificando toda una serie de factores de riesgo. Como criminólogos creo que no debiéramos dejar que estos instrumentos nos hagan olvidar el resto de "variables" e ignorar que estamos ante un problema social complejo que no se va a solucionar buscando la mejor fórmula matemática (p. xv).

Al hilo de la cita introductoria de este epígrafe, Cañas (2025) se pregunta:

> El reto, por tanto, no es técnico, sino profundamente jurídico y ético: ¿seremos capaces de mantener la dignidad del individuo como centro del sistema o acabaremos sometidos a una justicia predictiva y despersonalizada?

¿Serán los algoritmos predictivos en el sistema penitenciario meras opiniones insertadas en matemáticas (O'Neil's, 2016, p. 21), perdiendo la justicia procedimental necesaria respecto de la elaboración y procesamiento de los sistemas y su resultado e impacto (Heimstädt y Egbert, 2025)?

La conclusión de este capítulo es que podemos ser capaces de sistemas predictivos que no pierdan humanidad, siempre y cuando consideremos el contexto específico penitenciario con todas sus particularidades que son previas a los sistemas algorítmicos de predicción del riesgo.

La adaptación de la herramienta RisCanvi al contexto vasco puede suponer una oportunidad de mejora en la gestión del riesgo, si bien su valoración siempre debe recaer finalmente en una decisión humana, apoyada en instrumentos que permitan cumplir con los estándares éticos y normativos. De otra forma, una decisión basada únicamente en un juicio meramente probabilístico puede crear profecías de autocumplimiento[76], en línea con el texto clásico de Lemert sobre desviación secundaria.

La toma de decisiones en el ámbito penitenciario, sobre derechos fundamentales de las personas bajo su control, debe obedecer a criterios no solo técnicos y humanos, sino también interdisciplinares, donde no puede primar solo la Estadística o la Psicología, sino que estas deben conversar con la Criminología, el Derecho, la Sociología, el Trabajo y la Educación social, la Psiquiatría, la Medicina en general, etcétera, y donde cada persona merece una adaptación de criterios a sus particularidades. En esa toma de decisiones conjuntas intervienen aspectos complejos de las culturas profesionales y de las culturas organizacionales donde se adoptan.

La realidad de la aplicación de la normativa garantista de la Unión Europea sobre protección de datos y el uso de algoritmos e IA predictiva no parece que vaya en la dirección de un panóptico invertido en el sentido de una mayor transparencia o gobernanza democrática de la sociedad respecto de las políticas penitenciarias (Cliquennois, 2025), si bien al menos permite un marco de referencia y un control judicial a diferentes niveles.

Por otra parte, aunque se trata de dos temas distintos, conviene advertir la diversidad respecto de las prioridades en la digitalización destinada a la predicción del riesgo y la digitalización destinada a las personas internas y su vida dentro y fuera de prisión (Hofinger y Pflegerl, 2024; Romero-Seseña, Marsol-Gutiérrez y Tamarit-Sumalla, 2025).

76. El teorema de Thomas constituye una noción fundamental en Sociología. Formulado por William I. Thomas en su libro *The child in America: Behavior problems and programs* (1928), dicho teorema se expresa de la siguiente manera: "Si las personas definen las situaciones como reales, éstas son reales en sus consecuencias". Según Merton (1948), la desviación podría ser una de esas profecías autocumplidas si se estigmatiza o etiqueta a las personas.

Por todo ello, la adaptación de RisCanvi al contexto vasco debe enmarcarse en estudios más amplios de legitimidad ético-jurídica y transformación digital y estudiarse también el impacto en la propia cultura y actividad de los operadores jurídicos (funcionarios de diversas disciplinas y funciones, abogados, jueces de vigilancia penitenciaria y otros, fiscales…): cómo transforman estas herramientas su trabajo, sus prácticas y decisiones concretas, pero también las epistemologías penitenciarias, jurídicas y la investigación criminológica sobre las mismas.

Toda teoría o concepto que otorgue cobertura a las decisiones penitenciarias sobre las personas condenadas no deja de ser provisional o incompleto, como lo es el saber científico y técnico.

Un ejemplo de ello es el mismo concepto central de reincidencia (Loinaz, 2022; O'Connell y Laniyonu, 2025) que parece obedecer a binarismos o blancos y negros (¿reincidirá o no reincidirá?). Más allá de su definición jurídica, los estudios criminológicos señalan la necesidad de avanzar hacia conceptos que reflejan mejor la realidad, siempre una escala de grises y siempre interdisciplinar. Por ello, muchos expertos proponen centrarse en el concepto de desistimiento gradual (Maruna, 2025; Cid, 2024; Roldán, 2024) con un enfoque preventivo y ecológico de la comisión de nuevos delitos, campo favorable para una Criminología interdisciplinar por definición.

Este concepto criminológico de desistimiento se relaciona con las cuatro formas de rehabilitación que citamos en otro estudio (Varona et al., 2024), donde se expresaba que existen cuatro formas de intervención rehabilitadora productiva que siempre debemos considerar, y que muy a menudo deben combinarse, si queremos alcanzar el objetivo de la reintegración (véase Burke, Collett y McNeill, 2018; McNeill y Graham, 2020)[77]:

- La rehabilitación personal tiene como objetivo desarrollar una motivación nueva o existente para el cambio, así como crear nuevas habilidades, capacidades y aptitudes para vivir de otra manera.
- La rehabilitación judicial es un proceso de "des-etiquetado" formal y legal en el que el en el que se restablecen el estatus y los derechos del ciudadano. Garantiza el fin del castigo.
- La rehabilitación moral y política es más informal y se centra en el diálogo entre el ciudadano, la sociedad civil y el Estado: una con-

77. Estas citas pertenecen al estudio referenciado de Varona et al. (2024).

versación cívica y civil que no se limita a examinar la infracción, sino lo que hay detrás de ella, y que explora el daño y la reparación.

– La rehabilitación social se refiere a la posición social del individuo y a su identidad social. Se trata de sus conexiones y recursos, su capital social; la ayuda y la acogida que necesitan de otros ciudadanos en el camino hacia la reinserción (McNeill, 2016).

Siguiendo a McNeill (2016), el desistimiento no es un proceso lineal; suele implicar numerosos lapsus y retrocesos, pero supone que las personas, aunque puedan cometer otro delito, están en un camino de prevención de un mayor número de delitos o de delitos más dañinos. Tenemos que encontrar la manera de utilizarlos como oportunidades de aprendizaje, apoyando a las personas en su aprendizaje, con recaídas, para que cumplan las normas, en lugar de apresurarnos a aplicar medidas punitivas. El desistimiento es un proceso de desarrollo personal que cada persona experimenta de manera diferente; los estudios han explorado las diferencias, por ejemplo, relacionadas con el género y el origen étnico, así como las relacionadas con diferentes contextos sociales y culturales. Por tanto, según el autor citado, tenemos que individualizar las formas de apoyo que ofrecemos, respetando la diversidad. El desistimiento se asocia con el desarrollo de la esperanza y un sentido de agencia, o creciente control sobre la dirección de la propia vida. En definitiva, debemos trabajar de forma que se alimente la esperanza y se posibilite la propia agencia, por ejemplo, implicando a las personas en la planificación de su condena, y ayudándoles a desarrollar las capacidades necesarias para dirigir su vida. Las relaciones son fundamentales para el desistimiento; las relaciones sociales y el capital social desempeñan un papel clave, por lo que también debemos trabajar con las familias, los amigos y las comunidades para encontrar juntos formas de apoyar a las personas a través del desistimiento. El desistimiento implica cambios constructivos en las actividades rutinarias de las personas y en sus redes sociales. Esto significa que tenemos que proporcionar apoyo práctico para dichos cambios, por ejemplo, a través de ayudas públicas para cubrir necesidades económicas, vivienda, servicios sanitarios, educación y formación, etc. (Kanestrøm et al., 2025), incluyendo el impacto de haber sido víctimas, en particular en la infancia y adolescencia (Lee, 2025). El reconocimiento de los esfuerzos de las personas por cambiar tiene un efecto de refuerzo. Por el contrario, ciertas actitudes, lenguaje y prácticas de los profesionales de la justicia penal y de las comunidades pueden ser factores de riesgo al socavar el cambio (por ejemplo, reforzando la criminalización y la exclusión). Debemos reconocer y apreciar el cambio, algo que resulta difícil en el día a día y que, quizá, se ha pasado

por alto o no se le ha dado la importancia que merece en la construcción de herramientas predictivas de riesgo en el contexto penitenciario (McNeill, 2016). Lo importante es que el proceso de reinserción y la prevención de futuros delitos no se convierta en una gestión administrativa, sino que se considere lo que realmente es: una experiencia compleja vital (Arrigo, 2024) y relacional (Marder, Kenny y Kurz, 2025).

McNeill también alude a la criminóloga Alison Liebling, de Cambridge, que ha realizado algunos de los trabajos más importantes sobre los factores que limitan y favorecen el desarrollo personal en las prisiones, basado en una cuidadosa y sofisticada combinación de investigaciones etnográficas y encuestas en múltiples lugares de investigación en muchas jurisdicciones de todo el mundo. Liebling (2021) concluye que las "cinco grandes" dimensiones de la calidad penitenciaria que influyen en el desarrollo personal son:

- Legitimidad burocrática: es decir, la transparencia y capacidad de respuesta de la prisión y su reconocimiento moral del individuo.
- Humanidad: es decir, un entorno caracterizado por la consideración amable y la preocupación por la persona.
- Profesionalidad del personal: confianza y competencia del personal en el uso de la autoridad.
- Ayuda y asistencia: significa apoyo y estímulo para abordar problemas (incluidas las drogas y problemas de atención sanitaria) y poder controlarlos y continuar con la vida.
- Organización y coherencia: es decir, claridad, previsibilidad y fiabilidad del régimen penitenciario (McNeill, 2016).

Según el *HM Inspectorate of Probation*, las personas son más propensas a desistir de actividades delictivas cuando cuentan con algunos de los siguientes elementos:

- Fuertes lazos con la familia y la comunidad.
- Un empleo que les satisface.
- El reconocimiento de su valía por parte de los demás.
- Sentimientos de esperanza y autoeficacia.
- Un sentido de significado y propósito en sus vidas.

Paralelamente, se han identificado los siguientes principios para ayudar a facilitar un proceso de desistimiento:

- Respetar la individualidad: dado que el proceso de abandonar la delincuencia es diferente para cada persona, la prestación de servicios debe individualizarse adecuadamente[78].
- Construir relaciones positivas: las personas se ven más influidas para cambiar por aquellas personas cuyos consejos respetan y cuyo apoyo valoran. Las relaciones personales y profesionales son fundamentales para el cambio.
- Reconocer la importancia del contexto social: el abandono está relacionado con los aspectos externos/sociales de la vida de una persona, así como con factores internos/psicológicos.
- Abandonar la delincuencia requiere redes de apoyo y oportunidades en las comunidades locales
- Reconocer y desarrollar las fortalezas de las personas: promover una serie de factores de protección y un enfoque basado en las fortalezas debe formar parte del proceso de supervisión. Por ejemplo, unas relaciones familiares y íntimas sólidas y de apoyo pueden ayudar a las personas en su camino hacia el abandono de comportamientos delictivos.

A pesar de las dificultades en la evaluación del impacto de los programas y su variabilidad, a nivel internacional, se concluye que los programas sólidos que previenen la reincidencia van más allá de la educación y la formación (en prisión) y ofrecen asistencia tras la puesta en libertad para buscar trabajo y conseguir vivienda. Los programas adaptados a grupos específicos suelen ser muy eficaces, en particular los programas de tratamiento del abuso de sustancias (incluyendo enfoques de justicia terapéutica), enfocados en la salud mental y los programas específicos para delitos sexuales son ejemplos de ello porque, al mismo tiempo, promueven la confianza y la motivación (European Crime Prevention Network, 2024; O'Meara Daly et al., 2025). Las altas tasas de reincidencia se citan a menudo como prueba de las deficiencias actuales de las prácticas de encarcelamiento. Varios estudios

78. El grupo de expertos de la Oficina de las Naciones Unidas sobre las Drogas y el Delito (UNDOC), en relación con las estrategias para reducir la reincidencia, concluyó en enero de 2025, con vistas al Congreso de 2026, con la necesidad de promover la individualización y las alternativas, sin referirse, sorprendentemente, a sistemas automatizados de gestión y valoración del riesgo.

han demostrado que las penas de prisión, por sí solas, no son eficaces para reducir la delincuencia y prevenir la reincidencia, o al menos no son más eficaces que las alternativas no privativas de libertad. Más importante aún, los críticos sostienen que los enfoques punitivos, como las penas de prisión (largas), no abordan los problemas subyacentes y pueden incluso agravar las tendencias delictivas al aislar a las personas y exponerlas a entornos criminógenos.

La prisionización se refiere al efecto negativo que las penas de prisión pueden tener en la salud mental y el comportamiento de los reclusos, incluida la criminalidad. Las propias condiciones de la vida en prisión pueden tener un impacto negativo en la salud mental de los reclusos y, por ende, en la reincidencia. También es probable que los reclusos pierdan sus redes de apoyo social al mismo tiempo que se forjan nuevos vínculos con otros delincuentes. Las oportunidades tras la puesta en libertad, especialmente las oportunidades de empleo, se ven reducidas por el mero hecho de haber estado encarcelado.

En este sentido, el concepto de desistimiento puede ayudar a repensar las herramientas algorítmicas de predicción, con un enfoque integral, también en relación con procesos de reentrada o reintegración en la sociedad. Este enfoque incluye realizar un seguimiento de las medidas de empleo, vivienda, salud e integración social junto con la progresión de la reincidencia y, según expertos europeos, ofrece una visión más completa del éxito de la reinserción y ayuda a destinar los recursos donde más se necesitan (Willison, 2023). Cabe subrayar el enfoque de ciudades que apuestan por la recuperación de los individuos que salen de prisión, con una previsión de apoyo a más de cinco años (Best, Duffy y Colman, 2024). En relación con ello, también pueden citarse las conclusiones del Consejo de Ministros de Justicia y Seguridad de la Unión Europea, de junio de 2024, apoyando el movimiento de reescalamiento de las prisiones[79]. Este movimiento impulsa, para aquellas personas que sean adecuadas para ellas, prisiones pequeñas en la ciudad o casas de detención, integradas en la comunidad, lo cual beneficia no solo a las personas encarceladas, sino también a su red social, a las víctimas potenciales, al sistema de justicia penal, al personal de este sistema y a la sociedad en general. Según los Ministros de la Unión Europea, estas casas de detención o prisiones más abiertas en la ciudad[80] permiten

79. Véase el documento en https://data.consilium.europa.eu/doc/document/ST-10105-2024-INIT/en/pdf

80. En este sentido, no se identifican exactamente con los Centros de Inserción Social españoles.

una mejor aplicación de las leyes y los principios europeos en este sector y contribuyen a la rehabilitación, a la reducción de la reincidencia y a la sostenibilidad social.

La pregunta es cómo podemos integrar todo este conocimiento y recomendaciones sobre lo que funciona (Mourão et al., 2025) de cara a la prevención de nuevos delitos en un sistema algorítmico dinámico de predicción del riesgo que funcione como una herramienta para un juicio clínico grupal, ajustado al Derecho y a la Ética, en tiempos renovados de criminalización y punitivismo.

9. REFERENCIAS

Airoldi, M. (2022). *Machine Habitus: Toward a Sociology of Algorithms*. John Wiley & Sons.

Alemán Aróstegui, L. (2023). El uso del RisCanvi en la toma de decisiones penitenciarias. *Estudios Penales y Criminológicos*, 44. https:https://doi.org/10.15304/epc.44.8884.

Alonso García, M. C. (2025). Sistema VioGén: fallos y algunas propuestas de mejora. En *El Derecho Administrativo en la era de la inteligencia artificial: Actas del XVIII Congreso de la Asociación Española de Profesores de Derecho Administrativo. Vigo 25 a 27 de enero de 2024* (pp. 119-129). Instituto Nacional de Administración Pública (INAP).

Amoore, L. (2020). *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Duke University Press.

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior*, 17, 19-52.

Arrigo, B. A. (2024). Practice framework theorizing in correctional rehabilitation: Lessons from constitutive penology. *Legal and Criminological Psychology*, 29(2), 132-146.

Baratta, A. (1982). Observaciones sobre las funciones de la cárcel en la producción de las relaciones sociales de desigualdad. *Nuevo Foro Penal*, 15, 737.

Barrio, M. (2025). Los «lobbies" tecnológicos quieren frenar la aplicación de la ley europea de inteligencia artificial. *The Conversation*, 8 de julio.

Bataller, M. (2025). Riscanvi en el sistema penitenciario catalán, https://togas.biz/directorio/profesionales-bataller-abogados.

Baxter, K. & Dowling, S. (2025). Breaking down barriers to working restoratively in prison: The problem with ReHub. https://why-me.org/2025/rehub-breaking-down-barriers-to-working-restoratively-in-prison/.

Beaudry, G., Yu, R., Miller, O., Prescott-Mayling, L., Fanshawe, T. R., & Fazel, S. (2023). Prediction of violent reoffending in people released from prison in England: external validation study of a risk assessment tool (OxRec). *Journal of Criminal Justice*, 86, 102061.

Bosco, F., Creemers, N., Ferraris, V., Guagnin, D., & Koops, B. J. (2014). Profiling technologies and fundamental rights and values: regulatory challenges and perspectives from European Data Protection Authorities. *Reforming European Data Protection Law*, 3-33.

Braun, V. & Clarke, V. (2022). *Thematic Analysis: A Practical Guide*. Sage.

Cañas Pérez, P. A. (2025). IA en prisiones: ¿riesgo para los derechos fundamentales o herramienta de reinserción?, https://www.abogacia.es/publicaciones/blogs/blog-derecho-penitenciario/ia-en-prisiones-riesgo-para-los-derechos-fundamentales-o-herramienta-de-reinsercion/.

Capdeferro Villagrasa, O. (2018). El análisis de riesgos como mecanismo central de un sistema efectivo de prevención de la corrupción. En particular, el sistema de alertas para la prevención de la corrupción basado en inteligencia artificial. *Revista Internacional de Transparencia e Integridad*, 5, 1-6.

Cardoso, T. (2020). *Bias behind bars: A Globe investigation finds a prison system stacked against Black and Indigenous inmates*, https://www.theglobeandmail. com/canada/article-investigation-racial-bias-incanadian-prison-risk-assessments/.

Castro Liñares, D. (2021). *Los instrumentos de valoración y gestión de riesgos en el modelo de penalidad español*. Reus.

Castro-Toledo, F. J., Miró-Llinares, F. & Aguerri, J. C. (2023). Data-Driven Criminal Justice in the age of algorithms: epistemic challenges and practical implications. *Criminal Law Forum*, 34(3), 295-316.

Chelioudakis, E. (2020). Risk assessment tools in criminal justice: Is there a need for such tools in Europe and would their use comply with European data protection law? ANU Journal of Law and Technology, 1(2), 72-96.

Cid Moliné, J. (2024). ¿Son las teorías del desistimiento y sus implicaciones prácticas dependientes del contexto estructural? *Boletín Criminológico*, 30(228), 1-8.

Cid Moliné, J. & de la Encarnación Ordóñez, E. (2024). *Tercer grado y reinserción: ¿Cómo logra el tercer grado ser un punto de inflexión en la vida de las personas excarceladas?* CEJFE, Repositori del Departament de Justícia.

Clemente Díaz, M., Díaz Huamán, S., Auñón García, L. P., Baquedano Olave, S., Blanco Coello, R., Cabra Vázquez, E. et al. (1993). *Validación y depuración de la Tabla de Variables de Riesgo en el disfrute de permisos de salida*. Dirección General de Instituciones Penitenciarias.

Cliquennois, G. (2025). The Reinforced European Effects on Penal and Prison Policies: The Back-up of the EU and the CJEU. En *The Impact of European Human Rights Bodies on Penal Policies. Human Rights Interventions*. Palgrave Macmillan.

Cotino Hueso, L. (2025). ¿Cuándo "no es no"?: Criterios para definir los sistemas de inteligencia artificial prohibidos en la Unión Europea. *Revista General de Derecho Administrativo*, 69, 9.

Cueto Santa Eugenia, E. (2025). Herramientas de predicción criminal y su impacto en nuestro proceso penal. *Ius et Scientia*, 11(1), 14-31.

Departament de Justícia i Qualitat Democràtica (Generalitat de Catalunya). (2024). *Protocol de valoració del risc (RisCanvi),* https://justicia.gencat.cat/ca/ambits/reinsercio_i_serveis_penitenciaris/serveis_penitenciaris/RisCanvi/.

Díez Ripollés, J. L. & García España, E. (2020). RIMES: Un instrumento de comparación de políticas criminales nacionales desde la exclusión social. *Política Criminal*, 15(30), 670-693.

Digital Future Society. (2021). Gobernanza y algoritmos. https://digitalfuturesociety.com/es/report/governingalgorithms/.

Digital Future Society. (2022a). Capítulo 1. RisCanvi (I): el algoritmo de la cárcel. [pódcast]. https://digitalfuturesociety.com/es/podcasts/capitulo-1-riscanvi-i-el-algoritmo-dela-carcel/.

Digital Future Society. (2022b). Capítulo 2. RisCanvi (II): ¿Se puede predecir el próximo delito? [pódcast]. https://digitalfuturesociety.com/es/podcasts/capitulo-2-riscanvi-ii-se-puede-predecir-el-proximo-delito/.

Digital Future Society. (2023). *El uso de algoritmos en el sector público en España: cuatro estudios de caso sobre ADMS (Sistemas Automatizados de Toma de Decisiones)*. Digital Future Society.

Dribia Data & Direcció General d'Afers Penitenciaris. (2024). *Informe Tiresias: Auditoria de l'algorisme RisCanvi*. Generalitat de Catalunya, Departament de Justícia, Drets i Memòria.

Ebers, M. (2024). Truly risk-based regulation of artificial intelligence how to implement the EU's AI Act. *European Journal of Risk Regulation*, 1-20.

Esteve, J. S., & Domènech, C. B. (2024). Algunos riesgos para el estatus de las personas en la utilización de la inteligencia artificial: Aproximación a los avances de la Ley 15/2022. I*gualdadES* (10), 91-121.

European Commission (2018). *Coordinated Plan on Artificial Intelligence, Communication from the commission to the European Parliament, The European Council, The Council, The European Economic and social committee and the Committee of the Regions*. EC, COM (2018) 795 final.

European Crime Prevention Network. (2024). *Toolbox on Reducing Recidivism*. EUCPN.

European Institute for Gender Equality. (s.f.). *Risk Assessment and Risk Management by Police*. EIGE.

Europol. (2025). *AI Bias in Law Enforcement. A Practical Guide*. Publications Office of the European Union. https://www.europol.europa.eu/cms/sites/default/files/documents/AI_bias_in_law_enforcement_-_practical_guide.pdf

Fanega, M., Fresneda Portillo, C. & Beltrán Camacho, Á. (2025). Disclosing the TVR prison algorithm. Revista Española de Investigación Criminológica. *Sociedad Española de Investigacion Criminologica*. http:https://doi.org/10.46381/reic.v22i1.917

Fisher, E. (2022). *Algorithms and Subjectivity: The Subversion of Critical Knowledge*. Routledge.

Fry, H. (2018). *Hello World — Being human in the Age of Algorithms*. W.W. Norton & Company Ltd.

García, T. (2025). Las prisiones españolas usan un algoritmo sin actualizar desde 1993 para decidir sobre permisos de salida, https://civio.es/justicia/2025/02/26/las-prisiones-espanolas-usan-un-algoritmo-sin-ac-

tualizar-desde-1993-para-decidir-sobre-permisos-desalida/#:~:text=M%-C3%A1s%20de%20200%20resoluciones%20de,ning%C3%BAn%20factor%20de%20riesgo%20m%C3%A1s

Gimeno Beviá, J. (2023). Instrumentos actuales de policía y justicia predictiva en el proceso penal español: Análisis crítico y reflexiones *de lege ferenda* ante aplicaciones futuras. *Estudios Penales y Criminológicos*, 44, 1-20.

Gordo, Á., & Rubio-Martín, M. J. (2024). Incertidumbres algorítmicas en torno a las violencias de género. El caso del sistema VioGén y otros sistemas de predicción del riesgo. *Revista Española de Sociología*, 33(2), https:https://doi.org/10.22325/fes/res.2024.225.

Heilbrun, K., Dvoskin, J., Hart, S. & McNiel, D. (1999). *Violence risk communication: Implications for research, policy, and practice*. https://www.tandfonline.com/

Heimstädt, M. & Egbert, S. (2025). Actionable predictions: How designers of algorithmic systems calibrate criminal futures. *Big Data & Society*, 12(2), 20539517251340636.

Helmus, L. & Bourgon, G. (2011). *Taking Stock of 15 Years of Research on the Spousal Assault Risk Assessment Guide (SARA): A Critical Review*. https://www.tandfonline. com/doi/abs/https://doi.org/10.1080/14999013.2010.551709.

Henrich, S. (2025). Assessing Extremist Violence: What the Evidence Tells Us About Risk, Identity, and the Wider Ecosystem. *European Commission Newsroom*, 31.07.2025.

Hofinger, V., & Pflegerl, P. (2024). A reality check on the digitalisation of prisons: Assessing the opportunities and risks of providing digital technologies for prisoners. *Punishment & Society*, 26(5), 898-916.

Houser K. (2019, February 4). China is Installing «AI Guards" in Prison Cells. They'll make escape impossible — but the trade-*off* might be inmates" mental health. *Futurism.* https://futurism.com/chinese-prison-ai-guards-cells

Jiménez, A. (2025). *The Crimes of Digital Capitalism. Corporate Crime in an Age of Exploitation*. New York University Press.

Kanestrøm, H., Hernæs, Ø. M., Stallvik, M., Lydersen, S., Skokauskas, N., & Kaasbøll, J. (2025). Do Criminogenic Needs Matter in Non-Secure Settings? Assessing Change in Dynamic Risk Factors During Therapeutic

Residential Care and the Association With Prediction and Recidivism. *Crime & Delinquency*, 00111287251330776.

Larrauri, E. (2019). Se inventan sus leyes: ¿Qué criterios se deben valorar en la concesión de permisos de salida penitenciarios. *Jueces para la Democracia*, 94, 43-58.

Larsen, B. Ø., Larsen, K. S. & Berger, N. P. (2025). High-or low-risk prisoner, high-or low-risk ex-offender? Assessing the predictive validity of the level of service/risk-need-responsivity (LS/RNR) instrument among different subgroups of Danish prisoners. *European Journal of Criminology*, 22(2), 204-229.

Lee, Y., Bradford, B., & Posch, K. (2024). The effectiveness of big data-driven predictive policing: Systematic review. *Justice Evaluation Journal*, 7(2), 127-160.

Lee, Y. (2025). Continuity in Criminal Behavior: The Interplay of Early-Onset Delinquency and Victimization. *Crime & Delinquency*, 00111287251331572.

Lemert, E. M. (1972). *Human Deviance, Social Problems and Social Control*. Englewood Cliffs.

León, C. M. & Bartolomé, R. (2024). Valoración del riesgo de reincidencia en jóvenes infractoras:¿ Es el SAVRY un instrumento realmente neutro? *Boletín Criminológico*, 30(236), 1-26.

Libano Beristain, A. (2025). Viogén como herramienta algorítmica de protección frente a la violencia de género: análisis de impacto desde la perspectiva procesal penal. En *Conectadas y desiguales: la brecha de género en la era digital* (pp. 321-344). Reus.

Liebling, A. (2021). What is philosophy in prison? George Eliot and the search for moral insight. *Journal of Prison Education and Reentry*, 7(2), 104-114.

Loinaz, I. (2022). Inteligencia artificial y predicción del riesgo de reincidencia. En *La transformación algorítmica del sistema de justicia penal* (pp. 101-112). Thomson Reuters Aranzadi.

López de Mántaras, R. (2025). La trampa de los grandes modelos de lenguaje: ver conciencia donde solo hay palabras. *The Conversation*, 27 de julio, https://theconversation.com/la-trampa-de-los-grandes-modelos-de-lenguaje-ver-conciencia-donde-solo-hay-palabras-261278.

Lussier, P., Kouassi, P. L., & Frechette, J. (2025). Agency, Criminogenic Risk and Needs, and Recidivism: A Prospective Longitudinal Study Including 14,000 Adult Justice-involved Individuals. *International Journal of Offender Therapy and Comparative Criminology*, 69(13-14), 1978-2006.

Marder, I. D., Kenny, T. & Kurz, K. (2025). Restorative practices and relational prison leadership in Ireland. *Contemporary Justice Review*, 1-23.

Martín Ríos, P. (2024). Predictive algorithms and criminal justice: expectations, challenges and a particular view of the Spanish VioGén system. *Rivista Italiana di Informatica e Diritto*, 2, 547-562.

Martínez Garay, L. (2016). Errores conceptuales en la estimación de riesgo de reincidencia. *Revista Española de Investigación Criminológica*,14, 1-31.

Martínez Garay, L., Boix Palop, A., Briz Redón, A., Flores Giménez, F., García Ortiz, A., Molina Sánchez, M. et al. (2024). *Three predictive policing approaches in Spain: VioGén, RisCanvi and VeriPol. Assessment from a human rights perspective*. Universidad de Valencia. https://regulation.blogs.uv.es/files/2024/09/Tres-sistemas-de-policia-predictiva-26.12.24.pdf

Maruna, S. (2025). Redeeming desistance: From individual journeys to a social movement. *Criminology*, 63(1), 5-25.

McGoogan, C. (2016, December 6). Liverpool prison is using AI to *stop* smuggling drugs and weapons.*TheTelegraph*. https://www.telegraph.co.uk/technology/2016/12/06/liverpool-prison-using-ai-*stop*-drugs-weapons-smuggling/

McNeill F (2016) Desistance and criminal justice in Scotland. En Groall, H., Mooney, G., and Munro, R. eds. *Crime, justice and society in Scotland* (pp. 200-216). Routledge.

Merton, R. K. (1948). The self-fulfilling prophecy. *The Antioch Review*, 8(2), 193-210.

Ministerio del Interior. (2021). *Construcción y validación de una herramienta de clasificación y de valoración del riesgo de radicalismo violento en el ámbito penitenciario*. Ministerio del Interior.

Ministerio del Interior. (2022). *Estudio de reincidencia penitenciaria 2009-2019. Documentos Penitenciarios 30*. Ministerio del Interior.

Montoro Sánchez, J. A. (2023). La Directiva 2016/680/UE: un nuevo paradigma para el tratamiento de datos de carácter personal con fines penales.

En *El proceso penal ante una nueva realidad tecnológica europea* (pp. 103-130). Thomson Reuters Aranzadi.

Moore, R. and National Offender Management Service (July 2015), A compendium of research and analysis on the Offender Assessment System (OASys) 2009-2013. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/449357/research-analysis-offender-assessment-system. pdf (accessed 07/03/2024).

Mourão, A., Sousa, M., Ferreira, M., Gonçalves, L., Caridade, S. & Cunha, O. (2025). Beyond Recidivism: A Systematic Review Exploring Comprehensive Criteria for Successful Reintegration After Prison Release. *Criminal Justice and Behavior*, 00938548251335322.

Nellis, M. (2022). From electronic monitoring to artificial intelligence: technopopulism and the future of probation services. En *Reimagining Probation Practice* (pp. 207-224). Routledge.

Nellis, M. (2024). Artificial Intelligence and Correctional Practice: The Council of Europe Draft Recommendation on the Use of AI in Prisons and Probation. *Journal of Offender Monitoring*, 37(1).

Nellis, M. (2025). Artificial Intelligence and Probation Services, https://www.cep-probation.org/artificial-intelligence-ai-and-probation-services-by-mike-nellis/.

Newcomb, A. (2019, March 28). Finland Is Using Inmates to Help a Start-Up Train Its Artificial Intelligence Algorithms. Fortune. http://fortune.com/2019/03/28/finland-prison-inmates-train-ai-artificial-intelligence-algorithms-vainu/

Nussbaum, M. (1986). *The Fragility of Goodness. Luck and Ethics in Greek Tragedy and Philosophy*. Cambridge University Press.

O'Meara Daly, E., Dwane, J., Lewis, C. & Redmond, S. (2025). *An evidence-informed model and guide for effective relational working in youth justice*. HM Inspectorate of Probation.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Observatorio Sector Público e Inteligencia Artificial, Diez notas sobre la AI Action Summit de París, https://www.ospia.org/o-lab/diez-notas-sobre-la-ai-action-summit-de-paris

O'Connell, S. & Laniyonu, A. (2025). Race, gender, and risk assessments in Canadian federal prison. *Race and Justice*, 15(3), 428-453.

Pedro, R. (2025). Algoritmos digitais: uso público, transparência e litigância. En *Digital Transformation and Governance in the Judiciary* (pp. 149-167). Faculdade de Direito.

Pedrosa, A. (2025). Las Juntas de Tratamiento en prisión. ¿Cómo toman sus decisiones? *InDret*, 4, 314-346.

Pereira Puigvert, S. (2024). La justicia y el proceso penal en clave algorítmica: Nuevos enfoques, nuevos riesgos. *Ius et Scientia*, 10(2), 66-79. https:https://doi.org/10.12795/IESTSCIENTIA.2024.i02.03.

Pereira, A. (2020). Artificial Intelligence, Offender Rehabilitation & Restorative Justice. The "Good" Algorithm? Artificial Intelligence: Ethics, Law, Health. International Workshop organized by the Pontificia Academia Pro Vita. Date: 2020/02/26 — 2020/02/28. New Hall of the Synod, Vatican City. https://limo.libis.be/primo-explore/fulldisplay?docid=LIRIAS2960856&context=L&*vid*=Lirias&search_scope=Lirias&tab=default_tab&lang=en_US&fromSitemap=1.

Pérez Salazar, B. (2024). La criminología predictiva: ¿un futuro próximo o una ficción en lontananza? *Novum Jus*, 18 (3), 343-369. https:https://doi.org/10.14718/NovumJus.2024.18.3.13.

Perlin, M., Ellis Cucolo, H. y Kulesza, J. (2025). Artificial intelligence, algorithmic bias, and mental disability law: Ethical challenges and regulatory solutions, http://dx.doi.org/10.2139/ssrn.5979394.

Pires, A.R., Fernandes, A., Estalella, G., Zisiadou, M., Carrolaggi, P., Loja, S., & Leitão, T. (2021). The Potential for Virtual Reality for Education and Training in Prisons. https://www.researchgate.net/publication/357752509_THE_POTENTIAL_OF_VIRTUAL_REALITY_FOR_EDUCATION_AND_TRAINING_IN_PRISONS.

Portela, M., Castillo, C., Tolan, S., Karimi Haghighi M., & Andrés Pueyo, A. (2022). A Comparative User Study of Human Predictions in Algorithm— Supported Recidivism Risk Assessment. https://arxiv.org/pdf/2201.11080v2.pdf.

Puolakka P. (2020). RISE *AI: Reducing the Risk of Recidivism with AI*. Aalto Executive Education: Diploma in Artificial Intelligence. Unpublished.

Puolakka, P. & Van De Steene, S. (2021). Artificial intelligence in prisons in 2030: An exploration on the future of AI in prisons. *Advancing Corrections Journal*, 11, 128-138.

Puolakka, P., & Van De Steene, S. (2021). Artificial Intelligence in Prisons in 2030. An exploration on the future of Artificial Intelligence in Prisons. *Advancing Corrections Journal*, Edition # 11, ICPA.

Raynor, P. (2019). Development, critics and a realist approach. In Ugwudike, P., Graham, H., McNeill, F., Raynor, P., Taxman, F. S. y Trotter, C. (Eds.). *The Routledge Companion to Rehabilitative Work in Criminal Justice*. ProQuest Ebook Central.

Rodrigues, A. M. & Fidalgo, S. (2024). The role of Artificial Intelligence (AI) in rehabilitation and in the reduction of the use of imprisonment. *UNIO-EU Law Journal*, 10(1), 42-53.

Roldán, N. (2024). Tener un propósito: ¿por qué desistimos de seguir delinquiendo? *Cuestiones Criminales*, 7(13/14).

Romero-Seseña, P., Marsol-Gutiérrez, A. & Tamarit-Sumalla, J. M. (2025). The use of online tools and Internet access in prisons: insights from Catalonia. *Current Issues in Criminal Justice*, 1-21.

Rovira, M., Larrauri, E., & Alarcón, P. (2018). La concesión de permisos penitenciarios. *Revista Electrónica de Ciencia Penal y Criminología*, 20(02), 1-26.

Sachoulidou, A. (2023). Going beyond the "common suspects": to be presumed innocent in the era of algorithms, big data and artificial intelligence. *Artificial Intelligence and Law*, 1-54.

Simón Castellano, P. (2022). *La prisión algorítmica. Prevención, reinserción social y tutela de derechos fundamentales en el paradigma de los centros penitenciarios inteligentes*. Tirant lo Blanch.

Solar Calvo, M. del P. & Lacal Cuenca, P. (2025). Valoración del riesgo, uso de algoritmo y falsos positivos. El peligro de pasar de lo simple a lo complejo. *Diario La Ley*, Núm. 10669.

Sunstein, C.R. (2019, January 23). Algorithms, Correcting Biases. Oxford Business Law Blog. https://www.law.ox.ac.uk/business-law-blog/blog/2019/01/algorithms-correcting-biases

Tahirí Moreno, J. A. (2024). Una panorámica de los sistemas de inteligencia artificial desde la perspectiva del éticadministrativo. *Revista Aragonesa de Administración Pública* (61), 137-168.

Teijón Alcalá, M. (2022). La evaluación del riesgo de radicalismo violento en prisión: problemas de cientificidad y teorización en el instrumento dravy. En *Ejecución de las penas por delitos de terrorismo* (pp. 367-402). Dykinson.

Teng, M.Q. & Gordon, E. (2021). Therapeutic virtual reality in prison: Participatory design with incarcerated women. *New Media & Society*, 23(8), 2210-2229.

Törnberg, P. & Uitermark, J. (2025). The Ascendance of Algorithmic Tyranny, https://www.noemamag.com/the-ascendance-of-algorithmic-tyranny/?utm_source=tekneaudience&utm_medium=email&utm_campaign=711_lab_08_2025&utm_content=.

Transparencia e Integridad. http://www.encuentros-multidisciplinares.org/ revista-65/oscar_capdeferro-1.pdf.

Trickett, L. & Bryan, T. (2025). Aligning police practice with hate crime theory: The case for using risk assessments to improve police response to victims of hate. *International Review of Victimology*, 31(1), 22-38.

UNICRI. (2024). *Digital Rehabilitation in Prisons*. UNICRI.

Varona, D. (2024). Algoritmos e inteligencia artificial en el sistema de justicia penal. *InDret*, 4.

Varona, G. et al. (2024). *Confluencias entre procesos de reinserción y restauración en el ámbito penitenciario: provocar el nunca más*. UPV/EHU.

Wang, N., & Tian, M. Y. (2022), "Intelligent Justice": human-centered consideration in China's legal AI transformation. *Ai and Ethics*, 3(2): 349-354. https:https://doi.org/10.1007/s43681-022-00202-3.

Willison, J. B. (2023). *Measuring Reentry Success Beyond Recidivism*. Washington,DC: U.S. Department of Justice, Bureau of Justice Assistance.

Yampolskiy, R. V. (2025). On monitorability of AI. *AI and Ethics*, 5(1), 689-707.

Završnik, A. (2020). Criminal Justice, artificial intelligence systems, and human rights. *Academy of European Law — ERA Forum*, 20, 567-583.

Zempi, I. ,Trickett, L., Krulisova, K. & Hamilton, P. (2025). Community conversations as a strategy to prevent hate crime: facilitators" reflections. *International Review of Victimology*. https:https://doi.org/10.1177/02697580251340991.

Zeng, J., Ustun, B., & Rudin, C. (2017). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3), 689-722. https:https://doi.org/10.1111/rssa.12227.

Zottola, S. A., Stewart, K., Cloud, V., Hassett, L. & Desmarais, S. L. (2024). Predictive bias in pretrial riskassessment: Application of the Public Safety Assessment in a Native American population. *Law and Human Behavior*, 48(5-6), 398-414. https:https://doi.org/10.1037/lhb0000562.

Zouave, E. T. & Marquenie, T. (2017). An inconvenient truth: Algorithmic transparency & accountability in criminal intelligence profiling. *European Intelligence and Security Informatics Conference* (EISIC), 17-23. https:https://doi.org/10.1109/EISIC.2017.12.

Capítulo 8

Las dos caras de los algoritmos de estimación de riesgo en el ámbito penitenciario: recomendaciones para conjugar prevención y garantías[81]

Lucía Martínez Garay

Departamento de Derecho penal. Universitat de València

SUMARIO: 1. INTRODUCCIÓN. 2. LAS DOS CARAS DE RISCANVI. 3. LA PRIMACÍA DEL ASPECTO TRATAMENTAL EN EL DISEÑO, IMPLEMENTACIÓN Y SUPERVISIÓN DE RISCANVI. 4. PARADOJAS QUE PRODUCE LA TENSIÓN ENTRE LA DIMENSIÓN TRATAMENTAL/PREVENTIVA, Y LA DE GARANTÍAS. 4.1. *¿Puede la capacidad predictiva de RisCanvi ser buena y mala a la vez? ¿Puede afirmarse al mismo tiempo que RisCanvi minimiza los falsos positivos, y que produce demasiados?* 4.2. *La capacidad predictiva es mejor o peor según cuál sea el objetivo que perseguimos. ¿Cuál es el objetivo y quién debe establecerlo?* 4.3. *¿Produce RisCanvi efectos discriminatorios prohibidos, o no los produce?* 4.4. *¿Es RisCanvi una herramienta que mejora la transparencia de las decisiones, o un algoritmo opaco que vulnera el derecho de acceso a los datos personales y el derecho de defensa?* 5. CONCLUSIONES GENERALES. 6. RECOMENDACIONES CONCRETAS SI SE QUIERE IMPLANTAR ALGUNA HERRAMIENTA PARECIDA A RISCANVI EN EL

81. Este trabajo se ha realizado en el marco del Proyecto de I+D+i PID2021-123441NB-I00, financiado por MCIN/ AEI/10.13039/501100011033/ y "FEDER Una manera de hacer Europa", y del convenio C036/23-OT entre la entidad pública empresarial Red.es y una agrupación de entidades (Universidad del País Vasco/Euskal Herriko Unibersitatea (EHU), Universitat de València, Fundación Instituto de Investigación Sanitaria "Marqués de Valdecilla", Entidad de Gestión de Derechos de los Productores Audiovisuales LA (EGEDA), Universitat Rovira i Virgili, y Fundación Hay Derecho), para impulsar la implementación de la Carta de Derechos Digitales en el ámbito de entornos específicos.

SISTEMA PENITENCIARIO VASCO. 6.1. *Mantener la propiedad y el control sobre los modelos y sobre los datos*. 6.2. *Necesidad de evaluación previa, continua, y por resultados agregados*. 6.3. *Transparencia*. 6.4. *Conflicto con protección de derechos.* 7. REFERENCIAS.

1. INTRODUCCIÓN

Una vez que desde 2021 ha asumido la competencia en materia penitenciaria, el Gobierno Vasco está considerando la posibilidad de introducir en la gestión de sus prisiones un sistema similar al RisCanvi, que la Generalitat de Catalunya tiene en marcha desde hace ya varios años, para valorar de manera estandarizada y estructurada el riesgo de aparición de diferentes conductas, entre ellas la reincidencia. Como parte de esta fase de reflexión previa, el Gobierno Vasco ha invitado a especialistas en diversas materias a debatir los pros y los contras de esa eventual implementación antes de ponerla en marcha, cosa que a mi juicio es en sí misma ya un hecho muy digno de alabanza.

Son muchas las cuestiones sobre las que vale la pena reflexionar en este proceso de (eventual) implantación de un sistema parecido a RisCanvi en la gestión de las prisiones vascas. De entre todas ellas voy a abordar en este trabajo una que me parece especialmente importante, porque creo que afecta de manera transversal a muchas otras.

2. LAS DOS CARAS DE RISCANVI

RisCanvi es una herramienta que pretende estimar y cuantificar, de manera estructurada y estandarizada, el riesgo de que las personas internas en los centros penitenciarios lleven a cabo conductas delictivas (reincidencia, ya sea violenta o no, y quebrantamiento de permisos) o problemáticas (violencia autodirigida, conducta violenta dentro del centro). Además de estimar la probabilidad de que ocurran estos eventos, RisCanvi pretende también servir para gestionar esos riesgos, desde dos perspectivas diferentes: por un lado, la información que RisCanvi recoge de manera sistemática, y las estimaciones que riesgo que realiza, deben ser la base sobre la cual se articule la intervención penitenciaria, es decir, el plan de tratamiento y las actividades que se propongan a cada interno, con el fin de favorecer que puedan en el futuro vivir sin delinquir; por otro lado, esos riesgos, y en especial los de reincidencia y de quebrantamiento, deben tenerse en cuenta cuando se toman decisiones que afectan a la libertad de los internos, para minimizar la reincidencia de las personas que salen de prisión. Es decir,

RisCanvi, y cualquier otro sistema similar de gestión de riesgo en el ámbito penitenciario, es un instrumento que se utiliza tanto para la resocialización como para el control, una herramienta que refleja la dicotomía entre esas dos dimensiones que caracteriza al sistema penitenciario español, el cual ha de servir a la custodia eficaz de los presos, y con ello garantizar el cumplimiento de las penas y la seguridad pública, y a la vez favorecer la reinserción y resocialización que el art. 25.2 de la Constitución Española establece como fines de la ejecución de las penas privativas de libertad y de las medidas de seguridad.

En el caso de RisCanvi, esta doble finalidad es especialmente evidente y se ha reconocido de manera expresa en muchas ocasiones, explicando que su origen está ligado a la preocupación por controlar la reincidencia de los delincuentes peligrosos tras la excarcelación, especialmente delincuentes sexuales, tal como indicaron las conclusiones de la llamada "Comisión Mena"[82]. Junto a ello, además, se pretendió implantar un sistema de gestión de los centros penitenciarios basado en el modelo RNR de Andrews y Bonta, la evidencia empírica y las buenas prácticas (Subdirecció General de Programes de Rehabilitació i Sanitat, 2019, pp. 4 y 5).

Puesto que tiene esta doble faz, RisCanvi y cualquier otro sistema similar deben implantarse, gestionarse y supervisarse, teniendo en cuenta ambas dimensiones: por un lado, la intervención resocializadora, que busca la prevención de la reincidencia y del resto de riesgos que pretende gestionar; y, por otro lado, la dimensión jurídica, que debe asegurar que esa prevención se logra respetando los derechos fundamentales de los internos y salvaguardando el resto de garantías que deben rodear la ejecución de las penas. Son dos dimensiones diferentes, que en ocasiones pueden ser compatibles e incluso reforzarse mutuamente, pero que otras veces están enfrentadas. Conviene insistir en esto: el hecho de que una herramienta esté diseñada para favorecer la reinserción social (como sin duda creo que lo está RisCanvi, si se atiende a las publicaciones e informes de sus creadores y de los responsables de su gestión) no lo inmuniza frente al peligro de erosionar garantías y derechos. La prevención especial, también la positiva, puede —según

82. Por ej. Andrés Pueyo, 2017, p. 378; Andrés-Pueyo et al, 2018: 256 y s. Resulta reseñable que también la introducción de la Tabla de Variables de Riesgo (TVR), el algoritmo que se utiliza en el territorio gestionado desde la Secretaría General de Instituciones Penitenciarias, para valorar el riesgo de quebrantamiento de los permisos penitenciarios, estuvo condicionada por la preocupación que suscitó en aquel momento el terrible crimen de Alcàsser, en el que uno de los autores había escapado de prisión el año anterior usando un permiso penitenciario (Civio, 2025).

cómo se entienda— comportar injerencias ilegítimas en el libre desarrollo de la personalidad, la intimidad, o la libertad de opinión y expresión, puede generar discriminaciones y tensionar el derecho de defensa, y el mero hecho de buscar un beneficio para el interno no garantiza que sea legítimo en un Estado de Derecho el medio a través del cual se persiga ese fin.

3. LA PRIMACÍA DEL ASPECTO TRATAMENTAL EN EL DISEÑO, IMPLEMENTACIÓN Y SUPERVISIÓN DE RISCANVI

Pues bien, a pesar de que RisCanvi incorpora tanto la dimensión de intervención y tratamiento, como la de control, en mi opinión es la primera la que ha estado más presente desde el inicio en su diseño, implantación y gestión, y lo sigue estando aún. Es decir, han primado mucho más las consideraciones criminológicas dirigidas a lograr una prevención rigurosa y fundamentada en evidencia empírica, que las relativas a asegurar que en todo el procedimiento estuviera también presente la perspectiva de garantizar los derechos de las personas internas.

Que esto es así puede deducirse de varios datos. Por ejemplo, RisCanvi fue desarrollado por el Grupo de Estudios Avanzados en Violencia (GEAV) de la Universidad de Barcelona en colaboración con el Centro de Estudios Jurídicos y Formación Especializada (CEJFE) y con la Secretaría de Servicios Penitenciarios de la Generalitat de Catalunya. El GEAV está liderado por académicos expertos en Psicología y Criminología, prestigiosos especialistas en el análisis de la valoración del riesgo, y compuesto por miembros que tienen también formación en estas disciplinas, pero sin que hubiera —salvo error u omisión por mi parte— juristas involucrados en primera línea[83].

Es también relevante, a mi juicio, que tampoco participaran juristas en la auditoría externa sobre RisCanvi encargada en 2022 por la Generalitat y cuyos resultados se publicaron en 2024. Este trabajo, conocido como "Informe Tiresias" y que en adelante citaré por esta denominación, se encargó a una consultora experta en análisis de datos[84] y fue realizado por personas

83. El hecho de que en la creación de RisCanvi participaran también personal técnico del CEJFE y de la Generalitat catalana, algunos de los cuales seguramente tuvieran formación jurídica, no asegura que esté presente la dimensión de defensa de la garantía de los derechos de los internos, porque la administración penitenciaria es una parte directamente interesada, como es lógico, en asegurar los fines que tiene encomendados, esto es, la custodia y la orientación a la reinserción.

84. Dribia: puede verse la información de la empresa en https://dribia.com/es/ (consultada por última vez el 24/12/2025).

con títulos universitarios en ingeniería matemática y física, que también trabajaron en colaboración con personal de la administración penitenciaria catalana[85].

Si subrayo ambos datos no es porque me parezca equivocada la selección de estos perfiles, ni mucho menos: es muy oportuno que se acuda a expertos en psicología y en criminología, así como a expertos en matemáticas, física y ciencia de datos, para implantar, gestionar y auditar una herramienta que funciona con datos y algoritmos y que pretende estimar y controlar el riesgo de cometer delitos. Lo que quiero destacar es que se ha omitido incluir otro tipo de perfiles, que serían los adecuados para supervisar el funcionamiento de RisCanvi desde la perspectiva de su posible afectación a derechos fundamentales: expertos en derecho penal, derecho antidiscriminatorio, y derecho penitenciario, idealmente con experiencia en el asesoramiento jurídico de los internos ante la administración penitenciaria y en su defensa en la jurisdicción de vigilancia penitenciaria. Haber prescindido de estas personas creo que refleja lo que estoy señalando: que no se ha prestado la misma atención a la dimensión de limitación de derechos que RisCanvi también tiene, queramos verla o no, que a la de intervención y tratamiento.

Y puede que esta sea una de las razones por las que a veces no se comparten o no se entienden las críticas que se han hecho a RisCanvi desde la perspectiva jurídica. Recuerdo hace años una Jornada celebrada en Barcelona sobre valoraciones de riesgo en la que yo expuse algunos reparos a la manera en que se presentaba la capacidad predictiva de RisCanvi en un estudio sobre la herramienta. Y en una de las intervenciones una persona directamente implicada en la gestión del protocolo me contestó: "es que a nosotros no nos preocupa eso". La respuesta me hizo pensar. Porque, por un lado, esa persona tenía toda la razón: desde la óptica del tratamiento, las métricas de capacidad predictiva que para ellos resultaban relevantes eran otras, diferentes a las que yo reivindicaba. Y, sin embargo, creo que esa es precisamente una parte del problema: no haber prestado suficiente atención hasta ahora a que RisCanvi se utiliza para más cosas, además del tratamiento.

Otro ejemplo: en un artículo reciente sobre RisCanvi el profesor Andrés Pueyo, respetadísimo experto en valoraciones de riesgo y coautor de la herramienta, afirma que RisCanvi no "debe entenderse como una prueba judicial en sentido estricto, ya que sus resultados se utilizan para la gestión

85. Vid. Dribia Data & Direcció General d'Afers Penitenciaris (2024). El equipo autor del informe figura en la p. 57.

penitenciaria y la intervención, no como evidencia procesal vinculante"[86]. Sin embargo, por mucho que RisCanvi quizá no se diseñara *para* ser utilizado como prueba en juicio, la realidad demuestra que sí se utiliza para eso: los niveles de riesgo que marca RisCanvi se incluyen en los informes que se trasladan a los Jueces de Vigilancia Penitenciaria cuando estos han de tomar decisiones o revisar las tomadas por las juntas de tratamiento, y son un dato que tiene mucha influencia en la decisión[87]. En un informe publicado por Karimi-Haghighi y Castillo se observa que de las personas liberadas a las que RisCanvi asigna un riesgo bajo de reincidencia violenta, entre un tercio y aproximadamente la mitad reciben la libertad condicional, mientras que de las personas que salen de prisión a las que RisCanvi asigna un riesgo alto de reincidencia violenta, solo entre el 1 y el 3 % han salido en libertad condicional. La interpretación que hacen estos autores es que RisCanvi probablemente influye en el hecho de que una persona reciba o no la libertad condicional, pero no determina este resultado (Karimi-Haghighi y Castillo, 2022, p. 2). En mi opinión, RisCanvi parece ser más determinante cuando el riesgo es alto (ya que solo entre el 1 y el 3 % de las personas con esta puntuación de riesgo obtienen la libertad condicional), pero no tanto cuando el riesgo asignado es bajo, ya que a más de la mitad de las personas con esta puntuación de riesgo no se les concede, a pesar de ello, la libertad condicional.

Y un último ejemplo: una de las recomendaciones que contiene el informe Tiresias (el resultado de la auditoría encargada por la Generalitat de Catalunya sobre RisCanvi) es sustituir el algoritmo actual por una nueva herramienta basada en modelos de inteligencia artificial, y si ello no fuera posible, revisar los puntos de corte del algoritmo actual. Y aconseja que esto se haga en colaboración con expertos en reincidencia y expertos en algoritmos de inteligencia artificial (Dribia Data & Direcció General d'Afers Penitenciaris, 2024, p. 55). No menciona en ningún momento que fuera aconsejable incluir a juristas especializados en derecho penitenciario, en normativa de protección de datos, o en derecho antidiscriminatorio, y ello

86. Vid. Andrés Pueyo (2025, p. 24). En la p. 31 del mismo trabajo repite que no es "una prueba judicial con valor pericial vinculante".

87. En una investigación reciente hemos analizado sentencias dictadas por las Audiencias Provinciales de Catalunya cuando revisan resoluciones de los Jueces de Vigilancia, y consultado la opinión de varios abogados especialistas en Derecho penitenciario y de un magistrado. Concluimos que los Jueces de Vigilancia dan bastante importancia al resultado de RisCanvi, y que los Fiscales recurren sistemáticamente las resoluciones favorables a la libertad de los presos cuando el riesgo señalado por RisCanvi es medio o alto (Martínez Garay et al, 2024, pp. 100 y ss).

a pesar de que en el informe se analizan, entre otras cosas, posibles efectos discriminatorios de RisCanvi.

En definitiva, RisCanvi no es solo un protocolo para "estimar probabilidades y orientar estrategias de intervención preventiva en el contexto de la intervención rehabilitadora de las prisiones en Cataluña" (Andrés Pueyo, 2025, p. 31). Es, sin duda, eso, pero es también algo más: una herramienta que se utiliza para fundamentar decisiones que restringen de manera muy importante la libertad.

4. PARADOJAS QUE PRODUCE LA TENSIÓN ENTRE LA DIMENSIÓN TRATAMENTAL/PREVENTIVA, Y LA DE GARANTÍAS

A mi juicio, esta tensión que existe en RisCanvi entre la dimensión tratamental/preventiva y la de afectación a derechos fundamentales, y sobre todo el hecho de que hasta ahora no se haya reconocido ni abordado de manera suficientemente abierta, permite entender algunas aparentes paradojas, como las tres que paso a explicar.

4.1. ¿PUEDE LA CAPACIDAD PREDICTIVA DE RISCANVI SER BUENA Y MALA A LA VEZ? ¿PUEDE AFIRMARSE AL MISMO TIEMPO QUE RISCANVI MINIMIZA LOS FALSOS POSITIVOS, Y QUE PRODUCE DEMASIADOS?

El "informe Tiresias", resultado de la auditoría que la Generalitat encargó sobre RisCanvi, concluye que RisCanvi prioriza minimizar los falsos positivos (personas consideradas de riesgo alto pero que no cometen después un nuevo delito o una conducta problemática), y como consecuencia genera tasas altas de falsos negativos (personas consideradas de riesgo bajo o medio pero que sí cometen después un delito o incurren en una conduta problemática). Así, afirma que:

> "los puntos de corte escogidos en todas las violencias *priorizan tener una tasa de falsos positivos baja* (FPR entre 1 y 31%) para reducir los casos de predicción de riesgo alto erróneos. Esto quiere decir que Riscanvi tiene más tendencia a dar un riesgo bajo o medio. Este hecho *implica inevitablemente tener un número de falsos negativos alto*, como se ve en la tabla (FNR entre 40 y 100%). Esto, sumado a la extrema baja reincidencia (~5%), hace que el modelo no identifique como riesgo alto correctamente los internos que sí que acaban reincidiendo (TPR), sobre todo en el quebrantamiento de condena (donde

> la prevalencia es todavía inferior, 0,5%)." (Dribia Data & Direcció General d'Afers Penitenciaris, 2024, p. 5, cursivas añadidas).

Haciéndose eco de los resultados de ese informe, Andrés Pueyo (2025, pp. 25 y 27) ha afirmado, en la misma línea, que RisCanvi subestima el riesgo de reincidencia:

> "aunque [la] precisión [de RisCanvi] (AUC-ROC) es comparable a la de otros sistemas internacionales similares, los puntos de corte actuales resultan en una alta tasa de falsos negativos (FNR), lo que significa que el modelo subestima el riesgo de reincidencia y no logra identificar correctamente a muchos internos que reinciden. […].
>
> Los puntos de corte actuales priorizan mantener una tasa baja de falsos positivos (FPR entre 1% y 31% dependiendo del riesgo), lo que reduce la asignación errónea de riesgo alto a individuos que no reincidirán. Por ejemplo, en la escala Completa, el FPR para Reincidencia General es del 6.39%, y para Reincidencia Violenta es del 9.93%. La contrapartida de la baja FPR es una tasa alta de falsos negativos (FNR entre 40% y 100%). Esto implica que el modelo tiene una fuerte tendencia a clasificar incorrectamente como riesgo bajo o medio a un número considerable de internos que sí terminan reincidiendo (baja Tasa de Verdaderos Positivos, TVP)."

Sin embargo, otras voces han alertado de que RisCanvi genera un número muy alto de falsos positivos y sobreestima el riesgo de reincidencia. Así, en publicaciones anteriores yo misma he subrayado que los instrumentos estructurados de valoración del riesgo de violencia o de reincidencia "producen una notable y sistemática sobreestimación de la peligrosidad especialmente cuando se aplican a la predicción de fenómenos con bajas tasas de prevalencia como lo es la delincuencia violenta." (Martínez Garay, 2016, p. 25). Y, en concreto en relación con Riscanvi, que dentro del grupo de personas que RisCanvi califica como de riesgo alto o moderado solo reincide un escaso 17.94% (Martínez Garay, 2016, pp. 11 y s.).

En otro estudio más reciente, en el que hemos tenido en cuenta datos más completos y actuales de valoraciones de riesgo y de reincidencia de RisCanvi, hemos afirmado que:

> "RisCanvi no es un instrumento bien calibrado: según los resultados de Karimi-Haghighi y Castillo (2022), el porcentaje de sujetos clasificados como de alto riesgo en la muestra total (10,5 %) duplica el de los que realmente cometieron un nuevo delito violento en el período de seguimiento (4,9 %). Esta proporción se mantiene similar para muchos de los subgrupos analizados. Esto significa que RisCanvi asigna sistemáticamente más riesgo del que realmente existe en los diferentes subgrupos y en la muestra total" (Martínez Garay et al, 2024, p. 105).

Y si nos fijamos en los datos de capacidad predictiva que contiene el que es hasta ahora el estudio más completo sobre RisCanvi elaborado por la Generalitat de Catalunya, se puede comprobar que, dentro del grupo de riesgo alto, el porcentaje de personas que comete después un nuevo delito que comporta ingreso en prisión es del 42,9%, y el porcentaje que comete después un nuevo delito violento que comporta ingreso en prisión, es del 17,6 % (Capdevila et al, 2023, pp. 217 y s. Tablas 69 y 71)[88].

Por sorprendente que pueda parecer, todas las afirmaciones que acabo de reproducir son correctas. En efecto, según cómo se mire, RisCanvi subestima o sobreestima el riesgo de reincidencia, y tiene un porcentaje muy alto o muy bajo de falsos positivos. La explicación a esta aparente paradoja radica en la perspectiva desde la cual se analicen los datos. Pongamos como ejemplo una de las tablas que aparece en el estudio publicado en 2023 por el Centre d'Estudis Jurídics i Formació Jurídica Epecializtzada, sobre *Tasa de reincidencia penitenciaria en Cataluña 2020*:

Tabla 1. Tasa penitenciaria violenta agrupando riesgo alto frente a riesgo medio y bajo.

Valoración RisCanvi	**Reincidencia penitenciaria violenta observada**		
	Sí	**No**	**Total**
Riesgo alto	41 (a)	192 (b)	233 (a+b)
Riesgo bajo/medio	120 (c)	2378 (d)	2498 (c+d)
Total	161 (a+c)	2570 (b+d)	2731

Sensibilidad	25,5 %
Especificidad	92,5 %
Valor predictivo positivo	17,6 %
Valor predictivo negativo	95,2 %
Proporción de falsos positivos	7,5 %
Proporción de falsos negativos	74,5 %
Exactitud	88,6 %
Razón de probabilidades	4,23

Fuente: Capdevila Capdevila et al. (2023, p. 218).

88. Son los datos que salen cuando se agrupa el riesgo alto, por un lado, frente al medio y bajo por otro; las cifras cambian si se agrupa el riesgo alto y medio, por un lado, frente al bajo: en este caso los valores predictivos son inferiores: 32,3% para la reincidencia penitenciaria, y 13% para la reincidencia violenta penitenciaria. Indico en el texto los valores que resultan si se agrupa el riesgo alto frente al medio y bajo, porque esa es también la opción que escoge el Informe Tiresias.

Como puede observarse, la tasa (proporción) de falsos positivos es baja: solo del 7,5%. Y la tasa de falsos negativos es alta: 74,5%. El Informe de auditoría sobre Riscanvi (Informe Tiresias) también da cifras parecidas para la estimación de riesgo de reincidencia violenta penitenciaria: tasa de falsos positivos del 4,77% en el RisCanvi Screening y del 9.93% en el completo; tasa de falsos negativos de 84,25% y de 72,29%, respectivamente (Dribia Data & Direcció General d'Afers Penitenciaris, 2024, p. 5).

Ahora bien, ¿qué es lo que dicen estas tasas? Pues, si atendemos a la tabla que he reproducido en el texto, significan que, respecto del grupo de 2.570 sujetos (b+d) que tras ser puestos en libertad comprobamos que no han reingresado en prisión por un delito violento, RisCanvi calificó como de riesgo alto a 192 (b), es decir, el 7,5%. Y que respecto de los 161 sujetos (a+c) que tras ser puestos en libertad vemos que han reingresado en prisión por un delito violento, RisCanvi evaluó en su momento a 120 (c) como de riesgo medio o bajo, es decir, el 74,5%.

Estas tasas son sin duda relevantes. Porque de las 161 personas que han vuelto a prisión como consecuencia de la comisión de un nuevo delito violento, hay muchas que no tenían la valoración de riesgo alto (el 74,5%), y por tanto sobre esas no se habrá intervenido tanto como sobre las que sí tenían esa etiqueta: se ha perdido la oportunidad de aplicar programas o tratamientos que eventualmente habrían podido reducir esa reincidencia.

Sin embargo, estas tasas no son las únicas que importan. Porque cuando la junta de tratamiento valora si concede o no un tercer grado, o cuando el Juez de Vigilancia Penitenciaria tiene que pronunciarse sobre la libertad condicional, no saben aún —lógicamente— si la persona que tienen delante pertenece al grupo de reincidentes o de no reincidentes. Lo único que saben es qué nivel de riesgo tiene asignado en RisCanvi, y tienen que tomar la decisión con apoyo en ese dato. Por ello, las preguntas se plantean de manera un poco diferente:

– ¿cuál es la probabilidad de que, si el sujeto tiene en RisCanvi un riesgo alto, realmente reingrese en prisión por un nuevo delito violento? Es del 17,6% (valor predictivo positivo). Porque, de las 233 personas a las que RisCanvi asigna un riesgo alto de reincidencia violenta penitenciaria (a+b), tras el período de seguimiento solo reingresan en prisión por un nuevo delito violento 41 (a), es decir, el 17,6%. El resto, es decir, el 82,4%, son falsos positivos (los 192 sujetos de la celda b).

– ¿cuál es la probabilidad de que, si el sujeto tiene en RisCanvi un riesgo medio o bajo, no reingrese en prisión por un nuevo delito violento? Es del 95,2% (valor predictivo negativo). Porque, de las 2.498 personas a las que RisCanvi asigna un riesgo medio o bajo de reincidencia violenta penitenciaria (c+d), tras el período de seguimiento hay 2.378 que no han reingresado en prisión por un nuevo delito violento (d), es decir, el 95,2%. El resto, es decir, el 4,8%, son falsos negativos (los 120 sujetos de la celda c).

Estos porcentajes son también muy relevantes. El 82,4% de falsos positivos significa que, si a 100 internos que tienen todos ellos riesgo alto en RisCanvi, la junta de tratamiento o el Juez de Vigilancia les deniegan por ese motivo el acceso a un régimen de semilibertad o de libertad condicional, hay 82 personas a las que se está impidiendo salir de prisión que, sin embargo, no habrían vuelto a reingresar por un delito violento. En otras palabras: se les está restringiendo muy notablemente la libertad en aras de la protección de la seguridad pública cuando, si hubieran tenido la oportunidad, no habrían vuelto a cometer delitos violentos. Y en cuanto al porcentaje del 4,8% de falsos negativos, significan que, si a 100 internos que tienen todos ellos riesgo bajo o medio en RisCanvi, la junta de tratamiento o el Juez de Vigilancia les conceden por ese motivo el acceso a un régimen de semilibertad o de libertad condicional, solo 5 personas (redondeando el 4,8%) volverán a reingresar por un delito violento.

¿Es un 82,4% de falsos positivos una cifra alta? Personalmente, diría que sí. ¿Es un 4,8% de falsos negativos una cifra baja? También creo que sí. Al menos en el informe Tiresias porcentajes similares, de entre el 81 y el 84% (de falsos negativos) se han considerado altos, y de entre el 4 y el 10% (de falsos positivos) se han considerado bajos.

Pero entonces, ¿en qué quedamos: RisCanvi subestima el riesgo o lo sobreestima? La respuesta es que no cabe una respuesta tajante, de sí o no, sino que hay que matizar y diferenciar. RisCanvi produce pocos falsos positivos y muchos falsos negativos (como afirma el informe Tiresias de auditoría, y como sostiene Andrés Pueyo, 2025) si nos fijamos en su eficacia diagnóstica, y calculamos esos porcentajes sobre el total de reincidentes y no reincidentes. Pero produce muchos falsos positivos y pocos falsos negativos si nos fijamos en su eficacia pronóstica, y calculamos esos porcentajes sobre el total de personas clasificadas como de riesgo alto y medio/bajo.

Desde luego, estas afirmaciones hay que matizarlas todavía más: los números que salen no son los mismos si en lugar de agrupar el riesgo medio

con el bajo, agrupamos el riesgo medio con el alto, y consideramos que para ambos niveles la reincidencia confirma el pronóstico (es lo que se hace en la tabla 70 de Capdevila Capdevila et al, 2023, p. 217, y en la tabla de la p. 72 del Informe Tiresias). Las cifras también cambian si en lugar de fijarnos en la predicción de reincidencia penitenciaria violenta atendemos a la predicción de otros de los riesgos que RisCanvi también calcula, como el de reincidencia penitenciaria con cualquier tipo de delito (tablas 68 y 70 de Capdevila Capdevila et al, 2023, p. 217, y en la tabla de la p. 72 del Informe Tiresias). Y por supuesto la valoración de unas u otras cifras como "demasiado altas" o "demasiado bajas" dependerá de cuáles sean los errores que consideremos prioritario evitar, así como de cuál sea el tipo de conducta cuya ocurrencia queremos anticipar: por ejemplo, me parece que respecto de la escala de violencia autodirigida, esto es, prevención de suicidio y autolesiones, la perspectiva que debe adoptarse es muy distinta, ya que en estos casos no se está restringiendo la libertad al interno para proteger a otros del peligro que él representa, sino que se está intentando proteger su vida y su integridad física, en una situación en la que además la administración penitenciaria ostenta una posición de garante[89].

No es posible por razones de espacio efectuar en estas páginas un análisis detallado de cada una de estas dimensiones, pero creo que tampoco es necesario para la idea que quiero destacar, y que es la siguiente: la manera en que se suele analizar la capacidad predictiva evidencia que los responsables de RisCanvi han prestado mucha más atención a una de las dimensiones de RisCanvi que a la otra. Las publicaciones y los informes destacan las tasas de falsos positivos y de falsos negativos que preocupan desde la óptica tratamental, esa vertiente de RisCanvi que tiene como objetivo lógico y legítimo identificar, de entre las personas que luego reinciden y que no lo hacen, al mayor porcentaje posible para ajustar el tipo y el grado de intervención al riesgo real y a las necesidades de cada preso. Pero se siguen obviando los porcentajes de falsos positivos y falsos negativos que se producen cuando

89. Y aún habría que hacer más matizaciones: todas estas cifras son meras aproximaciones, porque sobre la base del riesgo que RisCanvi estima se toman decisiones para minimizarlo. Por ejemplo, el porcentaje de reincidentes sobre el total de personas calificadas como de riesgo alto quizá esté subestimado en todas estas tablas, ya que cuando se da la libertad o un permiso a alguien de alto riesgo seguramente se adoptan medidas de vigilancia y acompañamiento que, si son eficaces, pueden evitar reincidencias que de otro modo sí se habrían producido (tal como advertía —correctamente— la predicción de riesgo alto). En otras palabras: la gestión eficaz de riesgos estimados correctamente puede convertir en aparentemente equivocadas esas estimaciones, cuando las medidas consiguen evitar el peligro anticipado.

RisCanvi se utiliza para esa otra finalidad para la que también fue diseñado, y que es tomar decisiones sobre los presos peligrosos para minimizar el riesgo de reincidencia violenta. Si en estas decisiones se utilizara el riesgo alto de reincidencia violenta como un criterio decisivo para denegar la libertad, las tasas de falsos positivos superarían el 80%. Este es un dato muy preocupante desde el punto de vista de las garantías del preso, porque es probable que tanto Jueces como Fiscales desconozcan esa cifra (y piensen que el porcentaje de acierto es bastante mayor, a raíz entre otras cosas de las reiteradas afirmaciones de que RisCanvi "subestima" el riesgo), y porque el interno tiene muy difícil rebatir la calificación de riesgo, ya que no conoce el funcionamiento interno del algoritmo.

Dicho esto, hay que reconocer que los estudios publicados por la Generalitat de Catalunya y el CEFJE sí incluyen ahora (a diferencia del estudio de 2015), los valores predictivos positivo y negativo cuando informan sobre la capacidad predictiva de RisCanvi, cosa que desde mi punto de vista es una mejora metodológica importante. Sin embargo, me parece muy preocupante —porque revela lo poco que parecen haber cambiado a pesar de ello las cosas— el hecho de que la auditoría externa encargada por la Generalitat guarde absoluto silencio sobre estas métricas. En efecto, el informe Tiresias da cuenta de las tasas de falsos positivos y falsos negativos, y de verdaderos positivos y verdaderos negativos, para las 5 escalas de RisCanvi (Dribia Data & Direcció General d'Afers Penitenciaris, 2024, pp. 5 y 72). Pero no menciona en ningún momento los valores predictivos, es decir, los porcentajes de verdaderos y falsos positivos y negativos calculados sobre el total de predicciones de riesgo alto, medio o bajo. Es más, dichos porcentajes ni siquiera se pueden calcular a partir de los datos que contiene el informe, porque este solo ofrece porcentajes y no contiene tablas de contingencia con los totales (N)[90].

90. La ausencia de juristas o expertos en derecho penitenciario en la elaboración del informe Tiresias quizá explica otras afirmaciones contenidas en el mismo, como por ej. cuando en la p. 21, tras comentar que el porcentaje de reincidencia es en general muy bajo en todas las escalas y especialmente en la de quebrantamiento de condena, señala que "el porcentaje de internos susceptibles de cometer esta reincidencia es muy pequeño porque han de tener un permiso de salida o bien estar en un grado de clasificación que les permita salir de la prisión. Esto explicaría que la reincidencia detectada esté en torno al 0,58% de las evaluaciones". Sin embargo, el número de permisos ordinarios y extraordinarios concedidos a internos clasificados en segundo y tercer grado en Cataluña ha sido superior a 30.000 al año entre 2011 y 2016 (superior a 40.000 en ese último año), lo que a mi juicio difícilmente puede calificarse como un número "pequeño" (datos extraídos de Rovira, Larrauri y Alarcón (2018, p. 26), y la

4.2. LA CAPACIDAD PREDICTIVA ES MEJOR O PEOR SEGÚN CUÁL SEA EL OBJETIVO QUE PERSEGUIMOS. ¿CUÁL ES EL OBJETIVO Y QUIÉN DEBE ESTABLECERLO?

Para que se me entienda bien, y a riesgo de resultar reiterativa: no se trata de que la capacidad predictiva de RisCavni sea "mala". Como he afirmado en otros lugares, y han destacado también otros estudios, su capacidad predictiva es similar a la de otras herramientas parecidas en el contexto comparado[91]. Pero el problema es que la capacidad predictiva tiene muchas vertientes, y no puede resumirse en una única métrica[92]. Salvando todas las distancias, podría compararse a si la decisión de comprar un determinado modelo de coche es "buena". La calidad del automóvil puede medirse de muchas formas: consumo de gasolina, emisión de gases contaminantes, precio, potencia, calidades interiores, años de garantía, número de plazas, capacidad del maletero, dimensiones de la plaza de parking que tengo en el garaje, etc. Que sea una buena opción depende de las necesidades del comprador: una familia numerosa tendrá que priorizar el espacio, aunque sea a costa de un precio más alto; una persona con poca capacidad económica tendrá que elegir el más barato aunque el coche sea más contaminante de lo que querría, y así sucesivamente.

Pues con los algoritmos de valoración del riesgo pasa algo parecido. ¿Para qué quiero medir el riesgo, para priorizar a las personas que tienen más carencias y necesidades, y necesitan por tanto más intervención en el tratamiento penitenciario, o para tomar decisiones que implican limitación de derechos como la concesión o no del tercer grado? Si es lo primero, necesito un instrumento con una elevada sensibilidad; si es lo segundo, debo fijarme en los valores predictivos.

¿Qué es lo que quiero, asegurarme de que identifico correctamente a los sujetos de bajo riesgo, para acelerar su salida de prisión, o asegurarme de que identifico bien a los de riesgo alto, para demorar su salida o rodearla de controles adicionales? Si es lo primero, lo racional es situar el punto de corte entre riesgo alto y bajo bastante arriba, para minimizar los falsos positivos y no mantener en prisión a nadie cuyo nivel de riesgo no lo justifique, aunque sea a costa de asumir un porcentaje más grande de falsos negativos. Si es lo

población penitenciaria que disfruta de un permiso en algún momento de la condena ronda el 50% (los mismos, p. 14).

91. Martínez Garay, 2016; Dribia Data & Direcció General d'Afers Penitenciaris, 2024, p. 4 ("Los niveles de AUC-ROC son comparables a otros algoritmos en el panorama internacional (COMPAS 0,7. OASys 0,75)").

92. *In extenso*, Martínez Garay y Montes Suay (2018) y Martínez Garay (2020).

segundo, deberé situar el punto de corte mucho más abajo, para minimizar los falsos negativos y no dejar salir a nadie cuyo riesgo no lo aconseje, pero al precio de asumir un número de falsos positivos mucho mayor.

¿Quiero poder separar a los que tienen más riesgo que los demás para reservar a esos las plazas en programas de tratamiento que no alcanzan para todos? Entonces necesito una herramienta que discrimine bien entre niveles de riesgo mayores y menores, cosa que ocurre cuando el área bajo la curva ROC es elevada.

Insistir en que la capacidad predictiva de RisCanvi o del resto de herramientas de valoración del riesgo similares es "buena", o incluso "mejor" que las predicciones de riesgo tradicionales, no es decir ninguna mentira. Porque hay cosas que estas herramientas sí hacen bien, e incluso mejor. Por ejemplo, diversos estudios que han comparado las capacidades predictivas de unos y otros métodos coinciden en señalar que las herramientas estructuradas (ya sean actuariales o de juicio clínico estructurado) alcanzan generalmente mejores niveles de área bajo la curva ROC[93]. Esto puede ser una buena base sobre la que diseñar tratamientos y seleccionar a las personas para unos u otros programas de reinserción. Pero insistir sin matices en que la capacidad predictiva de RisCanvi y del resto de herramientas similares es "buena", o incluso "mejor" que las predicciones de riesgo tradicionales, es decir solo una media verdad. Primero, porque silencia información relevante. Por ejemplo, si lo que nos preocupa es identificar a los delincuentes que tienen más probabilidad de reincidir con nuevos delitos violentos, para aplicar regímenes penitenciarios más estrictos y demorar o dificultar la concesión de permisos, la progresión de grado o la libertad condicional, en este aspecto los niveles de acierto de las estimaciones de riesgo, como también han subrayado diversos autores, siguen siendo muy parecidos a los de hace 50 años, con porcentajes de falsos positivos superiores al 50% y en ocasiones al 80%[94].

93. El estudio más reciente y completo que conozco es el de Viljoen et al. (2024) que contiene ulteriores referencias y destaca también numerosas limitaciones. Véanse también Fazel, Burghart, Fanshawe et al. (2022), quienes subrayan el moderado rendimiento predictivo de estas herramientas, especialmente si se atiende a los —escasos— estudios independientes de validación, y la necesidad de reportar métricas distintas al área bajo la curva ROC, así como la necesidad de tener en cuenta consideraciones éticas y de transparencia.

94. Cfr. por todos Tonry (2019, pp. 273 y ss.) con ulteriores referencias. Como indica este autor, en los años 80 estos porcentajes fueron considerados excesivos e injustos por el sacrificio que suponen en la libertad de los internos. No deja de llamar la atención que en el debate actual sobre la capacidad predictiva de las herramientas de estimación del

En segundo lugar, porque oscurece una realidad incómoda: que no podemos tenerlo todo al mismo tiempo. Como explicaba antes con el ejemplo del coche: no puedo tener a la vez un coche muy pequeño pero en el que quepa una familia numerosa, un deportivo nuevo de alta gama que sea muy barato, o un vehículo eléctrico que me dé autonomía para circular durante muchos kilómetros a altas velocidades sin necesidad de recargar la batería. Aplicado a las herramientas de valoración de riesgo, podemos diseñar las herramientas para minimizar los falsos positivos, o bien los falsos negativos, pero no es posible reducir ambos a la vez y en la misma medida. En otras palabras, hay que elegir. Y esa elección no es una mera cuestión técnica: supone una opción valorativa —y política— porque implica establecer determinados equilibrios entre seguridad pública y derechos fundamentales.

4.3. ¿PRODUCE RISCANVI EFECTOS DISCRIMINATORIOS PROHIBIDOS, O NO LOS PRODUCE?

En los últimos diez años, y sobre todo a partir del caso COMPAS, se ha debatido mucho acerca de si las herramientas de estimación del riesgo tienen o no efectos discriminatorios.

Vaya por delante que la estandarización y estructuración del procedimiento de valoración del riesgo que suponen estos protocolos frente a las estimaciones clínicas de peligrosidad mitigan mucho la influencia de los sesgos o prejuicios subjetivos que pueden aquejar a estas últimas. Las personas tenemos sesgos de todo tipo: racistas, sexistas, clasistas, etc., y eso puede condicionarnos indebidamente a la hora de tomar decisiones. Protocolos como RisCanvi buscan más objetividad, y sin duda lo consiguen en alguna medida, porque minimizan los efectos discriminatorios que pueden derivar de estos sesgos o prejuicios subjetivos[95].

Pero aquí me refiero a un problema distinto, o mejor dicho, a dos problemas: el primero, si RisCanvi y herramientas similares producen discrimina-

---

riesgo de reincidencia se pase de puntillas sobre esta cuestión, si es que llega siquiera a mencionarse.

95. No obstante, hay ítems en RisCanvi cuya codificación abre la puerta a la subjetividad, porque son ambiguos o implican la previa realización de valoraciones de manera que según el profesional que haga la evaluación el resultado puede ser diferente. Como ha señalado Alemán Aróstegui (2023) en un trabajo sobre la aplicación práctica de RisCanvi, "Casi todos/as los/as profesionales entrevistados/as insisten en la ambigüedad y subjetividad de los distintos factores de riesgo […] consideran que el protocolo RisCanvi es un instrumento muy subjetivo y que, dependiendo de quién lo realice, puede salir un resultado u otro completamente distinto." (pp. 23 a 26; la cita es de la p. 25).

ción directa cuando hacen que aumente el riesgo debido a la presencia de factores que el sujeto no puede elegir, que le colocan en situación de desventaja, o que padece contra su voluntad; y, en segundo lugar, si RisCanvi y las herramientas parecidas producen discriminación indirecta cuando arrojan para determinados grupos de personas niveles mayores de riesgo que para otros, aunque las circunstancias que individualizan a esos grupos no estén expresamente contempladas como factores de riesgo por el algoritmo[96].

La preocupación por este problema en relación con RisCanvi es reciente. Salvo error u omisión por mi parte, la primera vez en que se ha analizado si RisCanvi produce discriminación ha sido en la auditoría que se encargó en 2022, y cuyos resultados se publicaron en 2024[97]. El informe concluye que no se han encontrado sesgos discriminatorios graves en el RisCanvi completo para los grupos protegidos de mujeres, extranjeros y menores de 30 años. En el RisCanvi Screening sí hay diferencias, porque ahí la tasa de falsos negativos es más alta para esos tres grupos protegidos por lo que hace al riesgo de violencia autodirigida, lo que significa que en estos grupos se subestima el riesgo de autolesión y de suicidio. También en el RisCanvi Screening, la tasa de falsos negativos es mucho mayor en los extranjeros para la estimación de riesgo de reincidencia general y de reincidencia violenta, lo que en este caso significa que, dentro del grupo de reincidentes y de reincidentes violentos, los extranjeros "escapan" más veces de la detección como personas de alto riesgo (aquí el algoritmo se equivoca "a favor" de estos grupos protegidos)[98]. Según este informe, en definitiva, RisCanvi no parece tener problemas importantes de discriminación en contra de grupos vulnerables, salvo la cuestión puntual del riesgo de violencia autodirigida en el RisCanvi Screening.

Sin embargo, también aquí las cosas aparecen bajo una luz distinta si se observan desde una perspectiva diferente. Sin poner en duda que los resultados del informe Tiresias serán correctos y estarán bien calculados, son un análisis insuficiente y ofrecen una imagen muy incompleta de la situación, por varias razones.

La primera es que al menos desde que ProPublica denunciara la supuesta discriminación de la minoría afroamericana en el caso de COMPAS, y del debate sobre la justicia algorítmica (*algorithmic fairness*) que le siguió, sabe-

96. Véanse las definiciones de discriminación directa e indirecta y de las causas por las cuales está prohibido discriminar en los arts. 6 y 2 de la Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación.

97. Tampoco la doctrina española ha prestado mucha atención a este problema, a pesar de que es una cuestión que se ha debatido intensamente en el contexto internacional. Una excepción es Julià Pijoan (2023).

98. Dribia Data & Direcció General d'Afers Penitenciaris (2024, pp. 5, 35-41).

mos que comparar tasas de falsos negativos y falsos positivos (sobre el total de reincidentes y no reincidentes, respectivamente) no es la única manera de analizar posibles sesgos discriminatorios. Por supuesto, es deseable que, entre las personas que finalmente reinciden, varones y mujeres (o jóvenes y mayores) hayan sido considerados de riesgo alto en una proporción parecida (tasa de verdaderos positivos o sensibilidad). Esto es lo que ha analizado el informe Tiresias en la auditoría de RisCanvi, con los resultados arriba señalados. Pero también es importante cómo se comportan los valores predictivos positivo y negativo entre los grupos, es decir, que entre las personas que son consideradas de riesgo alto, haya un porcentaje similar de varones y de mujeres (o de jóvenes y mayores, etc.) que finalmente reinciden después. Y es que puede ocurrir que entre las primeras no haya diferencias significativas, y sin embargo entre los valores predictivos sí, o al revés. De hecho, los diversos estudios que se han publicado sobre esta cuestión concluyen que es muy difícil conseguir una paridad de trato en ambas dimensiones, cuando las tasas de prevalencia del fenómeno son diferentes entre los grupos[99]. Puesto que el informe Tiresias analiza solo un aspecto de la posible discriminación algorítmica, es muy difícil valorar si RisCanvi produce o no otras clases de discriminación a partir de los resultados de ese informe.

En segundo lugar, las mujeres, los jóvenes y los extranjeros no son los únicos "grupos protegidos" que cabe identificar en la población cuyo riesgo evalúa RisCanvi. Esta herramienta utiliza como factores de riesgo muchas otras características personales que también pueden considerarse motivos discutibles para un trato desigual. Por ejemplo, padecer trastornos mentales o tener discapacidad intelectual, o circunstancias ligadas a la marginalidad socioeconómica. En un estudio elaborado por un grupo de investigadores de las Universidades de Valencia y Politécnica de Valencia (Martínez Garay et al, 2024) hemos intentado comprobar si RisCanvi produce discriminación directa o indirecta ligada a estos factores, y los resultados obtenidos, aunque provisionales y necesitados de validación con muestras más amplias, son preocupantes. Por razones de espacio mencionaré aquí brevemente solo los resultados relativos a la presencia de trastornos mentales, consumo de sustancias y discapacidad intelectual.

99. Berk et al. (2018). Un metaanálisis reciente sobre posible discriminación indirecta de herramientas algorítmicas de estimación de riesgo de reincidencia ha encontrado más evidencias en contra de que produzcan discriminación indirecta que evidencias a favor. No obstante, el estudio analizó sobre todo herramientas utilizadas para la adopción de medidas cautelares (pretrial risk assessments), y no herramientas complejas utilizadas para la gestión del sistema penitenciario como RisCanvi; se centró en estudios sobre discriminación indirecta por raza / etnia, y no la que es debida a otras características sensibles; y señala que los trabajos existentes hasta ahora tienen defectos metodológicos y que es necesaria más investigación al respecto (Lawson et al, 2024).

En una muestra de 2514 personas, todas las excarceladas en Cataluña en 2016 en libertad definitiva o libertad condicional, con un período de seguimiento de 5 años, establecimos cuatro grupos, en función de que los sujetos tuvieran o no alguno/s de estos cinco factores de riesgo: abuso de drogas (1), abuso de alcohol (2), baja capacidad cognitiva (3), trastorno mental severo (4), y trastorno de la personalidad relacionado con la ira, la impulsividad o la violencia (5). En el primer grupo englobamos a las personas en las que no concurrían ninguno de estos factores; en el segundo grupo ("solo trastorno mental"), a aquellos en quienes concurrían uno o varios de los factores 3, 4 o 5, pero ni el 1 ni el 2; en el tercer grupo ("solo abuso de sustancias"), a las personas en quienes concurrían uno o ambos de los factores 1 y 2, pero ninguno de los factores 3, 4 y 5; y finalmente, un cuarto grupo para las personas con un sí en uno o más de los factores 3, 4 y 5, y además con un sí en uno o ambos de los factores 1 y 2. Después comparamos los resultados de la estimación de riesgo de reincidencia penitenciaria violenta hecha por RisCanvi con los datos de reincidencia penitenciaria real por un nuevo delito violento[100]. Y los resultados fueron los siguientes:

Tabla 2: Porcentaje de internos clasificados como de riesgo alto y porcentaje de reincidentes, en función de la presencia de trastorno mental y abuso de sustancias.

	1. No trastorno mental ni abuso de sustancias	2. Solo trastorno mental	3. Solo abuso de sustancias	4. Trastorno mental y abuso de sustancias
% de reincidentes	3.4%	7.3%	8.7%	8.0%
% que RisCanvi clasifica como de alto riesgo	2.7%	25.3%	22.7%	55.2%
Sensibilidad o tasa de verdaderos positivos (riesgo alto entre quienes reinciden)	13.6%	36.4%	31.4%	60%
Tasa de falsos negativos (riesgo bajo entre quienes reinciden)	86.4%	63.6%	68.6%	40%
Valor predictivo positivo (reincidentes entre los clasificados como de riesgo alto)	16,67%	10,53%	11,90%	8,7%
N	1757	150	481	125

Fuente: extraído de la Tabla 16 en Martínez Garay et al. (2024, p. 87) y de la tabla 3.6 en Karimi Haghighi y Castillo (2022).

100. El periodo de seguimiento para comprobar la reincidencia fue de 5 años. Los detalles técnicos pueden consultarse en Karimi-Haghighi y Castillo, 2022, así como en Martínez Garay et al. (2024, pp. 72 y ss. y 87 y ss).

Como puede observarse (en la fila "% que RisCanvi clasifica como de alto riesgo"), las personas con trastorno mental (grupo 2) tienen un mayor riesgo de reincidencia violenta según RisCanvi, al igual que las personas que abusan del alcohol o de otras drogas (grupo 3 y grupo 4). En principio, esto es coherente con la constatación de que quienes se encuentran en uno de estos dos grupos reinciden más que quienes no lo están (cifras de la fila "% de reincidentes"). Sin embargo, se produce el siguiente efecto: los individuos con trastorno mental (grupo 2) reinciden algo más del doble (7,3%) que los que no padecen ni trastorno ni abuso de sustancias (grupo 1: 3,4%), y la reincidencia es ligeramente superior cuando los trastornos mentales se combinan con el abuso de sustancias (grupo 4: 8%). Pero el porcentaje de personas para las que RisCanvi predice un riesgo alto (cifras de la segunda fila) no se multiplica en la misma proporción, sino que es mucho mayor: es 8 veces mayor en el segundo grupo (25,3%) que en el primero (2,7%), y 18 veces mayor en el cuarto (55,2%). Es decir, RisCanvi tiende a sobrestimar el riesgo de reincidencia, en relación con el nivel real de reincidencia observado, entre las personas diagnosticadas con problemas de salud mental (tengan o no problemas de abuso de sustancias, pero en mayor medida cuando, además, sí los tienen).

Vale la pena atender también a cómo se comportan las tasas de falsos positivos y de falsos negativos en los distintos grupos, que es la perspectiva desde la cual se analizan los posibles efectos discriminatorios de Riscanvi en el informe Tiresias. La tasa de falsos negativos es mucho más alta entre quienes no tienen trastornos mentales, baja capacidad intelectual, ni consumo de sustancias (86,4%) que entre quienes sí padecen alguna de estas situaciones, y especialmente que entre quienes padecen ambas (40%). Esto significa que quienes no tienen estos problemas se "escapan" mucho más de ser detectados como de riesgo alto por el algoritmo. Teniendo en cuenta que la tasa de un grupo es el doble que la del otro, considero que se trata de una diferencia significativa.

Por otro lado, el valor predictivo positivo es más alto en quienes no tienen estos problemas (grupo 1) que en quienes sí los tienen. Esto quiere decir que entre quienes no tienen trastornos mentales ni problemas de drogas, reincide el 16,67% de los considerados de riesgo alto, pero entre quienes sí padecen tanto una cosa como la otra, solo reincide la mitad, el 8,7% de los considerados de riesgo alto. De nuevo, la magnitud de la diferencia me parece que la convierte en una diferencia de trato relevante.

En el estudio citado efectuamos un análisis similar en relación con variables asociadas a situaciones de marginalidad socioeconómica, con resultados

parecidos, pero en los que por razones de espacio no es posible entrar aquí[101]. Me limitaré por ello a señalar las cuestiones que estos datos ponen sobre la mesa. Hay indicios de que RisCanvi sí parece producir diferencias de trato relevantes. No en todos los grupos que pueden ser considerados vulnerables, como demuestra el informe Tiresias, pero sí en algunos, como sugieren nuestros datos. Y probablemente no en todas las métricas con las que puede medirse la discriminación algorítmica, pero sí en algunas. Por ello, en modo alguno puede resolverse la cuestión afirmando que RisCanvi no produce discriminaciones graves, sino que es necesario estudiar el tema con mucha más profundidad de lo que se ha hecho hasta ahora. Y al hacerlo hay que tener en cuenta tanto la discriminación directa como la indirecta. Los datos que he ofrecido *supra*, relativos a la sobreestimación del riesgo en los casos de personas que padecen trastornos mentales o abusan de sustancias, resultan problemáticos desde ambas perspectivas. Por lo que hace a la discriminación directa, ¿es admisible que tener una baja capacidad intelectual compute como factor que incrementa el riesgo de reincidencia violenta? Como hemos repetido varias veces a lo largo de las páginas precedentes, hay varias decisiones que afectan intensamente a la libertad de los presos y que dependen del riesgo de reincidencia (concesión de permisos de salida, clasificación penitenciaria, libertad condicional, revisión de la pena de prisión permanente, por ej.). A mayor riesgo, menor probabilidad de obtener estas ventajas, lo que constituye obviamente un trato desfavorable. Si una persona con discapacidad va a estar más expuesta a este trato desfavorable que una que no tenga discapacidad, *ceteris paribus*, porque la discapacidad está expresamente contemplada como factor que incrementa el riesgo, ¿no se trata de un caso de discriminación directa por razón de discapacidad?[102]. Y, por lo que hace a la discriminación

101. Cfr. Martínez Garay, et al. (2024, pp. 97 y ss.), así como Karimi-Haghighi y Castillo (2022), y Briz-Redón y Montes (2022).

102. El caso de la discapacidad intelectual es especialmente llamativo porque, además de que el art. 2.1. de la Ley 15/2022 prohíba la discriminación por "discapacidad", tras la aprobación de la Convención de Naciones Unidas sobre los Derechos de las personas con discapacidad en 2006, su ratificación por España, y la reforma del art 49 de la Constitución, la prohibición de discriminación por motivos de discapacidad está consagrada al más alto nivel. Pero son muchos los factores de riesgo de RisCanvi que pueden aparecer como problemáticos desde la óptica de la prohibición de discriminación, habida cuenta de que los motivos de discriminación prohibidos son muchos, y de que RisCanvi utiliza como factores de riesgo muchas circunstancias personales que el sujeto no puede elegir (ser varón, ser joven, ser extranjero), o que están relacionadas con opciones legítimas en un Estado de derecho (tener una vida sexual promiscua, tener opiniones y actitudes hostiles hacia el sistema penitenciario y las autoridades públicas), o que son circunstancias desfavorables que la persona probablemente sufre contra su voluntad (haber sufrido maltrato en la infancia, ser

indirecta, la escala de reincidencia violenta de RisCanvi sí incluye como factores que incrementan el riesgo el abuso de drogas y el de alcohol, así como la baja capacidad cognitiva. Sin embargo, ni el padecer un trastorno mental severo ni tener diagnosticado un trastorno de la personalidad relacionado con la ira, la impulsividad o la violencia están contemplados en esa escala[103]. No obstante, en estos grupos el riesgo de reincidencia violenta que da RisCanvi afecta a un porcentaje de personas mucho mayor. En otras palabras, las personas con trastornos mentales, que constituyen un colectivo especialmente vulnerable en el contexto penitenciario, reciben como grupo un trato más severo en la escala de riesgo de reincidencia violenta, aunque el padecimiento del trastorno no figura expresamente como factor que la escala tiene en cuenta para incrementar el nivel de riesgo. ¿Estamos ante un caso de discriminación indirecta por enfermedad?

Por supuesto, todas estas consideraciones[104] pueden parecer extrañas si la cuestión se mira, no desde la perspectiva de las garantías del preso, sino desde la del tratamiento que más le conviene. Porque si de lo que se trata es de identificar las necesidades criminógenas de la persona, y de diseñar el tipo de intervención que más le pueda ayudar a ser capaz de vivir en libertad sin reincidir, parece lógico que sea necesario conocer todas estas características. Difícilmente podría llevarse a cabo el tratamiento individualizado del que habla nuestra Ley Orgánica General Penitenciaria sin tener en cuenta la discapacidad, los trastornos mentales, el entorno familiar y social del preso, su situación socioeconómica, características de personalidad, etc.

Y así encontramos aquí, de nuevo, la tensión entre las dos caras de RisCanvi. Lo que es conveniente —imprescindible, incluso— conocer para diseñar programas e intervenciones que minimicen el riesgo de reincidencia y favorezcan que la persona esté en condiciones de vivir en libertad sin reincidir, resulta, sin embargo, problemático cuando se utiliza para calcular un

---

pobre, no tener apoyo social, tener una enfermedad mental o una discapacidad etc.). Hace un análisis detenido de varios de estos factores desde la perspectiva de la normativa europea Julià Pijoan (2023), concluyendo que RisCanvi produce efectivamente un trato discriminatorio prohibido.

103. Aunque no son públicos todos los algoritmos con los que funciona RisCanvi, sí hemos podido conocer el de la escala de estimación del riesgo de reincidencia violenta gracias al trabajo de Karimi-Haghighi y Castillo (2022, pp. 9 y s).

104. A estas consideraciones habría que añadir otras, como por ejemplo que a RisCanvi se la aplica la Ley Orgánica 7/2021, de protección de datos personales tratados para fines de prevención, detección, investigación y enjuiciamiento de infracciones penales y de ejecución de sanciones penales, que contiene exigencias especiales para el tratamiento de categorías especiales de datos personales; cfr. sobre ello Julià Pijoan (2023).

nivel de riesgo sobre la base del cual se toman decisiones que restringen la libertad. ¿Cómo resolver esta tensión? Es un problema complejo y no creo que tenga una solución fácil, pero una cosa que en mi opinión no ayuda desde luego a encontrarla es cerrar los ojos, e insistir en que RisCanvi se diseñó solo "para estimar probabilidades y orientar estrategias de intervención preventiva en el contexto de la intervención rehabilitadora", que no pretende "ser un oráculo", y que "no se limita a clasificar, sino que señala dianas de intervención ajustadas a las necesidades criminógenas de cada penado"[105]. Todo eso podrá ser, sin duda, muy cierto, pero también lo es que los niveles de riesgo que da RisCanvi se utilizan todos los días para tomar decisiones que suponen injerencias en derechos fundamentales. Y que este tipo de decisiones tienen unos requisitos de legitimidad que no son iguales a los del contexto en que se decide cuál es la mejor forma de diseñar y ejecutar el programa rehabilitador para cada persona privada de libertad.

4.4. ¿ES RISCANVI UNA HERRAMIENTA QUE MEJORA LA TRANSPARENCIA DE LAS DECISIONES, O UN ALGORITMO OPACO QUE VULNERA EL DERECHO DE ACCESO A LOS DATOS PERSONALES Y EL DERECHO DE DEFENSA?

Se ha criticado, y no sin razón, que las estimaciones de peligrosidad realizadas según el juicio clínico no son transparentes. La mente del juez, se dice (o la del profesional encargado de la valoración: psiquiatra, psicólogo, etc.) es una caja negra a la que no es posible acceder, por lo que no podemos saber los motivos reales que hay detrás de sus pronósticos (efecto *black box*). Aunque el juez motive su resolución, o el especialista detalle en su informe los argumentos que apoyan su conclusión, ésta podría obedecer (incluso de forma inconsciente) a prejuicios o suposiciones que no se mencionan expresamente pero que en realidad han resultado determinantes. Como destacaron con acierto Andrés Pueyo y Redondo Illescas, aquí la falta de transparencia es "propia del método, más que de la voluntad del clínico" y "no permite contrastar la fiabilidad de la decisión por medio de una réplica cuidadosa" (2007, p. 168). Frente a ello, se ha subrayado reiteradamente que las herramientas estructuradas mejoran la transparencia y la trazabilidad de las decisiones, porque contienen un listado de los factores que son tenidos en cuenta para valorar el riesgo, y los combinan de determinada forma

105. Andrés Pueyo (2025, pp. 30 y 31). En la p. 24 insiste: "[RisCanvi] No es tampoco un sistema para hacer un diagnóstico de la peligrosidad criminal en el sentido clásico del derecho penal español, pues no clasifica en categorías fijas y esencialistas, sino que evalúa riesgos dinámicos y modificables".

para obtener el resultado, de manera que se puede comprender cómo se ha llegado a la conclusión, y otras personas podrían replicar la valoración y verificar los resultados por sí mismas (Andrés Pueyo, 2025, pp. 23, 30, 31).

Ahora bien, paradójicamente, muchos de los algoritmos existentes en la actualidad no cumplen el requisito de la transparencia, pues su código no es público. Así ocurre en España con RisCanvi[106], pues, si bien se conocen cuáles son los factores que utiliza, no se sabe cómo se combinan, qué peso tiene cada uno de ellos en el resultado final, cuáles influyen en el nivel de riesgo calculado para cada escala, ni cuáles son los umbrales de discriminación que marcan la diferencia entre los distintos niveles de riesgo. Parafraseando la cita que hemos reproducido *supra* de Andrés Pueyo y Redondo Illescas, cabría decir que ahora el problema de transparencia no proviene del propio modelo, pues si este se abriera podría saberse exactamente cómo funciona, sino de la falta de voluntad de hacerlo público.

¿Cómo es que, a pesar de esta opacidad, los responsables de RisCanvi revindican que ha mejorado la transparencia y la trazabilidad de las decisiones sobre el riesgo? Pues porque sí es verdad que lo ha hecho, aunque solo para ciertas personas: quienes lo han diseñado o quienes supervisan su funcionamiento sí saben cómo funciona el algoritmo. Estas personas seguramente sí saben cuáles son los factores que tienen más peso en cada escala, y por tanto a qué se debe que en cada caso salga un resultado de riesgo alto o bajo. Además, como para valorar los factores de riesgo es necesario introducir en el sistema toda la información sobre las evidencias disponibles; estas personas conocen también cuál es la información con la que trabaja el algoritmo. Esto último también lo saben quienes forman parte de los equipos multidisciplinares, pues son los responsables de reunir esa información y meterla en el sistema. Sin embargo, estos profesionales no conocen el funcionamiento del algoritmo, y por eso no saben exactamente a qué factores se debe que los resultados sean los que son. Así, si a veces no están de acuerdo con el nivel de riesgo que da RisCanvi en alguna escala, pueden emitir una recomendación distinta de la que sería coherente con

106. Aunque RisCanvi no es, ni mucho menos, la herramienta más opaca entre las que se usan actualmente en el sistema de justicia penal español, más bien al contrario. También es opaco el algoritmo utilizado por la Secretaría General de Instituciones Penitenciarias para estimar el riesgo de quebrantamiento de los permisos penitenciarios, la TVR (aunque recientemente algunas investigaciones han abierto su código: Fanega, Fresneda-Portillo y Beltrán-Camacho, 2025; Civio, 2025). Y la opacidad es mayor aún respecto de VioGén, el algoritmo que utilizan las fuerzas y cuerpos de seguridad para estimar el riesgo de revictimización de las mujeres que denuncian violencia de género (cfr. Martínez Garay et al, 2024, pp. 26 y s., 157 y ss.).

el resultado del algoritmo (ya que este no es vinculante), pero no pueden explicar cuál es el motivo de la discrepancia porque no saben cómo calcula RisCanvi el nivel de riesgo.[107]

Tampoco los jueces, a quienes compete revisar las decisiones de las juntas de tratamiento, o tomar directamente algunas de ellas, conocen como funciona RisCanvi, ni las razones a las que se debe que en un caso el riesgo sea alto y en otro caso sea bajo. Les genera además bastante desconcierto recibir informes desde la prisión que incluyen los niveles de riesgo de RisCanvi pero sin ninguna explicación de por qué el algoritmo ha dado ese resultado, y más aún si el informe recomienda adoptar una decisión que no es coherente con el nivel de riesgo que indica RisCanvi (por ejemplo, un informe favorable a la concesión del permiso, cuando RisCanvi da niveles altos de riesgo de quebrantamiento y/o de reincidencia)[108]. Los fiscales además dan relevancia especialmente a los riesgos altos para oponerse a las decisiones favorables a la libertad, con independencia de que el informe que acompaña el nivel de riesgo explique por qué eso no debería ser determinante[109].

Y, por último, RisCanvi tampoco ha mejorado la transparencia para las personas directamente afectadas por las decisiones que se toman a partir de los niveles de riesgo: a los internos no se les explica que existe la herramienta ni cómo funciona. Los profesionales afirman que, si lo pidieran, habría que informarles de los niveles de riesgo, pero no está previsto en ningún caso que se les explique por qué en su caso el riesgo es el que es ni qué factores lo están condicionando, y mucho menos, que se les informe sobre la lógica del algoritmo. Tampoco los abogados de los internos tienen esta información, lo

107. Como señala Alemán Aróstegui: "En relación con el traslado de ese resultado a los órganos decisores y cómo este influye en sus decisiones, la mayoría explica que siempre se incluye el resultado del RisCanvi en los informes que se dirigen a los órganos judiciales y todos/as señalan que, normalmente, no se acompaña de ninguna explicación específica sobre ese resultado. Es interesante la reflexión que hace una psicóloga entrevistada, la cual afirma que no añaden explicaciones del resultado porque no saben cómo se genera, ya que no conocen cómo están ponderados los ítems en el algoritmo y esto hace que no puedan ni explicarlo ni cuestionarlo" (2023, p. 32).

108. Vid. Alemán Aróstegui (2023, pp. 32 y s).

109. "[…]Cuando se encuentran con casos en los que se aprecia dicha discordancia entre la valoración del equipo y el resultado del RisCanvi, dado que saben que el equipo tiene la posibilidad de modificar el nivel de riesgo, a veces, piden en su informe que se modifique el nivel de riesgo, y no consideran suficiente que se añada una explicación sobre los motivos por los que se mantiene ese nivel de riesgo. Insisten en que, mientras se mantenga el dato del resultado de riesgo alto, no es posible obviarlo, puesto que resulta contradictorio." (Alemán Aróstegui, 2023, p. 34).

que les dificulta argumentar frente a las decisiones restrictivas de libertad que se fundamentan en los niveles de riesgo[110].

En definitiva, RiCanvi ha mejorado la transparencia y la trazabilidad de las decisiones solo para un reducido grupo de personas: los responsables de la ejecución penitenciaria y los expertos que participan en el diseño y supervisión del algoritmo. Pero es muy discutible que la haya mejorado sensiblemente para el resto de trabajadores de prisión, ni para jueces y fiscales. Y desde luego no la ha mejorado en modo alguno para los penados ni para sus abogados, y tampoco para la ciudadanía, que está legítimamente interesada en el funcionamiento del sistema penitenciario, ni para los expertos externos a RisCanvi que tengan curiosidad por investigar su rendimiento y sus ventajas e inconvenientes. Hay una notable paradoja por tanto en esta cuestión de la transparencia, que se explica de nuevo por la doble dimensión de RisCanvi. Mientras la herramienta se considere exclusivamente un instrumento de trabajo interno de los profesionales penitenciarios, que sirve solo para diseñar los programas individualizados de tratamiento y facilitar el seguimiento de la evolución de los penados, es quizás lógico pensar que nadie más tiene por qué conocerla. Pero si RisCanvi es también uno de los argumentos sobre los que descansan decisiones que inciden muy intensamente en la libertad de los penados, es injustificable que estos últimos y sus abogados no puedan acceder a los detalles de esas valoraciones de riesgo. No tiene sentido tampoco que jueces y fiscales no tengan más información al respecto, para poder entender bien qué significan esos niveles de riesgo sobre los que están apoyando, siquiera sea en parte, sus decisiones. Y tampoco es de recibo que en una democracia los ciudadanos, y los expertos y académicos, no puedan conocer e investigar RisCanvi, que es al fin y al cabo un instrumento que usan las administraciones para ejercer y gestionar el poder público, en un ámbito tan sensible, y en el que ese poder es tan absoluto, como el de la ejecución de las penas privativas de libertad.

Dicho esto, es obligado reconocer que RisCanvi es, en comparación con otros algoritmos que utilizan las Administraciones públicas españolas en el ámbito de la justicia penal, uno de los más abiertos, pues el Centre d'Estudis Jurídics i Formació Especialitzada de la Generalitat de Catalunya lleva a cabo una política parcial de datos abiertos[111]. Pero esta política es solo parcial, y los códigos de los algoritmos siguen sin ser públicos. Incluso el resultado del informe de auditoría publicado sobre RisCanvi en 2024 recomienda que se

110. Martínez Garay et al. (2024, pp. 104 y s).

111. Vid. en https://cejfe.gencat.cat/ca/recerca/opendata/presons/ (último acceso el 20/12/2025).

haga público el código de RisCanvi y la documentación del modelo, y que se abran los datos del sistema, convenientemente anonimizados por supuesto. Para permitir a los profesionales del ámbito y a los internos entender la lógica que sigue el algoritmo, y para facilitar la revisión por parte de terceros y la detección de errores cuya corrección redunde en la mejora del sistema (Dribia Data & Direcció General d'Afers Penitenciaris, 2024, pp. 8, 53, 55).

5. CONCLUSIONES GENERALES

Primero: Articular la ejecución penitenciaria sobre el criterio del riesgo de reincidencia no es una obligación ni una necesidad; es una elección, una opción. Y el hecho de que existan herramientas para estimar ese riesgo (con mayor o menor grado de acierto, como ya hemos visto) no significa que sea la mejor opción.

El art. 25.2 de la Constitución no obliga a que la ejecución de las penas de prisión se diseñe tomando el riesgo de reincidencia como criterio principal. Lo permite, desde luego, pero igual que permite elegir otros criterios para organizar el sistema penitenciario. Dicho precepto dispone que "Las penas privativas de libertad y las medidas de seguridad estarán orientadas hacia la reeducación y reinserción social", pero de ahí no deriva que la búsqueda de la reinserción social tenga que pasar necesariamente por la medición del riesgo de reincidencia. La reinserción podría buscarse también por otras vías, por ejemplo, intentando reducir el riesgo de exclusión social que genera el sistema penitenciario, puesto que hay buenas razones para pensar que la exclusión social favorece la delincuencia, y que su reproducción o reforzamiento por parte del sistema penitenciario puede generar más delincuencia a medio y largo plazo. Necesitaríamos, entonces, herramientas para medir ese otro riesgo, que pueden construirse también[112]. Por otro lado, y en cualquier caso, nuestro Tribunal Constitucional se ha cansado de repetir que el art. 25.2 CE ni establece un derecho subjetivo del preso a la reinserción, ni obliga a que la reinserción social sea el único fin de las penas privativas de libertad, ni tampoco el principal. El art. 25.2 es perfectamente compatible, por tanto, con tratar de garantizar otras finalidades a través de la ejecución de las penas, como por ejemplo asegurar una ejecución no discriminatoria de las mismas.

Por ello, creo que la primera cuestión a considerar, antes de implementar en el sistema penitenciario vasco algo similar a Riscanvi, es pensar

112. Un ejemplo es RIMES; cfr. Díez Ripollés y García España (2020) y Díez Ripollés (2024).

cuánta importancia se le quiere dar al riesgo de reincidencia, y cuánta a otros criterios y a otros fines de la pena que son también legítimos. No estoy sugiriendo prescindir del riesgo: numerosos preceptos de nuestra legislación penal y penitenciaria lo establecen como un aspecto a tener en cuenta para tomar decisiones. Pero no es el único criterio a valorar, y hay un margen amplio para combinarlo con otras consideraciones también muy relevantes.

Segundo: Los algoritmos de valoración del riesgo de reincidencia y de violencia se han diseñado para estimar el riesgo de los individuos, pero nada impide utilizarlos para analizar críticamente el sistema y tomar medidas para mejorarlo. Esta es una ventaja de las estimaciones de riesgo sobre los "viejos" juicios clínicos de peligrosidad, que a mi juicio no ha sido suficientemente aprovechada hasta ahora.

Las valoraciones algorítmicas de riesgo extraen su riqueza del análisis de grandes conjuntos de datos, de diferentes muestras con muchos individuos. Esto plantea el problema de hasta qué punto es aplicable a cada individuo la probabilidad de riesgo que ha sido obtenida para el grupo[113], pero no es ese el problema al que ahora me quiero referir. Lo que quiero destacar es que al trabajar con grandes conjuntos de datos que se recogen y ordenan de manera sistemática, los estudios necesarios para diseñar herramientas como RisCanvi, y también los análisis que se hacen de sus resultados agregados, son en cierto sentido estudios epidemiológicos, porque analizan la frecuencia de la comisión de delitos y la variación de dicha frecuencia en grupos de personas, así como su asociación a una serie de factores. En Medicina, la epidemiología tiene relación con la clínica, pero también con la salud pública. Respecto de la clínica, los estudios epidemiológicos ayudan a formular hipótesis sobre los factores responsables de la aparición y evolución de las enfermedades, y sobre los tratamientos más adecuados, que luego pueden ser testados sobre individuos concretos para su curación. Respecto de la salud pública, la epidemiología ayuda en la prevención primaria, promoviendo prácticas que reduzcan la exposición de los sujetos sanos a los factores de riesgo que favorecen la aparición de las enfermedades. Piénsese solo en el caso del tabaco: una vez demostrado que es un factor de riesgo muy relevante para el cáncer de pulmón, las consecuencias no se han limitado a recomendar a los individuos que no se inicien en su consumo, o que lo abandonen (con ayuda de medicación específica si es necesario); también se han adoptado medidas como la prohibición de fumar en el lugar de trabajo, los transportes públicos y los bares y restaurantes.

113. Imrey y Dawid (2015).

En el sistema de justicia penal, las consecuencias que se han extraído hasta ahora de los estudios sobre estimaciones de riesgo se limitan a servir como apoyo para tomar decisiones sobre individuos concretos. Pero del conocimiento de los factores de riesgo asociados a la mayor o menor frecuencia de comisión de ciertos delitos en grupos de personas también pueden extraerse consecuencias para diseñar estrategias o políticas públicas que aborden esos factores de riesgo, no a nivel individual, sino poblacional. Los algoritmos de estimación del riesgo trabajan con la misma información que se utilizaba para hacer estimaciones clásicas de peligrosidad, pero tienen sobre estas la gran ventaja de que no son solo una "fotografía" de las necesidades criminógenas del individuo, sino también de las del sistema. Porque sistematizan un conjunto enorme de datos que, mirados de forma agregada, nos permiten ver, muchísimo mejor de lo que era posible sin los algoritmos, no solo las características individuales, sino también las dinámicas sociales y estructurales que favorecen la delincuencia y la reincidencia, pues muchos de los factores de riesgo son situaciones o circunstancias en las que el individuo se encuentra a consecuencia de condicionantes sociales.

Si la enfermedad mental —por seguir con el mismo ejemplo que he mencionado en el texto— está asociada a un incremento del riesgo de reincidencia violenta, es problemático utilizar ese dato para aplicar un régimen penitenciario más duro al individuo que la padece, porque puede generar discriminación (no solo porque padecer trastornos mentales sea un factor prohibido, sino también porque la asociación entre enfermedad mental y delito es muy compleja, y se debe en buena medida a la situación de mayor vulnerabilidad en que se encuentran estas personas en la sociedad y ante el propio proceso penal, y no tanto a que sean más violentas). Sin embargo, esa misma información puede utilizarse para hacer otras cosas que no plantean ningún problema de discriminación ni de vulneración de derechos fundamentales, como, por ejemplo, para repensar el sistema de atención a la salud mental en prisión, y el modelo de los hospitales psiquiátricos penitenciarios, y suplir de una vez las escandalosas carencias que uno y otro evidencian[114]. O también para mejorar el sistema público de salud mental, para todos los ciudadanos. De ello derivaría probablemente un efecto de prevención primaria, y también terciaria, pues como es sabido, la reinserción efectiva depende en buena medida de que al salir en libertad la persona cuente con un entorno y unos recursos favorables que le permitan consolidar la mejora que haya podido empezar a construirse en la prisión.

114. Sobre todo ello cfr. Calcedo-Barba, Antón-Basanta y Paz Ruiz (2023), Grupos de Estudios de Política Criminal (2023) y Martínez Garay (2024).

Tener información sobre los factores de riesgo de la reincidencia, obtenidos a partir de análisis sistemáticos de la conducta de muchos individuos, no obliga a utilizar esa información (solo) para tomar decisiones sobre los individuos; permite utilizarla también (¿o sobre todo?) para actuar sobre el sistema. Y cuando se utiliza para cambiar el sistema, muchas de las tensiones entre intervención y garantías a las que me he referido en este trabajo se difuminan.

Tercero: En el diseño, implementación y seguimiento de cualquier herramienta de estimación del riesgo parecida a RisCanvi han de intervenir no solo expertos en criminología y en ciencia de datos, sino también juristas especializados en derechos fundamentales, discriminación, protección de datos personales, derecho penitenciario y normativa reguladora de sistemas automatizados de decisión.

Como he repetido con insistencia en las páginas precedentes, RisCanvi (y las herramientas similares a él) tiene dos caras: es un instrumento de intervención al servicio de la rehabilitación y reinserción social, pero también un instrumento de control y limitación de derechos fundamentales al servicio del mantenimiento de la seguridad pública. Esta segunda dimensión no ha recibido la atención que merece ni cuando RisCanvi se creó e implementó, ni tampoco en las revisiones y auditorías de las que ha sido objeto hasta ahora. Hay razones que seguramente pueden explicar por qué esto ha sido así: la formación académica y experiencia profesional de quienes lideraron el proyecto era mucho más próxima a la psicología y a la criminología que al derecho; el contexto normativo era muy diferente hace 15 años; había mucha menos información acerca de los peligros que los algoritmos pueden suponer para los derechos fundamentales, y también mucha menos conciencia sobre estos riesgos, etc. Sea como fuere, no haber reconocido que RisCanvi tiene también, lo queramos o no, una dimensión de injerencia en los derechos fundamentales, ha hecho que la conversación (los debates acerca de su capacidad predictiva, de los sesgos y los efectos discriminatorios que puede producir, y de cuál es el nivel de transparencia que han de tener estos sistemas) haya parecido a veces por desgracia un diálogo de sordos, en el que a pesar de que todas las partes tienen (parte de) razón, paradójicamente ha resultado muy difícil llegar a los consensos que ineludiblemente hay que construir si se quieren seguir utilizando este tipo de herramientas en la gestión penitenciaria.

Por ello, sería imperdonable en el contexto actual actuar igual que cuando RisCanvi se implantó, y es imperativo atender desde el principio a las dos caras. Ser consciente de que están en juego los derechos a la intimidad,

a la defensa, a la no discriminación, y que hay que cumplir con todas las obligaciones que en la actualidad derivan de la normativa de protección de datos y de la que regula el uso de inteligencia artificial y de sistemas automatizados de decisión por parte de las Administraciones Públicas[115]. Ello requiere que juristas expertos en estos ámbitos participen desde el primer momento en el diseño y en la implementación, y también en la revisión y supervisión continua del sistema una vez se ponga en marcha.

Cuarto: El momento decisivo no es tanto el de estimación del riesgo por parte del algoritmo, sino aquel en el que una persona física toma la decisión con ayuda de ese resultado.

Para conocer y evaluar los efectos reales de la implantación de un algoritmo es imprescindible saber cómo lo van a utilizar las personas. Los sistemas como RisCanvi no pueden ser vinculantes y requieren que la decisión final esté en manos de personas físicas[116]. Pero un algoritmo diseñado para funcionar solo como apoyo en la toma de decisiones puede convertirse en materialmente decisivo si quienes lo utilizan confían por sistema en sus recomendaciones sin cuestionarlas ni tomar en consideración ningún otro parámetro. Hay que contar con el sesgo de automatización. Por otro lado, cabe también que ocurra lo contrario: si no se establecen pautas sobre cómo manejar la información que da el algoritmo, si esto se fía por completo al criterio personal de cada individuo, ese margen de actuación humana podría

115. La normativa vigente es además muy distinta de la que regía en el momento en que RisCanvi se creó, y algo más ambiciosa en cuanto a protección de derechos en estos ámbitos, aunque aún no alcance las cotas que serían deseables. Hay que tener muy en cuenta, por un lado, el Reglamento General de Protección de Datos y la Ley Orgánica 7/2021, de 26 de mayo, de protección de datos personales tratados para fines de prevención, detección, investigación y enjuiciamiento de infracciones penales y de ejecución de sanciones penales. También la Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno, y la importante Sentencia del Tribunal Supremo nº 1119/2025, de 11 de septiembre, dictada en el caso "Bosco", que ha interpretado de manera restrictiva algunos de los límites al derecho de acceso contenidos en esa ley. Y, cuando entre en vigor, el Reglamento de inteligencia artificial de la UE, que no es descartable que sea de aplicación a RisCanvi en la medida en que —como sabemos desde la publicación del Informe Tiresias— al menos las escalas de riesgo de reincidencia general y de quebrantamiento de condena son modelos de regresión logística, que se consideran sistemas de inteligencia artificial (Dribia Data & Direcció General d'Afers Penitenciaris, 2024, p. 24), y habrá que ver hasta qué punto entran en las definiciones del Reglamento (cfr.Comisión Europea, 2025).

116. Así lo exige la normativa vigente: art. 14 LO 7/2021, art. 14 del Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial.

dar al traste con la homogeneización de prácticas y la objetivación de criterios que se persigue con la introducción de las herramientas estandarizadas de estimación del riesgo. La interacción entre los algoritmos y las personas es complicada: los modelos pueden influir en la percepción de riesgo de las personas y condicionar las prácticas de trabajo, pero también las personas pueden manipular los modelos si no confían en sus resultados (por ejemplo, cambiando la información de entrada hasta conseguir que el sistema dé el resultado que la persona considera correcto)[117].

Las personas que van a tener que utilizar algoritmos o herramientas estructuradas de riesgo deben tener formación: saber para qué se han implementado, qué objetivos se persiguen, cómo funcionan, sus debilidades y sus puntos fuertes, y disponer de criterios para saber qué importancia dar a los resultados. Debe recogerse de manera sistemática información sobre la experiencia de los profesionales con el uso del modelo para poder conocer cómo impacta este en la realidad, y tenerse en cuenta su opinión en la supervisión del sistema.

6. RECOMENDACIONES CONCRETAS SI SE QUIERE IMPLANTAR ALGUNA HERRAMIENTA PARECIDA A RISCANVI EN EL SISTEMA PENITENCIARIO VASCO

6.1. MANTENER LA PROPIEDAD Y EL CONTROL SOBRE LOS MODELOS Y SOBRE LOS DATOS

Esto es algo que en RisCanvi se ha hecho a mi juicio muy bien[118]: asegurarse de que la Administración mantiene la propiedad y el control sobre el sistema y sobre los datos con los que se alimenta y que genera. Esto puede lograrse o bien construyendo la herramienta —como en RisCanvi— gracias a una colaboración entre la Administración pública y expertos académicos, o

117. Alemán Aróstegui (2023) ha hecho un estudio cualitativo muy interesante sobre la aplicación práctica de RisCanvi por diversas personas que lo han de utilizar en su trabajo. Otro estudio reciente compara las estimaciones de riesgo realizadas con y sin ayuda de una versión *machine learning* de RisCanvi, y concluye que son mejores cuando cuentan con el apoyo de la herramienta (Portela, M., Castillo, C., Tolan, S. et al.2025). En ocasiones la introducción de recomendaciones basadas en estimaciones de riesgo ha producido efectos contrarios a los deseados (Stevenson y Doleac, 2019).

118. Sin perjuicio de que el informe Tiresias indica algunas debilidades a este respecto, como que hay una excesiva dependencia en RisCanvi del personal que trabaja en el proveedor externo de servicios informáticos (Dribia Data & Direcció General d'Afers Penitenciaris, 2024, pp. 52 y s.). También es la Administración pública (Ministerio del Interior) quien tiene la propiedad sobre la TVR.

bien encargándola a un proveedor externo, pero asegurándose en este caso de redactar los pliegos de contratación de manera que la Administración sea la propietaria del sistema, tenga la propiedad intelectual exclusiva sobre los modelos y el control sobre todos los datos.

6.2. NECESIDAD DE EVALUACIÓN PREVIA, CONTINUA, Y POR RESULTADOS AGREGADOS

- Me parece necesario hacer una evaluación del funcionamiento del sistema penitenciario antes de implementar el algoritmo, identificar cuáles son las áreas que se quieren mejorar o los problemas que se desea resolver, y definir los objetivos a conseguir. Si no se tiene un punto de referencia, es complicado poder evaluar después si el uso del algoritmo ha mejorado o empeorado lo que había antes de él.

- Hay que garantizar también la supervisión continua de los resultados del modelo a medida que este va funcionando, e integrar en esa supervisión la opinión de todos los actores interesados (véase el siguiente apartado: transparencia). Deben evitarse situaciones como la de la Tabla de Variables de Riesgo, que no se ha actualizado desde que se introdujo en los años 90.

- Es frecuente que los algoritmos o protocolos estructurados de valoración de riesgo se introduzcan a consecuencia de casos dramáticos en los que parece que los individuos responsables no fueron capaces de identificar un riesgo que luego terminó en tragedia, y con el algoritmo se busca impedir que esos "errores" vuelvan a ocurrir (véase el inicio de este trabajo). Sin embargo, las estimaciones de riesgo son solo probabilidades. Todo sistema tiene siempre unas tasas de error, que se pueden optimizar y diseñar para minimizar errores de un tipo o de otro, pero no se puede garantizar el riesgo cero ni el 100% de acierto. Un algoritmo que, por ejemplo, consiguiera reducir la tasa de suicidios en prisión desde el 1 por mil al 0,5 por mil debería en principio valorarse favorablemente, y lo mismo cabe decir de la tasa de quebrantamiento de permisos, porque estaría mejorando los resultados agregados, a pesar de que de tanto en tanto un preso se quite la vida o un interno de permiso cometa un crimen muy grave. Es necesario resistir la presión de considerar como un fracaso del sistema cada nuevo incidente no evitado, porque de lo contrario se puede entrar en una espiral de endurecimiento no justificado. Por supuesto, cada "error" debe revisarse cuidadosamente, por si se

han hecho las cosas mal, y para aprender de cara al futuro, pero en muchos casos no se tratará de errores que hubieran podido evitarse, sino de las tasas de error inherentes al sistema[119], que deben ser conocidas, debatidas abiertamente y comprendidas[120].

6.3. TRANSPARENCIA

Es necesario hacer un modelo transparente: tanto para los trabajadores que tendrán que utilizarlo, como para las personas internas en los centros penitenciarios y sus abogados, para los jueces y fiscales, y también para el público en general y la comunidad académica en particular. Y del que se conozcan no sólo el código fuente y la documentación del algoritmo, sino también los conjuntos de datos garantizando, por supuesto, la imprescindible anonimización.

- La política parcial de datos abiertos que ya sigue la Generalitat es un ejemplo de valentía y de rigor, en un contexto como el español que hasta ahora se ha caracterizado por una opacidad absoluta —y absolutamente injustificada— de los sistemas utilizados por las administraciones públicas. Es un ejemplo que demuestra que se puede hacer sin comprometer ni la intimidad de los presos ni la integridad y seguridad del algoritmo.

- Un modelo transparente incrementa la confianza de las personas en una herramienta que saben cómo funciona, y sobre la cual pueden tener por tanto un mejor juicio crítico y unas expectativas razonables.

- La transparencia permite el análisis por parte de terceros, lo que posibilita que se detecten fallos y sea posible corregirlos, o que se sugieran mejoras; en definitiva, redunda en la mejora de la calidad del sistema.

- La transparencia implica también formación, para todos los actores que tengan que utilizar la herramienta al tomar decisiones, lo cual

119. Y en muchas ocasiones el problema no estará en el nivel de riesgo estimado, sino en las medidas de precaución que se hayan implementado a raíz del mismo. El tema tiene, obviamente, una complejidad en la que no es posible entrar aquí.

120. Una de las posibles cuestiones a debatir es la eventual responsabilidad que quepa exigir a la Administración por los "fallos", ya sean errores evitables y/o tasas estadísticas de error, o si es conveniente / justo / viable (o no) prever indemnizaciones para algunos de estos fallos (como existen para las víctimas de algunos delitos). Cfr. una reflexión detallada sobre esta problemática, no en relación con RisCanvi sino con VioGén, en Molina Sánchez (2025).

incluye no solo al personal de prisiones que trabaje día a día con él, sino también a otros actores interesados: abogados, jueces y fiscales.

– La reivindicación de la transparencia no es una manía de unos pocos académicos recalcitrantes, sino que cada vez más se asume como una exigencia elemental de garantía de los derechos fundamentales y como un instrumento habitual y normal de control del ejercicio del poder por parte de las administraciones públicas en las sociedades democráticas[121]. El informe Tiresias con los resultados de la auditoría de RisCanvi también lo recomienda, como hemos visto. La STS n.º 1119/2025, de 11 de septiembre, dictada en el caso "Bosco", ha obligado ya a una Administración Pública a abrir un algoritmo y sienta un importante precedente sobre la importancia del derecho de acceso a la información pública y la necesidad de interpretar de forma restrictiva las limitaciones al mismo, que siempre requieren justificación suficiente. Y hay algoritmos de estimación del riesgo de reincidencia de código abierto que se usan en otras jurisdicciones sin ningún problema de seguridad ni de confidencialidad[122].

– Si se opta por utilizar modelos de inteligencia artificial, la necesidad de garantizar la transparencia hará siempre preferibles modelos más sencillos y explicables frente a otros más oscuros. Solo debería ser admisible implantar estos últimos tras un examen riguroso de proporcionalidad: sólo valdría la pena sacrificar un mayor grado de explicabilidad si los modelos más opacos y complejos demuestran que mejoran de manera notable tanto la capacidad predictiva como la mitigación de sesgos y discriminaciones frente a modelos más sencillos.

6.4. CONFLICTO CON PROTECCIÓN DE DERECHOS

Muchos de los problemas que generan los algoritmos de estimación del riesgo derivan de la tensión a la que me he referido a lo largo de este trabajo entre la finalidad de mejorar la gestión de la intervención rehabilitadora en prisión, y la de proporcionar estimaciones fiables de riesgo sobre las cuales adoptar de manera bien fundamentada decisiones limitativas de derechos. No es fácil dar soluciones para todos estos problemas, y todavía

121. Boix Palop (2022 y 2025).

122. Por ejemplo el Public Safety Assessment (PSA), véase la información en https://www.advancingpretrial.org/how-the-psa-works/?tab=tab-factor-weights (último acceso el 20/12/2025), o OxRec en sus diferentes variantes (Fazel, Chang, Fanshawe et al, 2016; cfr. asimismo https://oxrisk.com/oxrec-9-3/, último acceso el 20/12/2025).

no se han ofrecido muchas propuestas, en parte debido precisamente a que se ha tardado en reconocer que estos problemas existen. Pero desde luego es ineludible empezar a pensar en cómo abordarlos, lo que requiere por parte de los juristas un esfuerzo de imaginación, ya que se trata de un escenario distinto al que hemos estado acostumbrados. A continuación propongo algunas ideas, que deben entenderse solo como puntos de partida para fomentar esta necesaria discusión, y no como propuestas acabadas.

– Diferenciar entre riesgos auto— y heterodirigidos.

  Pienso que en el sistema penitenciario la prevención de la violencia autodirigida (la prevención de suicidios y autolesiones) tiene claramente una dimensión tuitiva de la persona presa y no limitativa de derechos fundamentales. Por ello, considero que en los algoritmos diseñados para estimar este tipo de riesgo no es problemático, o no lo es tanto como en otros, utilizar factores de riesgo que constituyen motivos de discriminación prohibida, precisamente porque no se va a discriminar al preso (adoptando decisiones que le son desfavorables), sino que se le va a proteger incrementando la atención, precaución y seguimiento. Por la misma razón, tampoco considero tan problemático en estos casos que los números de falsos positivos sean elevados: las preocupaciones por la dimensión garantista no tienen tanto sentido aquí (salvo que redunden en medidas de protección muy drásticas). Al contrario, la prioridad debe ser minimizar los falsos negativos, habida cuenta de que la Administración penitenciaria tiene en este caso una posición de garante de la vida y de la integridad física de los internos.

– Establecer objetivos realistas.

  La prevalencia de algunas de las conductas cuyo riesgo estima RisCanvi es muy baja, especialmente en el caso del quebrantamiento de permisos. Esto hace necesario plantearse hasta qué punto vale la pena introducir un algoritmo para rebajarla más aún, ya que cuando un fenómeno tiene una prevalencia muy baja, cualquier intento de prevenir su aparición generará porcentajes muy elevados de falsos positivos. El propio informe Tiresias lo advierte, cuando explica que, puesto que la tasa de quebrantamiento de permisos es solo del 0,5 % de los permisos concedidos, un algoritmo que estimase bajo riesgo en todos los casos tendría una tasa de acierto del 99,5 %. No es que ese 0,5 % no se pueda rebajar, claro que se puede; pero si se hace con el criterio del riesgo, será al precio de denegar por riesgo

alto muchos permisos que en la actualidad se conceden y en los que no ocurre ninguna incidencia. La cuestión es: habida cuenta de que los permisos son un instrumento esencial del tratamiento y de la preparación para la vida en libertad, ¿merece la pena restringirlos a muchos internos que habrían hecho un buen uso de ellos solo para conseguir reducciones mínimas en los quebrantamientos cuando la tasa actual ya es muy baja?[123] Esto podría quizá "solucionar" un problema en el corto plazo, pero a costa de crear otro en el medio y largo plazo: si aumenta el número de personas que por no haber disfrutado de permisos tienen después menos vínculos con su entorno familiar y social, y cuando salen en libertad terminan volviendo a delinquir por no ser capaces de encontrar una red de apoyo que les ayude a subsistir en libertad por medios lícitos, podría acabar incrementándose la reincidencia. Hay que tener en cuenta no solo los efectos inmediatos de las decisiones individuales, sino los efectos a medio plazo que produce en el sistema la suma agregada de esas decisiones individuales.

– Usar el algoritmo para reformar el sistema cuando esto redunde en una mejor garantía de los derechos.

  Con ello me refiero a que el algoritmo, y especialmente los datos que se obtengan de su supervisión continua una vez implementado, puede evidenciar que el marco normativo actual restringe en exceso la libertad sin proporcionar a cambio ningún beneficio en términos de prevención. Por ejemplo: las investigaciones llevadas a cabo sobre RisCanvi evidencian que la salida progresiva a la libertad, a igual nivel de riesgo inicial, correlaciona con un nivel de reincidencia menor. Si esto es así, los datos del algoritmo deberían utilizarse para reclamar una reforma del sistema penal y penitenciario que elimine las barreras que impiden este acceso escalonado a la libertad, siempre que no haya otras consideraciones relevantes que justificasen mantenerlos. Esto significaría replantearse los límites que el Código Penal impone para el acceso al tercer grado en algunos tipos de delitos. Otro ejemplo: los algoritmos ofrecen siempre niveles de riesgo graduables en un continuo. RisCanvi tiene 3: bajo, medio y alto, pero las probabilidades son continuas y los niveles podrían ser

123. Huelga decir que estoy simplificando mucho una cuestión que es muy compleja, no solo porque la decisión final sobre los permisos recae en personas que utilizan los niveles de riesgo de RisCanvi solo como apoyo, sino también porque los criterios para la concesión de los permisos son complejos y problemáticos (cfr. Larrauri, 2025).

4, 5 o más. Sin embargo, muchas decisiones que se toman sobre los presos son dicotómicas: permiso sí o no, tercer grado sí o no, libertad condicional sí o no. Pues bien, quizá en vez de utilizar los niveles de riesgo para decidir si sí o si no, el sistema podría diseñarse de manera que algunas de estas decisiones fueran automáticas una vez cumplidos determinados requisitos legales ajenos al riesgo (determinado tiempo previo pasado en prisión, ausencia de expedientes disciplinarios, padecer determinado grado de enfermedad, etc.), y que el nivel de riesgo sirviera para graduar el nivel de intensidad en el acompañamiento o supervisión, o el tipo de medidas de seguimiento más adecuadas. Esto reduciría en cierta medida el problema de los falsos positivos y negativos porque todos podrían acceder a los beneficios con independencia del riesgo, y este se utilizaría para establecer el nivel de vigilancia o acompañamiento, que es también graduable, como el riesgo[124].

– Construir escalas diferentes para las finalidades tratamentales y para las de restricción de derechos.

  Quizá no todos los factores que hoy por hoy integran RisCanvi tienen la misma relevancia para la estimación de la reincidencia futura. El informe Tiresias apunta a que varios de los que actualmente conforman los algoritmos de las diferentes escalas son poco predictivos en este sentido[125]. Cabría quizá plantearse, por ello, eliminar los que resultan más problemáticos en términos de derechos fundamentales y comprobar si la capacidad predictiva del modelo sigue siendo lo suficientemente buena. Estos algoritmos "reducidos" serían los que se utilizarían para la predicción del riesgo cuando sobre esa base se tengan que tomar medidas restrictivas de la libertad. Al mismo tiempo, podrían mantenerse en paralelo algoritmos más extensos, que incluyeran todos los factores que, si bien resultan una base discutible sobre la cual tomar decisiones restrictivas de derechos, sí son necesarios para gestionar el riesgo: para diseñar los programas individualizados de tratamiento. Quizá esta propuesta no sea fácil de llevar a la práctica, por lo imbricadas que están ambas dimensiones en el sistema penitenciario. Podría también ocurrir que al excluir algunas variables la capacidad predictiva baje en exceso, y el algoritmo se vuelva inservible, aunque esto no necesariamente

124. He desarrollado una propuesta en este sentido, aplicada a la revisión de la pena de prisión permanente, en Martínez Garay (2023).

125. Vid. Dribia Data & Direcció General d'Afers Penitenciaris (2024, pp. 6 y s., 47 y s.).

tiene que pasar[126]. En cualquier caso, creo que merece la pena al menos pensar si podría ser viable lo que sugiero, de la manera en que lo planteo o en alguna otra diferente.

– Hacer un juicio de proporcionalidad explícito sobre los factores de riesgo problemáticos.

  Si se mantienen factores problemáticos en los cálculos de riesgo sobre los que se apoyan decisiones restrictivas de derechos, la presencia de esos factores debe ser analizada en términos de proporcionalidad. El art. 4.2 de la ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación, establece que "No se considera discriminación la diferencia de trato basada en alguna de las causas previstas en el apartado 1 del artículo 2 de esta ley derivada de una disposición, conducta, acto, criterio o práctica que pueda justificarse objetivamente por una finalidad legítima y como medio adecuado, necesario y proporcionado para alcanzarla." Es decir, que son admisibles las diferencias de trato por factores prohibidos, pero siempre que estén justificadas y sean proporcionadas. Debe estudiarse, entonces, cuánto aporta cada factor a la estimación de cada conducta problemática, cuánto empeoraría la predicción si se prescindiera de él, y, si no es posible eliminarlo o sustituirlo por otro menos sensible, valorar si lo que ganamos en prevención de la delincuencia compensa o no el efecto discriminatorio que produce. En EEUU por ejemplo la raza es un factor completamente prohibido, que no puede figurar en ninguna herramienta, aunque con ello se redujeran las tasas de errores. Hay que valorar si en nuestro contexto este puede ser el caso de algunos factores, como las enfermedades, la discapacidad, las preferencias sexuales u otros. Y, si no, abrir la discusión sobre hasta qué punto la prevención puede justificar según qué discriminaciones.

– Discriminación indirecta.

  Las consideraciones que he hecho en los dos párrafos anteriores no eliminarían el problema de la discriminación indirecta, que es la

126. Como afirman Fazel, Sariaslan y Fanshawe (2022, p. 398), "Parte de la justificación de estas herramientas largas y complejas es que identifican necesidades, lo que puede constituir la base para la gestión del riesgo. El problema es que los predictores y las necesidades no son intercambiables, y una herramienta que incluya muchas necesidades no necesariamente dará lugar a una mejora en su rendimiento predictivo, y hacerlo puede restar precisión a sus predicciones."

que se produce cuando las estimaciones de riesgo dan resultados diferentes para grupos sensibles, a pesar de que las características que los individualizan no estén expresamente contempladas como factores de riesgo. Para enfrentar este problema es necesario, en primer lugar, estudiarlo: identificar si hay diferencias en los resultados o no y, en su caso, cuantificarlas. Después, determinar a partir de qué punto (de qué porcentaje de trato dispar) podemos considerar que existe una discriminación prohibida[127]. Y, en caso de que efectivamente se dé, habrá que pensar estrategias para abordar esta situación, estrategias que probablemente pasen no solo por el rediseño de los algoritmos, sino también por el del contexto en el que operan (siguiendo con el ejemplo de la enfermedad mental: puede ser necesario reformar la atención a la discapacidad a lo largo del proceso penal, la atención a la salud mental en prisión, transferir esas competencias a las comunidades autónomas, etc.).

7. REFERENCIAS

Alemán Aróstegui, L. (2023). El uso del RISCANVI en la toma de decisiones penitenciarias. *Estudios Penales y Criminológicos*, número especial 44 (Inteligencia artificial y sistema penal).

Andrés-Pueyo, A. (2017). Predicción de la reincidencia penitenciaria en Cataluña por medio del RisCanvi. En E. Orts Berenguer, A. Alonso Rimo & M. Roig Torres (Dirs.), *Peligrosidad criminal y Estado de Derecho* (pp. 371-390). Tirant lo Blanch.

Andrés-Pueyo, A., Arbach-Lucioni, K., & Redondo, S. (2018). The RisCanvi: A new tool for assessing risk for violence. En J. P. Singh et al. (Eds.), *Handbook of recidivism risk/needs assessment tools* (pp. 255-268). John Wiley & Sons.

Andrés-Pueyo, A. (2025). El RisCanVi: Entre la predicción algorítmica y la decisión profesional para la prevención de la reincidencia penitenciaria. *ReCRIM: Revista de l'Institut Universitari d'Investigació en Criminologia i Ciències Penals de la UV*, número especial 25, 21-35.

Andrés-Pueyo, A., & Redondo Illescas, S. (2007). Predicción de la violencia: Entre la peligrosidad y la valoración del riesgo de reincidencia. *Papeles del Psicólogo*, 28(3), 157-173.

127. Cfr. consideraciones interesantes, a la luz de la jurisprudencia europea y de la propuesta de reglamento de inteligencia artificial, en Soriano Arnanz (2023).

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: *The state of the art. Sociological Methods & Research*, 50(1), 3-44. https://doi.org/10.1177/0049124118782533.

Boix Palop, A. (2022). Transparencia en la utilización de inteligencia artificial por parte de la Administración. *El Cronista del Estado Social y Democrático de Derecho*, 100 (Ejemplar dedicado a: Inteligencia artificial y derecho), 90-105.

Boix Palop, A. (2025). Transparencia y publicidad en el uso de algoritmos y de Inteligencia Artificial en el Reglamento europeo de Inteligencia Artificial: especial referencia a las obligaciones de transparencia referidas a los poderes públicos. En A. J. Sánchez Sáez (Ed.), *Claves del reglamento europeo de inteligencia artificial: análisis y comentarios* (pp. 39-64). Aranzadi La Ley.

Briz-Redón, Á., & Montes, F. (2022). *Análisis de los resultados del protocolo RisCanVi*. Universitat de València. https://www.uv.es/montes/informe%20riscanvi/informe.pdf.

Calcedo-Barba, A., Antón-Basanta, J., & Paz Ruiz, S. (2023). *Libro blanco sobre la atención sanitaria a las personas con trastornos mentales graves en los centros penitenciarios de España.* SEPL & SESP.

Capdevila Capdevila, M., et al. (2023). *Tasa de reincidencia penitenciaria 2020*. CEJFE.

Civio. (2025). Las prisiones españolas usan un algoritmo sin actualizar desde 1993 para decidir sobre permisos de salida. https://civio.es/justicia/2025/02/26/las-prisiones-espanolas-usan-un-algoritmo-sin-actualizar-desde-1993-para-decidir-sobre-permisos-de-salida/.

Comisión Europea. (2025). *Comunicación de la Comisión: Directrices relativas a la definición de sistema de inteligencia artificial establecida en el Reglamento (UE) 2024/1689. C(2025) 5053 final.*

Díez Ripollés, J. L. (2024). El modelo politicocriminal de la inclusión social. *Boletín Criminológico,* 30, número especial 230, 1-44.

Díez Ripollés, J. L., & García España, E. (2020). RIMES: Un instrumento de comparación de políticas criminales nacionales desde la exclusión social. *Política Criminal: Revista Electrónica Semestral de Políticas Públicas en Materias Penales*, 15(30).

Dribia Data & Direcció General d'Afers Penitenciaris. (2024). *Informe Tiresias: Auditoría del algoritmo RisCanVi.* Generalitat de Catalunya. https://

justicia.gencat.cat/ca/ambits/reinsercio_i_serveis_penitenciaris/serveis_penitenciaris/RisCanvi/.

Fanega, M., Fresneda-Portillo, C., & Beltrán-Camacho, Á. (2025). Publicación del algoritmo penitenciario TVR. *Revista Española de Investigación Criminológica*, 22(1), e917.

Fazel, S., Chang, Z., Fanshawe, T., et al. (2016). Prediction of violent reoffending on release from prison: derivation and external validation of a scalable tool. *The Lancet Psychiatry*, 3(6), 535-543.

Fazel, S., Sariaslan, A., & Fanshawe, T. (2022). Towards a more evidence-based risk assessment for People in the Criminal Justice System: the Case of OxRec in the Netherlands. *European Journal of Criminal Policy Research*, 28, 397-406.

Fazel, S., Burghart, M., Fanshawe, T., et al. (2022). The predictive performance of criminal risk assessment tools used at sentencing: Systematic review of validation studies. *Journal of Criminal Justice*, 81, 101902.

Grupo de Estudios de Política Criminal. (2023). *Una propuesta alternativa para un nuevo régimen penal aplicable a las personas con enfermedad mental o discapacidad intelectual.* https://politicacriminal.es/una-propuesta-alternativa-para-un-nuevo-regimen-penal-aplicable-a-las-personas-con-enfermedad-mental-o-discapacidad-intelectual.

Imrey, P. B., & Dawid, A. P. (2015). A commentary on statistical assessment of violence recidivism risk. *Statistics and Public Policy*, 2(1), 1-18.

Julià Pijoan, M. (2023). RisCanvi: ¿Un protocolo de discriminación algorítmica? Reflexiones desde la legislación y jurisprudencia europea. En F. Bueno de Mata (Coord.), *Justicia y derecho en datos* (pp. 123-141). Tirant lo Blanch.

Karimi-Haghighi, M., & Castillo, C. (2022). *Quantitative analysis of disparate effects of RisCanvi for estimating the risk of violent recidivism. Technical Report*, Web Science and Social Computing Research Group, Universitat Pompeu Fabra.

Larrauri, E. (2025). Los permisos de salida, un primer paso al régimen abierto y las dificultades de los colectivos de inmigrantes para acceder a ellos. *ReCRIM: Revista de l'Institut Universitari d'Investigació en Criminologia i Ciències Penals de la UV*, número especial 25, 123-138.

Lawson, S. G., Narkewicz, E. L., & Vincent, G. M. (2024). Disparate impact of risk assessment instruments: A systematic review. *Law and Human Behavior*, 48(5-6), 427-440.

Martínez Garay, L. (2016). Errores conceptuales en la estimación de riesgo de reincidencia: La importancia de diferenciar sensibilidad y valor predictivo, y estimaciones de riesgo absolutas y relativas. *Revista Española de Investigación Criminológica*, 14.

Martínez Garay, L., & Montes Suay, F. (2018). El uso de valoraciones del riesgo de violencia en Derecho Penal: algunas cautelas necesarias. *InDret*, 2.

Martínez Garay, L. (2020). Evidence-based sentencing y evidencia científica: A la vez, algunas consideraciones sobre las "políticas basadas en la evidencia" y el derecho penal. *Teoría y Derecho: revista de pensamiento jurídico*, 28, 238-276.

Martínez Garay, L. (2023). Revisión con riesgo bajo, y también con riesgo alto: razones para que las valoraciones de riesgo no impidan la revisión de la prisión permanente. *Revista General de Derecho Penal*, 39.

Martínez Garay, L. (2024). Un nuevo régimen penal aplicable a las personas con enfermedad mental o con discapacidad intelectual: la propuesta alternativa del Grupo de Estudios de Política Criminal. En V. Cervelló Donderis et al. (Eds.), *Salud mental y género: el debate entre el punitivismo y la despenalización* (pp. 413-464). Tirant lo Blanch.

Martínez Garay, L., et al. (2024). *Three predictive policing approaches in Spain.* Publicacions de la Universitat de València. https://omp.uv.es/index.php/PUV/catalog/book/626

Molina Sánchez, M. (2025). *La responsabilidad patrimonial de la Administración por el uso de VioGén en la prevención de la violencia de género*. [Tesis doctoral, Universitat de València]. https://producciocientifica.uv.es/documentos/687545586a0dd13b65b2bde9

Portela, M., Castillo, C., Tolan, S., et al. (2025). A comparative user study of human predictions in algorithm-supported recidivism risk assessment. *Artificial Intelligence and Law*, 33, 471-517.

Rovira, M., Larrauri, E., & Alarcón, P. (2018). La concesión de permisos penitenciarios. *Revista Electrónica de Ciencia Penal y Criminología*, 20-02, 1-26.

Soriano Arnanz, A. (2023). Creating non-discriminatory Artificial Intelligence systems: balancing the tensions between code granularity and the general nature of legal rules. IDP: *Revista de Internet, Derecho y Política*, 38.

Stevenson, M. T., & Doleac, J. L. (2019). Algorithmic risk assessment in the hands of humans. IZA-Institute of Labor Economics Discussion Paper No. 12853. https:// www.iza.org/publications/dp/12853/algorithmic-risk-assessment-in-the-hands-of-humans

Subdirecció General de Programes de Rehabilitació i Sanitat. (2019). *Manual d'aplicació del protocol de valoració RisCanVi. Generalitat de Catalunya*. https://justicia.gencat.cat/ca/ambits/reinsercio_i_serveis_penitenciaris/serveis_penitenciaris/RisCanvi/

Tonry, M. (2019). Sentencing and prediction: Old wine in old bottles. En J. de Keijser, J. Roberts & J. Ryberg (Eds.), *Predictive sentencing* (pp. 269-298). Hart Publishing.

Viljoen, J. L., et al. (2025). *Are risk assessment tools more accurate than unstructured judgments in predicting violent, any, and sexual offending? A meta-analysis of direct comparison studies.* Behavioral Sciences & the Law, 43, 75-113.

PARTE III: ANÁLISIS CRÍTICO

¿Qué riesgos conlleva predecir algorítmicamente
el riesgo de reincidencia, violencia en general
o quebrantamiento? ¿Vulnera derechos fundamentales?
¿Es beneficioso para el sistema penitenciario?
¿Y para la población penitenciaria?
¿Ayuda a proteger a la ciudadanía?

Capítulo 9

Algoritmo, Algoritmo… ¿me podrías decir qué puedes aportar mediante la predicción de riesgo de reincidencia en la misión de facilitar que todas las personas que deban cumplir una pena y medida penal en Euskadi lo hagan en condiciones de dignidad, respetando sus derechos y ayudándoles a prepararse bien para su rápida reintegración a la sociedad?

Juan Aldaz Arregui

Profesor Agregado de Sociología de la EHU
Instituto Vasco de Criminología (IVAC / KREI)

"1. Ninguna decisión de la Administración penitenciaria que implique la apreciación del comportamiento humano de los reclusos podrá fundamentarse, exclusivamente, en un tratamiento automatizado de datos o informaciones que ofrezcan una definición del perfil o de la personalidad del interno".
(Artículo 6. *Limitación del uso de la informática penitenciaria*, CAPÍTULO III, Protección de los datos de carácter personal de los ficheros penitenciarios, Reglamento penitenciario)[128]

SUMARIO: 1. INTRODUCCIÓN. 2. RIESGOS DE LA PREDICCIÓN DE RIESGOS. 2.1. *Una mirada general al marco de la sociedad del riesgo*. 2.2. *El paradigma del riesgo en el ámbito penitenciario.* 3. LA DIFICULTAD DE CUANTIFICAR Y PREDECIR EL FUTURO. 4. ALGUNOS DE LOS DILEMAS CENTRALES DE LA POTENCIAL APLICACIÓN

128. Real Decreto 190/1996, de 9 de febrero, por el que se aprueba el Reglamento Penitenciario. https://www.boe.es/buscar/pdf/1996/BOE-A-1996-3307-consolidado.pdf

DE ALGORITMOS PREDICTIVOS AL ÁMBITO PENITENCIARIO. 4.1. *Datificación de la realidad. Entre la calidad y la calidez de los datos*. 4.2. *El procesamiento de los datos. Entre la búsqueda de la agilización y el riesgo de la precipitación*. 4.3. *Decisiones informadas. Entre sesgos y aprendizajes.* 5. RETOS ORGANIZATIVOS PREVIOS A LA POTENCIAL IMPLEMENTACIÓN DE UN ALGORITMO PREDICTIVO. 6. CONCLUSIONES. 7. REFERENCIAS.

1. INTRODUCCIÓN

Al igual que en el cuento de Blanca Nieves, puede que, como la bruja del cuento, estemos pidiendo al espejito (en nuestro caso al algoritmo predictivo[129]) que nos muestre una imagen que no es verdad. Quizás, queremos que nos mienta mostrando una imagen de fiabilidad, objetividad y de una certeza que, camuflada con desarrollos tecnológicos incomprensibles para los no legos en la materia, aparenta una rigurosidad que se mezcla con la verdad. No obstante, como sucede con los espejos, cualquier solución tecnológica, en el mejor de los casos, hará las veces de un espejo, más o menos distorsionado, donde observar el reflejo de nuestra imagen, es decir, de la apariencia. ¿Pero qué sucede con lo que no es fácilmente observable? ¿Con todo lo que queda oculto al espejo-algoritmo?

En este trabajo nos vamos a centrar en aspectos relativos a la construcción y gestión de datos empleados para informar la toma de decisiones vinculadas a la gestión de la misión de la administración penitenciaria, que no es otra que *facilitar que todas las personas que deban cumplir una pena y medida penal en Euskadi lo hagan en condiciones de dignidad, respetando sus derechos y ayudándoles a prepararse bien para su rápida reintegración a la sociedad*[130].

Los algoritmos predictivos, en general, y los relativos a la predicción de riesgo de reincidencia, en particular, se apoyan en mayor o menor medida en la inteligencia artificial para poder procesar grandes cantidades de datos —Big Data— a enormes velocidades. Ahora bien, estos desarrollos tecnológicos siguen sin poder anticipar, y consideramos que así va a seguir siendo en el futuro, si una persona está lo suficientemente reeducada y preparada para su reinserción en la sociedad, de acuerdo con el mandato de la Constitución Española sobre las penas privativas de libertad y las medidas de seguridad (artículo 25 de la CE). Es más, por muchos y fantásticos

129. Cuando nos refiramos al algoritmo predictivo en este trabajo, lo haremos teniendo en cuenta que, en mayor o menor medida, está asistido por inteligencia artificial.

130. https://www.euskadi.eus/administracion-penitenciaria/web01-a2justic/es/

datos que tengamos para tratar de predecir el comportamiento futuro de las personas, y por muchos medios que se pongan para neutralizar posibles "factores precipitadores" del comportamiento delictivo, siempre va a ser necesario un acto de confianza. No solo eso, basta que una persona reincida en su comportamiento delictivo durante un permiso para que se cuestione el conjunto de la política penitenciaria que, por lo demás, y sin negar el dolor que puede causar a las personas afectadas[131], puede estar alcanzando sus fines de resocialización de forma significativa.

Frente a planteamientos tecnófilos y tecnófobos enfrentados sobre la aplicación de algoritmos predictivos, en este trabajo planteamos una serie de retos y oportunidades con una base en el enfoque de evaluación de programas y políticas públicas orientada al aprendizaje (Preskill y Torres, 1999a, y 1999b). Este enfoque se centra en la elaboración de sistemas de seguimiento integrados en la labor del personal técnico que facilite la toma de decisiones informada y el aprendizaje individual como organizacional. Es decir, lejos de abordar la utilidad de la aplicación de algoritmos predictivos como artefactos sustitutorios de la labor técnica, se basa en la generación de un sistema de información que posibilite una suerte de conciencia reflexiva individual y colectiva que funcione a modo de espejo donde poder valorar el desempeño de la intervención.

¿Por qué plantearse, entonces, la posibilidad de implementar un algoritmo predictivo de riesgo de reincidencia en la administración penitenciaria de Euskadi? El origen de la puesta en marcha de esta propuesta reside en una preocupación absolutamente central para la administración penitenciaria de Euskadi: la evaluación del éxito de su intervención en el desistimiento de la población reclusa respecto del comportamiento delictivo, principalmente violento, que origino su privación de libertad.

Ahora bien, este propósito, lejos de ser un reto netamente técnico, tiene un claro componente ético que nos confronta al sentido mismo de la labor de la administración penitenciaria y del propio *ius puniendi*. Este hecho, necesariamente, lejos de plantear la posibilidad de una puesta en marcha de un algoritmo en términos meramente tecnológicos, nos obliga a analizar la utilidad y el interés real de un planteamiento de este tipo en el contexto

131. Un ejemplo reciente lo podemos encontrar en el caso de Remedios Sánchez, quién asesino a una anciana estando de permiso (https://www.lavozdegalicia.es/noticia/coruna/2025/10/14/remedios-sanchez-mataviejas-condenada-varios-asesinatos-serie-detenida-nueva-muerte-coruna-durante-permiso-penitenciario/00031760453460932363625.htm).

específico de Euskadi para, posteriormente, valorar la viabilidad del mismo. Así pues, la pregunta central a la que ha de dar respuesta este proyecto no es cómo desarrollar un protocolo asistido por un algoritmo, sino, previo a ese posible desarrollo, responder a la pregunta de si, realmente, es necesario este planteamiento, y de serlo, en qué condiciones ha de ser implementado.

Por eso mismo, en este trabajo nos planteamos, más allá de las posibles funcionalidades de estas tecnologías aplicadas a la gestión penitenciara en tanto que posible solución, cuál es el problema de fondo: la gestión de la información para la toma de decisiones informadas sobre aspectos tan sensibles como las progresiones-regresiones con base a las valoraciones de riesgo de reincidencia de las personas privadas de libertad.

No cabe duda de que el planteamiento mismo de esta posibilidad se da a la vez que se da la irrupción de la inteligencia artificial en el conjunto de la sociedad y en la reactivación de un enfoque de gestión de lo público basado en lo que Morozov (2015) define como *solucionismo tecnológico*. Lejos de caer en la *tecnofobia* ni la *tecnofilia*, la cuestión es no perder de vista cuáles son los retos éticos a los que nos enfrentamos para, posteriormente, valorar las respuestas técnicas más adecuadas.

En este trabajo, vamos a comenzar por contextualizar el marco cultural en el que el constructo del riesgo destaca como el marco explicativo principal, dando pie a epistemologías predictivas basadas, cada vez en mayor medida, en la inteligencia artificial. A continuación, vamos a profundizar en la dificultad de cuantificar y predecir el futuro. Posteriormente, vamos a centrarnos en algunos de los dilemas centrales de la potencial aplicación de algoritmos predictivos al ámbito penitenciario para, finalmente, esbozar una serie de retos organizativos previos a la potencial implementación de un algoritmo predictivo.

2. RIESGOS DE LA PREDICCIÓN DE RIESGOS

> *"La definición del riesgo depende de una construcción social, posee el carácter a la vez histórico y polémico. La percepción del riesgo no es un calco en la conciencia del peligro corrido que debe preservarse en una conducta, sino que siempre se mezcla con la subjetividad de las representaciones sociales y culturales. Las formas de peligro señaladas remiten en profundidad a modos de vida, a valores colectivos, reflejan una moral en acto, una visión del mundo. La disposición para asumir riesgos o para rechazarlos, incluso ignorarlos completamente, no de distribuye de la misma forma de una clase o de una cultura a otra. La medida "objetiva" del riesgo constituye, por lo tanto, una ficción política y social, no es la misma según*

los criterios de evaluación moral, se nutre con un debate permanente entre los diferentes sujetos concernidos, ya que implica consecuencias económicas y sociales, en ocasiones, considerables".

(Le Breton, 2021: *82)*

2.1. UNA MIRADA GENERAL AL MARCO DE LA SOCIEDAD DEL RIESGO

El marco cultural, o la sensibilidad social, en la que se plantea la posibilidad de implementar un algoritmo predictivo de riesgo de reincidencia, se enmarca dentro de lo que Ulrich Beck (1998), de forma muy acertada, denominó como sociedad del riesgo, es decir, una sociedad donde la producción de riesgos va de la mano de los supuestos avances sociales, propiciando una nueva forma de percepción y gestión, basadas en un tipo de reflexividad[132] sobre los mismos, propia de nuestra época. Así, incluso las propias tecnologías, incluidas aquellas diseñadas para la predicción y el control de riesgos, en una carambola rocambolesca y paradójica, propia de nuestra sociedad, se convierten en fuente de riesgo. En este sentido, compartimos la afirmación de Lupton (1999: 13) al considerar que "los significados y las estrategias de riesgo son intentos de controlar la incertidumbre, pero a menudo tienen el efecto paradójico de aumentar la ansiedad sobre el riesgo a través de la intensidad de su enfoque y preocupación".

En todo caso, el marco explicativo del riesgo como gran constructo a través del cual tratamos de gestionar nuestra existencia en sociedad no viaja solo, va de la mano de conceptos tales como el *peligro*, la *seguridad*, las *decisiones* (incluidas las predictivas), la *responsabilidad*, o el *arrepentimiento*, entre otras (López y Luján, 2000). Nociones que conectan con el deseo de predecir y controlar situaciones más o menos lejanas en el tiempo que pueden resultar dañinas.

Tal y como señalan López y Luján (2000: 23) refiriéndose a Bechmann, "El riesgo es una modalidad atenuada de la inseguridad: donde hay inseguridad, ésta trata de ser controlada mediante el cálculo del riesgo. El riesgo representa así una medida de la seguridad que puede alcanzarse razonablemente", por lo que, como indica Luhmann (1991, en López y Luján, 2000: 23) se abre la posibilidad a hacer algo frente al mismo: "recogerse más información, invertirse más recursos en tiempo o dinero, promulgarse nuevas leyes".

132. Lo que Beck, Giddens y Lash (2001), entre otros autores, han denominado como modernidad reflexiva.

Se podría decir, pues, que el conocimiento sobre el riesgo es un saber que no se sabe todo, por lo que se convierte en una suerte de ignorancia atenuada, tal y como plantea Wynne (1992) y, por lo tanto, más aceptable y asumible emocionalmente.

A este respecto, siguiendo con López y Luján, cabe conectar el concepto de riesgo con los conceptos de decisión y responsabilidad, porque, tal y como indica Luhmann (en López y Luján, 2000: 23) los riesgos, a diferencia de los peligros, más azarosos, son vistos "como una consecuencia de la propia actuación u omisión". Así, si los daños "son vistos como daños fortuitos o contratiempos, serán entendidos socialmente como peligros; pero si se perciben como fruto de decisiones (u omisiones), entonces serán entendidos como riesgos que conllevan imputabilidad (respecto al responsable de la acción, sea un tercero o uno mismo" (López y Luján, 2000: 24).

Aquí emerge una figura clave en el marco de la sociedad del riesgo, la persona experta, aquella a la que se le otorga la responsabilidad de estimar los riesgos y tomar las mejores decisiones. Ahora bien, en este contexto cultural en el que nos encontramos, tal y como apunta Beck (en Beck, Giddens, y Lash: 2001: 25) ya "no se trata únicamente de tomar decisiones; es vital restablecer las reglas y las bases de las decisiones, relaciones de validez y críticas de las consecuencias imprevisibles y carentes de responsable (conceptualizadas desde la pretensión de control)".

Y esto es así porque, tal y como apunta Beck (en Beck, Giddens y Lash: 2001), la sociedad del riesgo es una sociedad tendencialmente autocrítica donde las expertas y los expertos se contradicen entre sí, cuestionándose mutuamente y, lejos de articular y armonizar explicaciones diversas sobre realidades complejas, se enfrentan deslegitimándose unos a otros y amplificando la incertidumbre y la sensación de riesgo de la ciudadanía que, perdiendo la esperanza en el conocimiento científico, recurre a otras fuentes de conocimiento y certidumbre que les acojan.

2.2. EL PARADIGMA DEL RIESGO EN EL ÁMBITO PENITENCIARIO

Llevado al ámbito penitenciario que nos ocupa, el paradigma del riesgo ha supuesto, en mayor o menor intensidad, un giro en las políticas penales de un enfoque basado en un modelo rehabilitador individualizado (modelos del "bienestar"), a otro más administrativo centrado en la gestión de poblaciones y basado en técnicas actuariales de cuantificación y evaluación del "riesgo" de ciertos perfiles (Feeley y Simon, 1994). La implementación de

algoritmos predictivos para calcular el riesgo de reincidencia de las personas privadas de libertad se enmarcaría en este cambio de paradigma que, en mayor o menor medida, también parece tener presencia en nuestro contexto (Aleman, 2023).

Y esto es así debido a que, como ha sucedido con otra serie de problemas sociales, la criminalidad —y la consiguiente intervención punitiva— "son vistas como un riesgo calculable, evitable y gobernable", de tal manera que grupos de población penada diferentes "son subdivididos, categorizados y clasificados de forma rutinaria de acuerdo a sus niveles de riesgo", y ciertos grupos de delincuentes, en concreto, son percibidos como excepcionalmente peligrosos y, consiguientemente, destinatarios de un control legislativo especial (tal es el caso de los agresores sexuales, los delincuentes reincidentes, etcétera (Hannah-Moffat, 2013: 129).

Tal y como advirtiera David Garland en su ya clásico *The Culture of Control* (2001: 175), las personas privadas de libertad son, cada vez más, "vistas como riesgos que han de ser gestionados, más que rehabilitados. En vez de enfatizar los métodos rehabilitadores que respondan a las necesidades de los delincuentes, el sistema enfatiza los controles efectivos que minimicen los costes y maximicen la seguridad". Así, todo indica que las preocupaciones sociales sobre el supuesto riesgo que suponen algunas personas privadas de libertad para el conjunto de la sociedad son las que dan forma a la política penitenciaria que, tratando de buscar un equilibrio entre dar respuesta a sus funciones resocializadoras y no alarmar a la sociedad frente a algunos casos especialmente sensibles, parece intensificar el peso de la balanza hacia un modelo penal donde se dan intervenciones más incapacitadoras, expresivas y centradas en la conciencia de riesgo, para "tranquilizar" al conjunto de la sociedad.

Tomando como referencia, principalmente, los avances dados en medicina en relación con las prácticas *basadas en la evidencia*, este marco se ha ido extendiendo a otros ámbitos hasta alcanzar también el ámbito penitenciario, donde se han centrado, especialmente, en la búsqueda de evidencias de lo que "funciona" y no "funciona" a la hora de reducir la reincidencia (Naughton, 2005; Warren, 2008). En concreto, tal y como recoge Martínez (2020: 240), se trataría de "[…] políticas, procedimientos y programas que la evidencia científica ha demostrado que reducen la reincidencia para grupos específicos de delincuentes, como personas sometidas a *probation*, en *parole*, o delincuentes adictos a drogas" (National Conference of State Legislatures, 2011: 28). (…)."

Es por ello que, al estar basadas en "evidencia científica" informada por herramientas actuariales, estas aportaciones son caracterizadas como moralmente neutrales y apolíticas (Hannah-Moffat, 2013: 131), cuando en realidad, ahondando de forma crítica en la idea de política basada en la evidencia, en la línea de Pawson (2006), uno de los autores de referencia en la materia, y sirviéndonos de las palabras de Martínez (2020: 241): "(…). (…), la apelación a la "evidencia científica" corre el riesgo de convertirse en un recurso para legitimar prácticas que no han demostrado su eficacia en la reducción de la delincuencia y para presentar como meras cuestiones técnicas lo que no dejan de ser opciones políticas". Es decir, más que de política basada en la evidencia, puede que estemos hablando de evidencia basada en política (Martínez, 2020).

En el origen de esta búsqueda de evidencia científica para lograr una mayor certeza en las valoraciones de riesgo de reincidencia se encuentra el cuestionamiento del juicio clínico y de los instrumentos empleados —principalmente la Tabla de Variables de Riesgo (TVR) y la Tabla de Concurrencia de Circunstancias Peculiares (TCCP)— por el personal técnico de tratamiento para dicha tarea, en tanto que "poco rigurosos" y potencialmente tendentes a los sesgos. De hecho, en el origen de estos instrumentos, tanto de la TVR como de la TCCP, se encontraba la finalidad de implementar una mayor racionalidad al juicio clínico, pero la realidad es que, hasta la fecha, no parecen haberse mostrado de gran utilidad a la hora de dar respuesta al fin para el que fueron diseñados.

Quizás es por ello que nos encontramos ante un resurgir de un "solucionismo tecnológico" (Morozov, 2015) de una labor, la tratamental, difícilmente sustituible por ninguna tecnología. Es más, cabe cuestionar la premisa sobre la que descansan estos instrumentos, a decir, que la reincidencia delictiva es predecible y puede ser cuantificada, clasificada y ordenada numéricamente. Y todo ello sirviéndose de estudios basados en datos sobre grandes muestras de población que permiten calcular la influencia de factores estáticos como dinámicos en la probabilidad de reincidencia de personas concretas por su perfil, que por suerte o desgracia, coincide con los establecidos estadísticamente como de mayor o menor riesgo de reincidencia, de tal manera que, como nos advierte Hannah-Moffat (2013: 139), "en la práctica esta posible correlación se convierte en causalidad, a la vez que el riesgo potencial se convierte en certeza administrativa"[133].

Quizás, por ello, parece que muchos de los técnicos que trabajan en primera línea de las labores tratamentales reniegan de estos instrumentos

133. La traducción es nuestra.

técnicos que parecen no servir para captar la potencial evolución de las personas privadas de libertad, constriñendo el conocimiento atesorado por los técnicos a variables cuantificables sin darles ninguna certeza ni seguridad en sus decisiones técnicas. Y ante esta situación, se presenta la potencial implementación de algoritmos predictivos basados, en mayor o menor medida, en inteligencia artificial, como el RisCanvi. Pero para ello, antes que nada, solemos obviar una fase previa crítica para la potencial elaboración de predicciones: la construcción de los datos que han de alimentar cualquier modelo algorítmico, porque sin datos, concretamente sin datos masivos, no hay posibilidad de ningún algoritmo predictivo. Y esta fase esencial, y compleja donde las haya, suele ser obviada.

3. LA DIFICULTAD DE CUANTIFICAR Y PREDECIR EL FUTURO

"No todo lo que se puede contar cuenta,
ni todo lo que cuenta se puede contar"
(Albert Einstein)

El potencial principal de los algoritmos predictivos es poder procesar de forma neutral y a una gran velocidad una ingente cantidad de datos, de tal manera que se pueda predecir la probabilidad de reincidencia de una persona, pero es precisamente ahí, donde reside su supuesto potencial de mejora, donde reside su gran falacia, los algoritmos predictivos no pueden predecir el futuro, en el mejor de los casos pueden generar prospecciones basadas en los resultados de la evolución del comportamiento de personas con perfiles similares al de la persona a la que se está valorando (con todas las dificultades apuntadas sobre la datificación de la realidad de las personas penas y teniendo en cuenta que es una cuestión esta que contradice el sistema de individualización científica previsto en el ordenamiento penitenciario español), y en el peor de los casos, puede que no solo no sean capaces de predecir el futuro, sino más bien, puede que lo produzcan, dado que, tal y como apunta Innerarity (2025: 164):

> "El conocimiento anticipatorio no trata de representar la realidad sino de producir un futuro deseado, ya sea en materia de seguridad, en relación con los riesgos (…). Las predicciones analíticas tienden a crear el futuro que están intentando predecir. La política llevada a cabo con estos instrumentos convierte un efecto aún no producido en una causa, lo que podría llamarse una <<causa futura>> (Massumi, 2007). Ese futuro anticipado es el incentivo para adoptar acciones preventivas inmediatas que cambien el curso de la realidad y esa es la razón de que el acontecimiento previsto no suceda nunca. La predicción acierta cuando impide que suceda lo que se había predicho. En este nuevo orden del mundo, la verdad es de naturaleza retroactiva".

Así, tal y como indicara un técnico de tratamiento entrevistado en el marco de este proyecto, la mejor manera para no correr ningún riesgo es no dar ningún permiso ni propiciar ninguna progresión. Ahora bien, poniendo en riesgo el sentido esencial mismo recogido en la CE de las penas privativas de libertad: la reeducación y reinserción social.

Supuestamente, la incorporación de algoritmos predictivos basados en IA puede mejorar el rigor de las decisiones técnicas en el ámbito penitenciario, avanzando hacia una mayor objetividad y sistematicidad de las decisiones, frente a los posibles sesgos y la discrecionalidad de las decisiones técnicas adoptadas actualmente por el personal técnico, tanto individual como colegiadamente en el marco de las Juntas de Tratamiento. De hecho, se plantea como una suerte de novedad cuando, en realidad, tal y como argumenta Agar (2003, en Innerarity, 2025: 18) en relación a los ordenadores, pero que resulta igual de válido para la IA, no son los gobiernos y las administraciones públicas las que adoptan una nueva lógica de gestión, sino que es la IA la que se adapta a la que ha sido la función principal de los gobiernos, que no es otra que realizar "operaciones que miden, planifican y establecen procesos para la toma de decisiones conforme a cierto orden". La particularidad de los algoritmos predictivos es que estos son capaces de procesar una cantidad ingente de datos a velocidades descomunalmente rápidas.

El actual fenómeno de la gobernanza algorítmica, tal y como define Innerarity (2025) a este desarrollo de algoritmos basados en IA para la gestión de diferentes ámbitos sociales, "forma parte de una tendencia más amplia hacia la matematización y la mecanización de la gobernanza que viene de antiguo" (Innerarity, 2025: 19), donde "los procedimientos de cálculo y algoritmización prometen neutralizar los prejuicios subjetivos mediante procedimientos exactos de decisión (Innerarity, 2025: 20). El objetivo, en último término, es "*hacer el mundo calculable*", según Horkheimer y Adorno (1998: 63), y así hacer "comparable lo heterogéneo reduciéndolo a grandezas abstractas" de tal forma que, "todo lo que no se agota en números, en definitiva en el uno, se convierte para la Ilustración en apariencia", y, consiguientemente, una realidad de segundo orden, menos científica y, consiguientemente, menos valida.

La tentación podría ser, servirse de artefactos que permitan la cuantificación y la monitorización de las personas privadas de libertad. De hecho, cada vez más, las personas cuantificamos una gran cantidad de ámbitos de nuestro día a día, incluidos los más íntimos, sirviéndonos de las diferentes tecnologías a nuestra disposición, tales como los teléfonos móviles inteligentes, pulsómetros o demás artefactos sin que nadie nos obligue (Lupton,

2016). Esto, obviamente, tiene grandes implicaciones éticas y cabe abordarse con gran prudencia y escepticismo, evitando caer en la tentación de una suerte de modelo panóptico postmoderno. Este, sin duda, es un tema que requeriría un trabajo específico que supera los objetivos de nuestro trabajo. En todo caso, queda anotado como riesgo a tener en cuenta, como el deseo del Big Data, puede convertirse en una suerte de Big Brother.

Ahora bien, una pregunta que sí nos ocupa aquí es preguntarnos si todo es cuantificable y de qué manera influye la sensibilidad de quién realiza los registros y, consiguientemente, la *datificación* de la realidad, en la construcción de los datos, entre otras preguntas.

En el seminario desarrollado el 13 de octubre en Donostia en el marco del presente proyecto de investigación y bajo el título *Círculo de diálogo: retos criminológicos ante los instrumentos de valoración y gestión de riesgos en el ámbito penitenciario. una exploración de su adaptación al ámbito vasco*, Iulen Ibañez[134], valiéndose de un ejemplo escatológico, explicaba de forma gráfica que, si metes mierda en el algoritmo, sale mierda. Es decir, si los datos con los que se entrena el algoritmo son malos, las predicciones que pueda realizar serán malas.

Downes y Rock (2011: 82) ya nos advertían al respecto cuando afirmaban que

> La búsqueda de evidencias criminológicas es trabajosa: implica una inversión de dinero, tiempo y energía. Excepto por unos pocos casos de evidencias indirectas, los datos [data] en realidad son una denominación errónea, porque rara vez son *datos*. Lo más apropiado sería llamarlos *capta*, pues se trata de hechos que son *captados* con dificultad. Por lo tanto, las deficiencias y los sesgos de los indicios particulares no pueden ser subsanados de manera automática mediante la acumulación de un amplio espectro de evidencias.

No obstante, parece extenderse una creencia, una suerte de neo-positivismo, donde la imagen de *gestionabilidad* que proporcionan las realidades datificadas hace que creamos que la realidad es más gestionable y predecible de lo que lo es en realidad. Esto puede que se deba al hecho de que los datos se perciban como neutrales y, consiguientemente, se los confiere una autoridad que no se le otorga a otra serie de percepciones o codificaciones no numéricas.

Ahora bien, tal y como nos recuerda Innerarity (2025: 119-120)

> (…). Los datos no son algo <<hecho>> sino construcciones epistémicas (generalmente, a partir de algo dado). Los datos remiten siempre a su descripción,

134. Vid. Datacy: https://datacy.es/nosotros/

> a una atribución activa de significado. Toda cuantificación implica una previa cualificación, en otras palabras, una interpretación acerca de qué cuenta como dato. Los eventos y sucesos no se convierten en datos hasta que no han sido interpretados como tales.

Siguiendo con Innerarity (2025: 120)

> A pesar de su nombre, los datos no son algo dado sino hecho. Las tecnologías de análisis de datos determinan lo que puede ser recogido como dato y cómo se convierte en algo analizable. <<Los datos tienen que ser *imaginados como* datos para existir y funcionar como tales>> (Gitelman / Jackson 2013, 3). Ocurre lo mismo con las bases de datos y sus infraestructuras: no son medios técnicos neutrales que recogen y comparten datos, sino que su diseño y configuración establecen ciertos tipos de datos y posibilitan ciertos tipos de análisis; el modo en el que están estructurados tiene profundas consecuencias en qué clase de preguntas y análisis pueden ser llevados a cabo. Las bases de datos son creadas por actores con determinados objetivos y que trabajan en determinadas comunidades de aprendizaje, y expresan ciertas relaciones de poder que, a su vez, reproducen. Los datos no son independientes del sistema de pensamiento y de los instrumentos en los que se basa su producción (Bowker / Star 1999).

Así pues, el acto de datificar para hacer cuantificable una realidad compleja, lo reconozcamos o no, no deja de ser un proceso de simplificación —e incluso de simplismo en algún caso—, donde tratamos de convertir en datos realidades francamente complejas, generando, consiguientemente, pérdidas de información en dicho proceso.

Si a este hecho le sumamos la particular realidad de las personas privadas de libertad que, como analizara Goffman (2001) en un contexto como el carcelario se presta a propiciar una "obediencia simulada" a las normas dados los consiguientes beneficios que se derivan de dicho comportamiento, pensar que la evolución de la persona privada de libertad en el contexto penitenciario y su traslación al exterior son lo mismo, sería hacer nuestra la más que probable ficción de la evaluación que se da en este singular contexto, y que puede encajar con lo que el subdirector de tratamiento de Gipuzkoa afirmaba en una entrevista en el Diario Vasco (14/12/2025) como una realidad bastante habitual y que es que "En la cárcel hay estupendos presos y malos ciudadanos"[135]. Sin duda, este es otro de los grandes nudos gordianos relativos a la calidad de los datos que habrían de alimentar un posible algoritmo predictivo.

135. Vid. en https://www.diariovasco.com/gipuzkoa/carcel-estupendos-presos-malos-ciudadanos-20251214030028-nt.html

En último término, y a modo de conclusión para este epígrafe, es bueno volver a la cita de Einstein que le da inicio recordando, sobre todo, que no todo es cuantificable, y que, en algún caso, lo cuantificable no es de gran interés. De hecho, cabe traer a colación el chiste del borracho al que se encuentra un policía agachado bajo la luz de una farola y que, al preguntarle este último por qué está así, el primero le contesta que porque ha perdido las llaves y, como ahí es el único lugar donde hay luz, es más fácil buscarlas. Es decir, que algo sea fácilmente traducible a dato, no quiere decir que sea la información más relevante y, en lo que a nosotros nos ocupa en este trabajo, que podamos valernos de la luz que puedan arrojar los algoritmos predictivos no nos confunda del hecho de que las respuestas puede que estén en la oscuridad.

4. ALGUNOS DE LOS DILEMAS CENTRALES DE LA POTENCIAL APLICACIÓN DE ALGORITMOS PREDICTIVOS AL ÁMBITO PENITENCIARIO

El debate en torno a la utilización de IA a diferentes contextos sociales parece moverse entre las filias y las fobias. El debate sosegado y racional parece brillar por su ausencia. Parece darse una especie de "complejo de Frankenstein", es decir, un miedo a los desarrollos tecnológicos, en general, y a la IA, en concreto (Coeckelbergh, 2021), por un lado, o todo lo contrario, una fe ciega en las soluciones tecnológicas (lo que Morozov califica como solucionamos tecnológico, 2015).

En términos generales, puede que los riesgos y las oportunidades que nos presenta la potencial aplicación de un algoritmo predictivo al contexto penitenciario sean los mismos. Es decir, depende de cómo los abordemos podrán suponer la solución o la amplificación de los problemas a los que se viene enfrentando la Administración Penitenciara a la hora de tratar de cumplir con su misión de la mejor de las maneras posibles. Es más, los retos a los que nos hemos referido, tal y como hemos podido observar, no son característicos de la potencial aplicación de algoritmos predictivos, sino de la labor propia de la Administración Penitenciara, y de las Juntas de Tratamiento, en concreto, de evaluar la evolución de las personas privadas de libertad y de tratar de tomar, de la forma más sustentada y rigurosa posible, las mejores decisiones en relación a la progresión-regresión de las mismas.

El planteamiento de una potencial implementación de algoritmos predictivos, por lo tanto, no parece que genere nuevos problemas, sino que modifica la intensidad de los problemas que ya existían. Ni siquiera la mistificación de los algoritmos predictivos como portadoras de soluciones mágicas al

problema de la evaluación del desarrollo de las personas privadas de libertad es nueva. Ya se ha dado en relación a los instrumentos de valoración antes citados, como las TVR y las TCCP, y que las compañeras y compañeros que me acompañan en este trabajo colectivo describen en mayor detalle en la presente obra. El imaginario mitológico de los algoritmos predictivos, en la línea de lo indicado por Cabrera (2021), no es nuevo, se intensifica de forma específica, pero sin aportar nada nuevo a la ilusión de dar respuestas tecnológicas a dilemas éticos. Y esto es así porque, adelantándonos a la propuesta de trabajo que realizaremos en el siguiente epígrafe, los dilemas que nos confronta la potencial puesta en marcha de un algoritmo predictivo vienen de lejos, y nos confrontan a las fases previas a la toma de decisiones informadas (ver imagen a continuación), a decir la *datificación de la realidad*, *el procesamiento de los datos*, y *las decisiones basadas en la información extraída del procesamiento de datos*, que darán unos resultados que, a su vez, serán datificados, reiniciando el ciclo de procesamiento de información (ver **figura 1**, a continuación). Vamos a analizar los dilemas de cada una de estas fases.

Figura 1. Fases para la construcción de sistemas de información orientado a la toma de decisiones informadas.



Fuente: elaboración propia.

4.1. DATIFICACIÓN DE LA REALIDAD. ENTRE LA CALIDAD Y LA CALIDEZ DE LOS DATOS

Sobre esta primera fase de la datificación ya hemos hablado de forma significativa en el epígrafe anterior incidiendo en el hecho de que ningún dato es neutro. Todo dato es construido y el resultado de un interés previo hacia unos aspectos de la realidad y no otros, por ser considerados más interesantes o, simplemente, por ser cuantificables. Dicho de otro modo más llano, se podría decir que vemos-datificamos lo que miramos, pero ¿cómo miramos?

Como decíamos, estas cuestiones epistemológicas han sido suficientemente abordadas en el epígrafe anterior y son, como hemos advertido en la

introducción a este epígrafe, problemas epistemológicos que no solo afectan a la potencial aplicación de algoritmos predictivos al ámbito penitenciario, sino que son previas y propias de la labor de evaluar la evolución de las personas privadas de libertad.

Un elemento transversal para la mejora de los datos, o, mejor dicho, de los captados, reside en poder transitar desde una idea de calidad meramente técnica, "fría", a otra que también reconozca la calidad ética, más cálida, de los mismos. O lo que podríamos denominar como "poner rostro" a las diferentes realidades presentes en prisión. Esto solo es posible mediante el establecimiento de un vínculo de confianza donde se haga posible conocer a la persona concreta y sus circunstancias, más allá de la categoría o las variables que utilicemos para clasificarla.

Creo que se podrá ver claramente a qué me refiero con el siguiente ejemplo que utilizo habitualmente para incidir en la necesidad de ir más allá del dato. Tenemos tres personas que se han presentado a un examen de, pongamos, sociología de la desviación. Las tres han obtenido un "7". Yo, como profesor responsable de la asignatura, podría tener una disposición distante sobre la asignatura y sentirme significativamente satisfecho por la valoración media del grupo. Ahora bien, si fuera un profesor más incisivo, si tuviera una relación más cercana con el alumnado, si estuviera más pendiente y encima del alumnado, hubiera detectado que una de ellas ha obtenido el "7" habiendo pasado una semana nefasta, con grandes problemas familiares e, incluso, estando enferma, gracias a un gran esfuerzo personal. La segunda apenas se ha esforzado. Ha venido ha clase habitualmente y, dada su facilidad para con la materia, ha sacado un "7" sin ningún tipo de esfuerzo. Por último, la tercera persona, que apenas ha venido a clase, que cuando ha venido apenas ha participado, le he descubierto copiando en el examen. De no haber detectado esta mala *praxis*, su calificación hubiera sido de un "7". La pregunta y la respuesta a la pregunta son obvias: ¿merecen la misma respuesta las tres notas? Evidentemente no.

Esto es lo que sucede cuando tratamos de reducir realidades complejas a respuestas simples, perdemos una gran cantidad de información de vital importancia, sobre todo, cuando nos referimos a realidades sobre las que hay que tomar decisiones para intervenir sobre estas personas.

Las consecuencias de esta suerte de ceguera epistemológica son muchas, pero, sobre todo, cabe destacar el riesgo del sesgo, tanto si nos servimos de algoritmos predictivos como si no. La base de cualquier sesgo es tomar decisiones con información parcial y limitada, y eso es un riesgo tanto en

un escenario mediado por algoritmos predictivos como por personas. De hecho, se nos olvida que en los diseños de los algoritmos predictivos también hay personas que deciden qué cuenta y qué no para llevar a cabo un cálculo supuestamente neutro (O'Neil, 2017).

Esto nos lleva a uno de los problemas más destacados de cualquier instrumento orientado a la valoración de riesgo, el sesgo. Ahora bien, aunque sea un riesgo tanto para las personas como para los casos en los que se utilicen algoritmos predictivos, la especificidad de estos últimos reside, como venimos insistiendo, en la creencia de que son neutros. Volviendo a O'Neil (2017: 38-39), nos advierte de lo siguiente:

> Muchas personas dirían que los sistemas estadísticos como el LSI-R[136] son efectivos para estimar el riesgo de reincidencia —o, al menos, que son más precisos que la conjetura aleatoria de un juez—. El problema es que, si dejamos de lado, aunque sea durante un segundo, la cuestión esencial de la justicia, nos sumiremos sin darnos cuenta en el bucle de retroalimentación pernicioso de un ADM[137]. Es probable que una persona calificada <<de alto riesgo>> esté en paro y provenga de un barrio en el que muchos de sus amigos y familiares hayan tenido roces con la ley. Gracias en parte a su alta puntuación en la evaluación como consecuencia de sus orígenes, esa persona será condenada a una pena más larga, lo que la encerrará en la cárcel durante más años, donde estará rodeada de otros delincuentes, lo que incrementa la probabilidad de que vuelva a prisión. Cuando finalmente sea puesta en libertad, volverá al mismo barrio pobre, aunque ahora con antecedentes penales, con lo que le resultará mucho más difícil conseguir un empleo. Si comete otro delito, el modelo de reincidencia se atribuirá otro éxito, pero lo cierto es que el propio modelo está alimentando un círculo tóxico y ayuda a mantenerlo. Esa es una cualidad distintiva de un ADM.

Así pues, el riesgo del sesgo es propio de cualquier mirada, más o menos tecnificada. La cuestión es estar abiertos a romper con los juicios previos, los prejuicios, y estar abiertos a modificarlos en cada caso particular, pero para ello, como apuntábamos al comienzo de este epígrafe, es imprescindible, entre otras cuestiones, el tiempo para poder establecer relaciones significativas, para poder observar la evolución de las personas y, en definitiva, para poder conocer a las personas más allá de las categorías y variables a las que, supuestamente, "pertenecen". Esta idea encaja con lo expuesto por

136. (Level of Service Inventory-Revised) es una herramienta estandarizada y actuarial utilizada en el sistema de justicia penal para evaluar el riesgo de reincidencia de adultos.

137. O'Neil se refiere a los algoritmos como Armas de Destrucción Matemática. De ahí el acrónimo ADM, por sus iniciales.

Innerarity (2025: 72-73), quien considera que "el problema del aprendizaje de los sistemas inteligentes es que son incapaces de ver cada caso como un caso particular; están concebidos para construir estereotipos a través de una gran cantidad de datos. (…). El sistema no nos ve sino a través de las propiedades por las que no somos un caso único".

Sin negar la utilidad que tienen estas etiquetas y los prejuicios a la hora de hacer cognoscible la realidad en nuestro día a día, como una suerte de "economía emocional" para gestionar la compleja realidad social, cuando esto afecta a derechos fundamentales, es lo primero que tenemos que tratar de controlar.

4.2. EL PROCESAMIENTO DE LOS DATOS. ENTRE LA BÚSQUEDA DE LA AGILIZACIÓN Y EL RIESGO DE LA PRECIPITACIÓN

Una de las grandes virtudes de los algoritmos predictivos es que son capaces de procesar una cantidad enorme de datos a una gran velocidad. Cantidades y velocidades que son inalcanzables por los humanos. Por lo menos para cierto tipo de información calculable. Ahora bien, los humanos somos capaces de procesar otra serie de informaciones que nos son fácilmente calculables. De hecho, hasta la fecha, tal y como afirma Innerarity (2025: 43-44)

> La inteligencia artificial únicamente simula algunos aspectos concretos de la inteligencia humana, pero no lleva a cabo todas las tareas de la inteligencia humana, que no incluyen sólo cálculo y rapidez, sino también comprensión y reflexión. Los anunciadores del futuro *sorpasso* no están hablando de la inteligencia integral sino de las capacidades analíticas de la inteligencia instrumental, desde una epistemología estrechamente empírica, calculadora y que ignora el contexto histórico de la vida humana.

Es mas, siguiendo con Innerarity (2025: 45)

> (…). Por muy actualizados que estén los datos con los que se entrenan y aunque pudiéramos personalizarlos en ámbitos específicos, se alimentarán siempre de algo que ya existe y no tendrán esa capacidad de generar novedad, que entendemos como una propiedad de la verdadera inteligencia.

Así, una de las grandes virtudes de la IA es, a su vez, su gran defecto: la incapacidad de equivocarse. Es decir, la incapacidad de dar con un error acertado, o lo que comúnmente entendemos como serendipia. No es que aboguemos por promover los errores, pero es cierto que en la realidad a la que nos referimos, el reto de la resocialización, es probable que muchos

"errores" hayan propiciado avances en el proceso de resocialización de las personas privadas de libertad, y también lo contrario, que muchos "aciertos" hayan impedido avances en la resocialización. Este es un aspecto que desarrollamos en el próximo subepígrafe.

No obstante, el aspecto crítico principal en relación al procesamiento de datos tiene que ver con el fenómeno de la *caja negra*, o dicho de otro modo más llano, la opacidad que existe sobre el procesamiento de datos, tanto en el caso de las personas como en el caso de los algoritmos predictivos. La cuestión es como poder trazar y hacer explicable los cálculos desarrollados a través de los modelos empleados en cada algoritmo predictivo porque, de lo contrario, según algunos expertos, tal y como se deriva de la revisión sistemática realizada por Scaria, Subramanian, George, y Sengupta (2024), podemos caer en injusticias derivadas de sesgos en el diseño de los algoritmos predictivos.

En el caso de los procesos analíticos y decisorios también nos encontramos con la idea de la caja negra que, en muchos casos, se explica —tanto para ponerlo en valor, como para restárselo— en términos de intuición. En ambos casos, tanto en el caso de los algoritmos predictivos como de las personas, lo que explica esa inexplicabilidad, esas cajas negras, es que hay aspectos que escapan a nuestra capacidad cognitiva más racional, en el caso de los algoritmos predictivos por la ingente cantidad de datos que se procesa, y en el caso de las personas, por la incapacidad de expresar cuantitativamente percepciones y sentimientos que nos hacen dudar de los datos fríos (Gigerenzer —2008— desarrolla de forma maravillosa esta idea en su ensayo *Decisiones instintivas: la inteligencia del inconsciente*).

Esta falta de transparencia y explicabilidad en el diseño algorítmico tiene su particular significancia e implicaciones en relación al artículo 6, relativo al derecho a un proceso equitativo, recogido en el *Convenio Europeo de Derechos Humanos*[138], tal y como nos advierte Ugwudike (2022: 95) "un problema con la opacidad algorítmica y los derechos de propiedad que protegen a los creadores de publicar su código para su escrutinio es que dichas herramientas no pueden ser refutadas ni cuestionadas por los acusados".

Ahora bien, tal y como indica Innerarity (2025: 316), quizá, el problema de la falta de transparencia y explicabilidad no es tanto en términos tecnológicos y de procedimientos tecnológicos, sino "que expliquen sus decisiones de forma que estas tengan sentido para sus interlocutores". De hecho,

138. Vid. https://www.echr.coe.int/documents/d/echr/convention_spa

siguiendo con Innerarity (2025: 316) es esto mismo lo que sucede cuando pedimos a las personas que expliquen sus decisiones:

> Cuando a uno se le da una explicación obtiene información sobre la decisión sin ser informado sobre los procesos neurofisiológicos o psíquicos experimentados por quien ofrece la explicación. Explicar nuestras decisiones no requiere revelar nuestro proceso de pensamiento, ni mucho menos las conexiones de nuestras neuronas. Sólo necesitamos escuchar razones que permitan que la comunicación se desarrolle de forma controlada y no arbitraria.

He ahí, a nuestro entender, la clave, en el control y la no arbitrariedad sobre las informaciones empleadas y la manera de relacionarlas. De algún modo, y de formas diferentes, tenemos que asumir un mínimo de cajas negras sobre una arquitectura de datos que sea explicable y que sea comprensiblemente transparente.

4.3. DECISIONES INFORMADAS. ENTRE SESGOS Y APRENDIZAJES

En último término, vendría la toma de decisiones basada en los datos procesados y que, *a priori*, habría de posibilitar una toma de decisiones informada. Ahora bien, además de la calidad (en términos técnicos y éticos) de los datos empleados, de la transparencia y explicabilidad de los procesos, el reto final es tomar las mejores decisiones posibles en relación al proceso terapéutico de las personas privadas de libertad.

El artículo 6 del Reglamento Penitenciario es claro al respecto, en su primer punto:

> 1. Ninguna decisión de la Administración penitenciaria que implique la apreciación del comportamiento humano de los reclusos podrá fundamentarse, exclusivamente, en un tratamiento automatizado de datos o informaciones que ofrezcan una definición del perfil o de la personalidad del interno.

Ahora bien, ¿hasta qué punto la existencia de un algoritmo predictivo, puede conllevar que, de facto, las decisiones las tome una inteligencia artificial y no una persona y, por consiguiente, que nos enfrentemos a un riesgo real de lo que Bauman (2009) denominó como *adiaforización*, es decir, una realidad donde los sistemas y los procesos, en este caso penales, se alegan de cualquier consideración moral.

Valiéndonos, una vez más, de las aportaciones de Innerarity (2025: 23-24)

> El problema fundamental de la inteligencia artificial es la creciente externalización de decisiones humanas en ella. La automatización generalizada

> plantea el problema de qué lugar corresponde a la decisión humana, si se trata simplemente de un suplemento, de una modificación o un remplazo. (…). El hecho de que automaticemos ciertas decisiones individuales o colectivas implica grandes beneficios en términos de efectividad. Sin embargo, este potencial puede constituir una amenaza si implica una rendición absoluta de nuestra soberanía.

Es decir, la pregunta a la que nos podemos terminar enfrentando en caso de darse una delegación excesiva en los algoritmos predictivos, es "¿quién decide cuándo, aparentemente, nadie decide?" (Innerarity, 2025: 24). No obstante, la respuesta a esta pregunta tan solo puede ser una: nosotras, las personas, y en nuestra sociedad, en el ámbito de la Administración Penitenciaria, las personas que componen los equipos técnicos y las Juntas de Tratamiento. Y esto ha de ser así porque las y los profesionales de las Juntas de Tratamiento han de representar la incomodidad moral que nos ha de generar la aplicación-modulación de un castigo, una suerte de mala conciencia, como planteaba Christie (1988), que nos permita castigar con una mirada humanizadora incluso ante los hechos más inhumanos que hayan podido realizar algunas personas privadas de libertad. Y esto no parece que, ni actualmente ni en un futuro cercano, nos lo pueda dar un algoritmo predictivo.

Pero cuidado, la intervención de humanos en la toma de decisiones no garantiza, automáticamente, la humanidad de las decisiones. Volviendo a Bauman, y concretamente a su trabajo clásico *Modernidad y Holocausto* (1998), no podemos olvidar las lecciones de nuestra historia porque supondría olvidar que algunas de las decisiones supuestamente más racionales y modernas de nuestra historia, a la postre, se han mostrado como las más inhumanas y, consiguientemente, como las más irracionales, si tenemos en cuenta, como nos advertiría Atienza (en Bergalli, 2003: 265), la racionalidad ética de las decisiones.

5. RETOS ORGANIZATIVOS PREVIOS A LA POTENCIAL IMPLEMENTACIÓN DE UN ALGORITMO PREDICTIVO

Toda vez que hemos planteado de forma teórica los aspectos epistemológicos que consideramos críticos en relación a la potencial implementación de un algoritmo predictivo, basándonos en el análisis de las diferentes investigaciones sobre la materia, la documentación propia de la Administración Penitenciara de Euskadi, el seminario desarrollado el 13 de octubre en Donostia, dos entrevistas en profundidad llevadas a cabo con profesionales

de la Administración Penitenciara de Euskadi con responsabilidades en el proceso de tratamiento, entre otras acciones llevadas a cabo en el marco del proyecto *Evaluación de necesidades para la potencial puesta en marcha de un algoritmo para la evaluación del riesgo/probabilidad de reincidencia/desistimiento*, queremos plantear una serie de retos más organizativos a los que se ha de enfrentar la Administración Penitenciaria de Euskadi para poder avanzar hacia lo que, entendemos, se esconde detrás del planteamiento de la potencial puesta en marcha de un algoritmo predictivo: conseguir que las decisiones que han de tomar los diferentes profesionales penitenciarios estén cada vez mejor informadas.

Si bien no vamos a desarrollar en profundidad el proceso de construcción de organizaciones con una cultura evaluadora y de aprendizaje en el sentido de lo planteado por Preskill y Torres (1999a, y 1999b), creemos que previo al desarrollo potencial de un algoritmo predictivo en la Administración Penitenciaria de Euskadi, se encuentra la necesidad de construir un buen sistema de datos que permita optimizar los aprendizajes de la institución en cuestión. El día a día, el volumen de trabajo, y otra serie de realidades elementales condicionan desde la base misma la realidad de toda organización, incluida la Administración Penitenciaria, por eso, antes del abordaje de un potencial algoritmo, sobre cuyos resultados somos modestamente optimistas, hay que trabajar la misma arquitectura de datos que, progresivamente, se integra en la labor cotidiana de las y los profesionales de la Administración Penitenciaria, y que permita extraer aprendizajes que abran la puerta a una intervención cada vez más basada en la toma de decisiones informadas.

Con todo, entendemos que hay unos retos elementales que son imprescindibles para la Administración Penitenciaria con el fin de avanzar en este sentido.

Reto 1. Clarificar las necesidades informativas de los diferentes stakeholders

Un paso previo a cualquier tipo de sistema de información es clarificar cuáles son las necesidades informativas de las partes implicadas (los *stakeholders*) en una política o programa concreto. Esto es determinante de cara a optimizar los siguientes retos. Todas las acciones que se lleven a cabo de cara a poner en marcha un sistema de seguimiento y evaluación que permita mejorar la toma de decisiones de las diferentes partes, ha de partir de la clarificación de cuáles son sus necesidades. Esto se clarifica, por supuesto, teniendo en cuenta las responsabilidades de los diferentes profesionales, el tipo de decisiones que han de adoptar y, por lo tanto, cualquier tipo de

acción orientada a obtener datos y procesarlos, ha de tener en cuenta estos aspectos si aspira a tener algún tipo de utilidad (Patton, 2008).

Reto 2. Integrar la recogida de datos en la labor cotidiana

Uno de los retos principales para poder transitar a un sistema de datos que posibilite diferentes usos, incluidos un potencial algoritmo predictivo, es que la recogida de los datos necesarios para alimentarlo esté integrada en la labor cotidiana de las y los profesionales. Es decir, que no sea una labor añadida. Esto, además, tiene que posibilitarse garantizando que los datos que se recogen, a la postre, permiten obtener información de utilidad para las y los profesiones, evitando cualquier tarea que resulte innecesario. De lo contrario, corremos el riesgo de que las y los profesionales no se apropien del proceso. A este respecto, el principio de *menos es más* habría de adoptarse como directriz principal.

Reto 3. Importancia del dato único para el conjunto de la Administración Penitenciaria de Euskadi

Otro de los retos tiene que ver con la necesidad de trabajar por el dato único, por lo menos dentro de la Administración Penitenciaria. Es decir, que toda la información sobre una persona este centralizada en una única base de datos que permita poder hacer diferentes análisis, evitando la perdida de energía y de aprendizajes que parecen darse actualmente porque cada profesional técnico registra y almacena los datos de forma diferente. Esto implicaría establecer protocolos y procedimientos comunes a las tres prisiones en materia de recogida y gestión de datos para coordinarse de la mejor manera posible.

El ideal sería, toda vez que se garantice el dato único a nivel penitenciario, poder ampliar la lógica del dato único al conjunto de la administración pública de Euskadi.

Reto 4. Contar con un Sistema Informático Penitenciario (SIP) propio.

El cuarto y último reto sería poder contar con un SIP propio. Actualmente se trabaja con el sistema heredado de la Administración Penitenciaria Central, pero, de cara a tener un control total sobre los datos y sus posibles análisis, sería imprescindible poder contar con un SIP propio.

Si bien es cierto que hay otra serie de retos de interés para poder avanzar en una mejor gestión de la información en el ámbito penitenciario de

Euskadi, entendemos que, en estos momentos, poder abordar estos cuatro retos a los que nos referimos sería un gran avance.

6. CONCLUSIONES

No cabe duda de que el planteamiento de la potencial puesta en marcha de un algoritmo predictivo por parte de la Administración Penitenciaria de Euskadi reside en una preocupación absolutamente central para la misma: la evaluación del éxito de su intervención en el desistimiento de la población reclusa respecto del comportamiento delictivo, principalmente violento, que origino su privación de libertad.

Ahora bien, tal y como hemos tratado de mostrar a lo largo de este trabajo, este deseo, lejos de suponer un reto netamente técnico, tiene un claro componente ético que nos confronta al sentido mismo de la labor de la administración penitenciaria y del propio *ius puniendi*. De hecho, en el fondo, creemos que dar respuesta a este reto ético conlleva elegir entre hacer una política penitenciara desde la confianza o desde la desconfianza hacia las personas.

Los datos, "fríos" o "calientes", nunca van a poder responder plenamente a esta preocupación sobre el comportamiento futuro de las personas, en el mejor de los casos, nos podrán dar algunas pistas sobre el comportamiento de personas con características similares en condiciones similares. En todo caso, los algoritmos predictivos habrían de incorporar una visión cada vez más sistémica, estructural y compleja sobre la realidad de las personas penadas, con una mirada más corresponsabilizadora sobre el castigo y no tan centrada en aspectos meramente individuales y psicológicos (Aldaz, 2014).

Al planteamiento de la implementación de un potencial algoritmo predictivo le precede la necesidad de construir un sistema de datos optimizado, que permita compartir aprendizajes y mejorar la toma de decisiones informada (basada en registros compartidos y validados). Por eso, aunque, probablemente, la puesta en marcha de un algoritmo puede que no aporte demasiado a la realidad de Euskadi, puede servir de palanca para abordar este reto previo de mejora del sistema de datos.

Por eso, quizás, y en el mejor de los casos, dando respuesta a la pregunta que da título a este trabajo, la implementación de un algoritmo predictivo podrá hacer las veces de un espejo para la Administración Penitenciaria de Euskadi. Un espejo que nos pone ante una realidad que hace visible lo visi-

ble, pero que sigue sin arrojar luz sobre los dilemas morales más profundos a los que nos confronta modular la función del castigo en nuestra sociedad. Y no es poco. Poder mirarnos al espejo y gustarnos lo que vemos sería, probablemente, la gran aportación que puede realizar un algoritmo predictivo.

7. REFERENCIAS

Aldaz, J. (2014). "La institución del castigo en la sociedad del control social incontroladouna propuesta desde la evaluación para una institución de castigo corresponsabilizadora". *International e-journal of criminal sciences*, N.º 8.

Alemán, L. (2023). "El uso de RISCANVI en la toma de decisiones penitenciarias". *Estudios Penales Y Criminológicos*, *44*(Ext.), 1-43. https://doi.org/10.15304/epc.44.8884.

Bauman, Z. (1998). *Modernidad y Holocausto*. Sequitur.

Bauman, Z. (2009). Ética Posmoderna. Siglo XXI

Beck, U. (1998). *La sociedad del riesgo. Hacia una nueva modernidad*. Paidós Básica.

Beck, U., Giddens, A., y Lash, S. (2001). *Modernización reflexiva. Política, tradición y estética en el orden social moderno*. Alianza Universidad.

Bergalli, R. (2003). *Sistema Penal y Problemas Sociales*. Tirant lo Blanch.

Cabrera, D.H. (2021). "El algoritmo como imaginario social", *Zer*, 26(50), 125-145. https://doi.org/10.1387/zer.22206.

Coeckelbergh, M. (2021). Ética de la inteligencia artificial. Cátedra.

Downes, D. y Rock, P. (2011). *Sociología de la desviación. Una guía sobre las teorías del delito*. Editorial Gedisa.

Goffman, E. (2001). *Internados. Ensayos sobre la situación social de los enfermos mentales*. Amorrortu Editores.

Gigerenzer, G. (2008). *Decisiones instintivas: la inteligencia del inconsciente*. Ariel.

Horkheimer, M. y Adorno, T.W. (1998). *Dialéctica de la Ilustración. Fragmentos filosóficos*. Editorial Trotta.

Le Breton, D. (2021). *Sociología del Riesgo*. Prometeo libros.

López, J. A., y Luján, J.L. (2000). *Ciencia y Política del Riesgo*. Alianza Editorial.

Lupton, D. (1999). *Risk*. Routledge.

Lupton, D. (2016). *The Quantified Self*. Polity.

Morozov (2015). *La locura del solucionismo tecnológico*. Katz Editorial.

Naughton, M. (2005). "'Evidence-based policy" and the government of the criminal justice system — only if the evidence fits!" *Critical Social Policy*, *25*(1), 47-69. https://doi.org/10.1177/0261018305048967 (Original work published 2005).

O´neil, C. (2017). *Armas de destrucción matemática. Cómo el Big Data aumenta la desigualdad y amenaza la democracia*. Capitán Swing.

Patton, M.Q. (2008). *Utilization-Focused Evaluation*. Sage Publications.

Preskill, H. y Torres, R.T. (1999a). *Evaluative Inquiry for Learning in Organizations*. Sage.

Preskill, H. y Torres, R.T. (1999b). "Building Capacity for Organizational Learning through Evaluative Inquiry". *Evaluation*, Vol. 5(1): 42-60. https://doi.org/10.1177/13563890990050010.

Scaria, A. G., Subramanian, V., George, N. K., y Sengupta, N. (2024). "Algorithms and Recidivism: A Multi-disciplinary Systematic Review". *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*(1), 1292-1305. https://doi.org/10.1609/aies.v7i1.31724

Warren, Roger K. (2009). "Evidence-Based Sentencing: The Application of Principles of Evidence-Based Practice to State Sentencing Practice and Policy." *University of San Francisco Law Review*: Vol. 43: Iss. 3, Article 4.

Wynne, B. (1992). "Uncertainty and environmental learning: Reconceiving science and policy in the preventive paradigm". *Global Environmental Change, Volume 2, Issue 2*. https://doi.org/10.1016/0959-3780(92)90017-2.

Capítulo 10

Aproximación crítica al uso de instrumentos algorítmicos en la ejecución penitenciaria desde la perspectiva del Derecho penal

Lorena Alemán Aróstegui

Profesora Ayudante Doctora de Derecho penal.
Universidad Pública de Navarra

SUMARIO: 1. INTRODUCCIÓN: EL USO DE ALGORITMOS Y LA DETERMINACIÓN DEL CASTIGO PENAL. 2. EL DERECHO PENAL Y LA INCIDENCIA DE LA PELIGROSIDAD EN LA DETERMINACIÓN DEL CASTIGO. 2.1. *Derecho penal de la culpabilidad por el hecho*. 2.2. *El ámbito de valoración de la peligrosidad: su extensión a las penas y a los sujetos imputables.* 3. PELIGROSIDAD, RIESGO Y USO DE ALGORITMOS EN LA EJECUCIÓN PENITENCIARIA. 3.1. *Peligrosidad versus valoración del riesgo*. 3.2. *Aspectos positivos y negativos del uso de instrumentos algorítmicos frente al juicio clínico en la ejecución penitenciaria.* 4. CONCLUSIONES. 5. REFERENCIAS.

1. INTRODUCCIÓN: EL USO DE ALGORITMOS Y LA DETERMINACIÓN DEL CASTIGO PENAL

En las últimas décadas, estamos asistiendo a avances científico-tecnológicos muy importantes que están cambiando nuestras vidas rápidamente y en algunos casos de manera sustancial. Entre estos avances, se encuentran por supuesto los sistemas de inteligencia artificial y los basados en algoritmos. Y no hay duda de que para el Derecho —y, en concreto, para el Derecho penal— la incorporación de este tipo de sistemas genera tanto oportunidades como innumerables retos, lo que está dando lugar a un intenso debate presente en casi cada espacio de discusión académica.

Teniendo en cuenta este escenario, en este abordaje crítico sobre el uso de algoritmos en la ejecución penitenciaria, parto de dos premisas. En primer lugar, el reconocimiento y la asunción de que este tipo de tecnologías están ya implantadas en muchos sectores y que la tendencia no es hacia su reducción, sino hacia la ampliación de su uso. Y, en segundo lugar, la reivindicación de que es posible y necesario establecer límites a su utilización, diferenciando ámbitos de aplicación y alcances, y advirtiendo tanto de su incidencia en derechos fundamentales como de las nuevas oportunidades que generan. Por tanto, se trata de innovaciones tecnológicas cuya relevancia y futuro resultaría ingenuo y absurdo negar. Pero ello no implica plegarse a ellas, sino que es posible —e ineludible— establecer límites jurídicos a su utilización y, también, imaginar y desarrollar formas de ponerlas al servicio de políticas públicas sociales y transformadoras.

Desde la perspectiva del Derecho penal, el análisis de estos nuevos desarrollos científico-tecnológicos conecta con un tema de discusión de largo recorrido y que sigue en ciernes: qué relación debe establecerse entre el Derecho penal y lo empírico. En este sentido, comparto la opinión de un número considerable de autores que han planteado que el Derecho penal no puede permanecer de espaldas a la realidad (Hassemer & Muñoz Conde, 1989; Díez Ripollés, 1997; Sotomayor Acosta, 2008; Donini, 2011; Miró Linares & Fuentes Osorio, 2021). El Derecho penal debe atender a los cambios sociales, tecnológicos y culturales. Además, habida cuenta del sustrato empírico de muchas de las categorías jurídico-penales (imputabilidad, dolo, etc.), de las finalidades preventivas del sistema penal y de los límites que no deben sobrepasarse para evitar vulneraciones de derechos, el Derecho penal, no solo no debe ignorar el conocimiento empírico que pueden ofrecerle otras disciplinas, como la sociología, la criminología o la psicología, sino que lo necesita.

No obstante, ello no debe conducir a que se borren los contornos que permiten distinguir los dos planos en los que se mueven unas y otras disciplinas (Ferrajoli, 1986; Sotomayor Acosta, 2008; Donini, 2019): el plano empírico (descriptivo), en el caso de las ciencias que desarrollan estas tecnologías, las ciencias del comportamiento y las sociológicas y criminológicas; y el plano jurídico (valorativo y prescriptivo), en el caso del Derecho penal. Con ello, el objetivo es doble: evitar incurrir en falacias naturalistas —que deriven lo que *debe ser* de lo que *es*— y vacíen de contenido las normas y principios jurídicos; y, viceversa, evitar la falacia normativista que se produce cuando, ignorando la realidad, se deriva lo que *es* de lo que *debe ser* (Ferrajoli, 1986).

Por tanto, el Derecho penal debe abrirse a la innovación tecnológica y al conocimiento empírico sin renunciar a sus fundamentos, principios y reglas jurídicas. Especialmente, teniendo en cuenta que es la disciplina jurídica que se ocupa de regular la mayor restricción de derechos que puede realizar el Estado, puesto que determina en qué casos, cómo y con qué alcance pueden y deben imponerse los castigos penales.

Como se viene reivindicando desde la Ilustración (Beccaria, 1764) y se recoge en los manuales de Derecho penal (Roxin, 1997; Mir Puig, 2003 y 2015; Luzón Peña, 2025), para que el ejercicio del *ius puniendi* pueda ser considerado legítimo, deben fijarse unos límites que impidan que la lógica preventiva derive en excesos punitivos. En la actualidad, el marco constitucional ofrece un fundamento positivo a esos límites y vías para el reclamo de su necesidad, su respeto y su vigencia. Además, desde mi punto de vista, estos límites —infranqueables aun en detrimento de los fines preventivos— deben operar en toda manifestación del ejercicio del *ius puniendi* para evitar que este devenga ilegítimo, entendiéndose aquí incluida la ejecución penitenciaria (Alemán Aróstegui, 2025, pp. 241 y ss.). Si bien en la fase de ejecución la presunción de inocencia ha quedado obviamente vencida, el resto de las garantías jurídico-penales deben seguir operando respecto de toda decisión penitenciaria que afecte al grado de privación de libertad y que, con ello, modifique su aflictividad y module el castigo penal; esto es, lo determine, aunque no en su clase o duración, pero sí en su intensidad (Alemán Aróstegui, 2025, pp. 340 y ss.).

2. EL DERECHO PENAL Y LA INCIDENCIA DE LA PELIGROSIDAD EN LA DETERMINACIÓN DEL CASTIGO

A partir de la aprobación de la Constitución en 1978, nuestro Derecho penal se configuró como un Derecho penal de la culpabilidad por el hecho, vertebrado por los principios constitucionales y respetuoso con los derechos fundamentales. Ello implicó derogar y dejar atrás la imposición de medidas de seguridad predelictuales que atendían a la peligrosidad social, y establecer el hecho delictivo como límite absoluto de toda intervención penal. Se restringió la posibilidad de aplicar medidas de seguridad basadas en la peligrosidad criminal a los supuestos de merma de la imputabilidad, siempre que se hubiera cometido un hecho delictivo (art. 6 CP). Sin embargo, el Código penal de 1995 abrió desde el principio otros espacios a la valoración de la peligrosidad (suspensión de la ejecución y libertad condicional) y, posteriormente, las tendencias político-criminales de las últimas décadas han ampliado ese margen que en el ámbito punitivo se deja a la peligrosidad.

2.1. DERECHO PENAL DE LA CULPABILIDAD POR EL HECHO

En España, históricamente, se ha otorgado un lugar preeminente al estudio de la personalidad criminal. En este sentido, destaca el desarrollo de la frenología[139] por parte de la intelectualidad española de principios del siglo XIX (Quintero Olivares, 2017, pp. 37 y ss.), y, desde la segunda mitad del siglo XIX, de la mano de la filosofía krausista y la influencia del positivismo italiano, la construcción de la teoría correccionalista (Quintero Olivares, 1999, pp. 27 y ss.). El correccionalismo español —cuyos principales referentes fueron Salillas, Dorado Montero y Silvela— no llegó a los extremos del positivismo en la abolición de garantías (Quintero Olivares, 1999, p. 55), pero puso en el centro la peligrosidad y, con ello, enfrentó los postulados clásicos del Derecho penal presentes en los Códigos penales de 1822, 1848 y 1870.

El Código penal de 1928, elaborado durante la dictadura de Primo de Rivera, acogió la tesis correccionalista, y la Ley de Vagos y Maleantes de 1933 abrió paso a "la idea positivista de la peligrosidad sin delito" (Roldán Barbero, 1988, p. 184), que se extendió más allá de los enfermos mentales —tradicionalmente etiquetados como peligrosos—, abarcando a todos los sujetos "marginales" por distintas causas, de cuyo estudio se ocupó la sociología criminal que vinculaba el delito con la "mala vida" y que empezó a desarrollarse en España en aquella época (Quintero Olivares, 2017, pp. 45 y ss.). La Guerra Civil introdujo en este escenario la cultura de la guerra, la militarización, la brutalidad y la represión política (Gargallo Vaamonde, 2011). Y, posteriormente, durante la dictadura franquista se combinaron todos estos elementos con la represión y el control de la marginalidad a través de la persecución de los "estados peligrosos" que contemplaba la Ley de 1933 —una de las pocas leyes de la II República que no fueron derogadas durante la dictadura—. En 1970, dicha norma se sustituyó por la Ley sobre Peligrosidad y Rehabilitación Social (LPRS), "manteniendo sustancialmente sin modificación los principios en que la Ley de mil novecientos treinta y tres se inspiró"[140].

El Derecho penal de la culpabilidad por el hecho se configuró a partir de la instauración del Estado democrático y de Derecho, en el que se rechazó

139. La frenología sostenía que distintas zonas cerebrales se relacionaban con rasgos específicos de la personalidad y que el desarrollo o tamaño de determinadas regiones del cerebro estaba vinculado con la intensidad de las capacidades mentales correspondientes. Marià Cubí i Soler fue uno de sus representantes más importantes (Quintero Olivares 2017, p. 39).

140. Así se recoge en el Preámbulo de la Ley 16/1970, de 4 de agosto, sobre peligrosidad y rehabilitación social.

la posibilidad de privar de libertad a una persona por considerar que su forma de ser, su modo de vida o las circunstancias en las que le ha tocado vivir denotan su peligrosidad (Roxin 1997, pp. 176 y ss.). Así quedaron superadas las medidas de seguridad predelictuales basadas en la declaración de estados peligrosos, como tratarse de "vagos habituales", "ebrios habituales" o toxicómanos, realizar actos de homosexualidad, mendigar, ser menores de 21 años abandonados por la familia o tratar asiduamente con "delincuentes o maleantes" (art. 2º LPRS). Así, la culpabilidad por el hecho constituye la garantía para evitar la defensa de "la conveniencia de la actuación del sistema represivo en nombre de la peligrosidad *sin delito*", que pueda conducir de nuevo a la justificación de los mayores abusos y a formas de Derecho penal de autor (Quintero Olivares, 1999, pp. 187-188).

En el marco constitucional, la culpabilidad por el hecho es uno de los elementos esenciales del Derecho penal. De tal forma que al sujeto solo se le puede castigar en función de aquello que ha hecho y se le puede reprochar jurídicamente (el injusto culpable), y no por lo que se considera que es o el modo de vida que lleva y el supuesto peligro que ello representa (Pérez Manzano, 1990; Cobo del Rosal & Vives Antón, 1999[141]; Lascuraín Sánchez, 2021), puesto que ello atentaría contra el libre desarrollo de la personalidad (art. 10.1 CE), y contra la igualdad y la prohibición de toda discriminación (art. 14 CE). Por tanto, el Derecho penal de la culpabilidad por el hecho entraña una concepción del sujeto que delinque como un sujeto responsable por sus actos delictivos —independientemente de cómo piense, viva, etc.— y al que reconoce la capacidad y el derecho de autodeterminarse.

En consecuencia, se configuró un sistema penal en el que la apreciación de la peligrosidad quedaba, en principio, restringida a los supuestos de aplicación de medidas de seguridad a quienes cometieran un delito encontrándose en situación de semi o inimputabilidad. Por tanto, se hizo del delito un presupuesto ineludible, y la valoración de la peligrosidad solo tenía que ver con las medidas de seguridad —que no podían imponerse a sujetos imputables— y en ningún caso con las penas.

141. "En el ámbito del Derecho penal no se juzga la persona, sino solamente sus actos. Esta limitación al acto y a la atribuibilidad del acto desempeña una primordial función de garantía. No incumbe al Derecho penal establecer si las inclinaciones del individuo, sus facultades y disposiciones, son dignas de censura o de elogio. Para el Derecho cualquier modo de ser personal es igualmente respetable. Toda concepción de la culpabilidad que la configure como un juicio sobre la persona traspasa los límites propios del Derecho penal para invadir los de la Moral" (Cobo del Rosal & Vives Antón 1999, pp. 536-537).

2.2. EL ÁMBITO DE VALORACIÓN DE LA PELIGROSIDAD: SU EXTENSIÓN A LAS PENAS Y A LOS SUJETOS IMPUTABLES

El Código penal español deja poco espacio a la peligrosidad como fundamento autónomo de las consecuencias jurídico-penales, sin embargo, varias decisiones punitivas incorporan su valoración (Martínez Garay, 2019, p. 124; Guzmán Fluja, 2022, p. 46).

Desde la aprobación del Código penal, el ámbito de valoración de la peligrosidad fue más allá de las medidas de seguridad y los supuestos de semi o inimputabilidad, teniéndose en cuenta para tomar decisiones respecto de las penas, como la suspensión de la ejecución de la pena de prisión o la concesión de la libertad condicional. Asimismo, ha tenido siempre una importante incidencia en la toma de todas las decisiones penitenciarias, a pesar de que —como se verá— la normativa penitenciaria no prevé su valoración en relación con varias de ellas. Finalmente, en los últimos años, en la línea de las tendencias político-criminales securitarias e inocuizadoras, la valoración de la peligrosidad se ha extendido todavía más allá, llegando a fundamentar la posibilidad de imponer una medida de seguridad a un sujeto imputable (libertad vigilada). E, incluso, se ha introducido una pena de duración indeterminada cuyo término dependerá de un pronóstico de peligrosidad (prisión permanente revisable).

Para mostrar qué extensión tiene en la actualidad el ámbito de valoración de la peligrosidad en el sistema penal, he examinado el Código Penal, la Ley Orgánica General Penitenciaria (LOGP, 1979) y el Reglamento Penitenciario (RP, 1996), con el objetivo de encontrar preceptos en los que se incluyan términos que hagan referencia a criterios de valoración de la peligrosidad[142]. Curiosamente, es el Código Penal la norma que más referencias incluye en este sentido. La LOGP y el RP contienen pocas menciones y solo respecto de algunas decisiones penitenciarias.

142. Si bien, en el apartado siguiente, se abordará la discusión respecto de la diferencia entre el concepto de peligrosidad y el de valoración de riesgo, en este análisis, se parte de un concepto amplio de peligrosidad, en el que se entienden incluidos los siguientes términos utilizados en las normas penales y penitenciarias: "peligrosidad" ("valorar la peligrosidad" o la "escasa peligrosidad", la "menor peligrosidad", "falta de peligrosidad" o la "peligrosidad extrema"); "pronóstico" ("pronóstico individualizado y favorable de reinserción social", "pronóstico positivo de reinserción", "pronóstico de falta de peligrosidad", "pronóstico de comportamiento futuro que revele la probabilidad de comisión de nuevos delitos" o "pronóstico de reincidencia o reiteración delictiva"); "probabilidad" ("probabilidad de comisión de nuevos delitos", "probabilidad de comisión de delitos futuros", "juicio de probabilidad sobre el comportamiento futuro").

Se incluyen términos que abren la puerta a la valoración de la peligrosidad en los siguientes ámbitos: en relación con la imposición de medidas de seguridad —incluida la libertad vigilada a imputables por determinados delitos— (arts. 6, 95, 97, 98 y 106 CP); para la suspensión de la ejecución de la pena de prisión y la imposición de prohibiciones o deberes (arts. 80 y 83 CP); en relación con la libertad condicional (arts. 90.5 y 91.3 CP, art. 67 LOGP y arts. 195, 199 y 205 RP); para decidir sobre la posibilidad de revertir los efectos del período de seguridad (art. 36.3 CP) y del cumplimiento efectivo (art. 78.2 CP); en relación con la imposición obligatoria de la libertad vigilada a sujetos imputables (arts. 192.1 y 579 bis.2 CP); para determinar la posibilidad de clasificar en tercer grado o conceder la suspensión de la ejecución en condenas a prisión permanente revisable (arts. 36.1 y 92 CP); para la clasificación en primer grado e imposición del régimen cerrado en prisión (arts. 10.1 LOGP y 74.3, 91.3 y 102.4 RP); y en relación con una serie de aspectos específicos del ámbito penitenciario, como las conducciones y las medidas de seguridad interior (arts. 38 y 65 RP), el tercer grado por motivos humanitarios (art. 104.4 RP) y las salidas a centros para la realización de programas de tratamiento (art. 117 RP).

A la luz de este análisis de previsiones normativas que incorporan la valoración de la peligrosidad, caben varias consideraciones. En primer lugar, ha quedado de manifiesto que las normas penales abren espacios a la valoración de la peligrosidad y el riesgo más allá de las medidas de seguridad aplicables a semi e inimputables, y que la tendencia que se aprecia es hacia la ampliación del ámbito de la peligrosidad (Quintero Olivares, 2017, p. 54).

En segundo lugar, a pesar de que en el ámbito penitenciario se ha utilizado ampliamente la estimación del riesgo de reincidencia para la toma de decisiones (Martínez Garay, 2024), en realidad, son escasas las decisiones penitenciarias en las que se prevén términos que directamente introduzcan este tipo de valoraciones. La normativa penitenciaria restringe la valoración de la peligrosidad casi exclusivamente a la aplicación del régimen cerrado —de hecho, en la LOGP hay una única mención expresa a la peligrosidad, contenida en el art. 10.1 que prevé la posibilidad de imponer dicho régimen de vida— y a la libertad condicional, sin que se incluya mención alguna ni en relación con la concesión de permisos de salida (arts. 46 LOGP y 154 y ss. RP) ni en la actividad ordinaria de clasificación (art. 63 LOGP)[143]. Sin embargo, la escasez de normas legales y el abuso de la utilización de con-

143. Únicamente se prevé en caso de clasificación en primer grado (art. 102.5 RP), en caso de regresión de grado (art. 106.3 RP) y para la modalidad extraordinaria de concesión del tercer grado por motivos humanitarios (art. 104.4 RP).

ceptos jurídicos indeterminados, ambiguos y susceptibles de valoración subjetiva en el ámbito penitenciario han dado entrada a la valoración de la peligrosidad en casi cualquier decisión penitenciaria (Alemán Aróstegui, 2025, pp. 536 y ss.)[144].

En tercer lugar, en el análisis de la regulación se puede apreciar que el ámbito de la valoración de la peligrosidad se ha ampliado progresivamente y ha llegado a la determinación del castigo penal, no solo en cuanto a su "intensidad" como cuando se toma en cuenta en decisiones penitenciarias (Martínez Garay, 2019, p, 180), sino en relación con la extensión del control penal. Por un lado, la libertad vigilada posibilita la imposición de una medida de seguridad a sujetos imputables, de tal forma que, con base en la peligrosidad, habilita una extensión del control punitivo más allá de la medida de la culpabilidad[145]. Y, por otro lado, a pesar de que, según el legislador, la prisión permanente revisable[146] se introdujo como "una respuesta penal ajustada a la gravedad de la culpabilidad"[147], como apunta Martínez Garay, "su régimen jurídico gira alrededor de la peligrosidad" (2019, p. 142). Por tanto, tenemos ya un castigo penal indeterminado y cuya posibilidad de determinación depende de la valoración de la peligrosidad.

3. PELIGROSIDAD, RIESGO Y USO DE ALGORITMOS EN LA EJECUCIÓN PENITENCIARIA

Partiendo del reconocimiento de que, efectivamente —como se ha visto—, las normas penales abren espacios a la valoración de la peligrosidad y, en general, a la realización de pronósticos, debe admitirse que, en el ámbito punitivo, existe un campo abonado para la utilización de instrumentos algo-

144. Si esto es posible es porque la ejecución penitenciaria ha permanecido olvidada e ignorada, y se ha tolerado que sea un ámbito en el que se rebajan las garantías jurídicas contra el *ius puniendi* y se produce una "devaluación" de los derechos fundamentales de las personas presas (Rivera Beiras, 1997).

145. La introducción de la libertad vigilada por la LO 5/2010, de 22 de junio, fue muy criticada por la doctrina penal por romper con la exclusión de la posibilidad de imponer medidas de seguridad a sujetos imputables. Sobre las discusiones acerca de la libertad vigilada, vid. Otero González (2015) y Salat Paisal (2015).

146. La prisión permanente revisable fue recibida por la doctrina penal con un mayoritario rechazo por considerarla una pena contraria a los principios y derechos constitucionales. En consecuencia, más de cien catedráticos/as de Derecho penal firmaron un *Manifiesto contra la prisión permanente revisable* y el Grupo de Estudios de Política Criminal también se posicionó en contra de esta pena (2016, pp. 173 y ss.).

147. Preámbulo Ley Orgánica 1/2015, de 30 de marzo, por la que se modifica la Ley Orgánica 10/1995, de 23 de noviembre, del Código Penal.

rítmicos de valoración de riesgos en el que tradicionalmente el método de valoración que se venía utilizando era el juicio clínico (Andrés Pueyo, 2013; Varona Gómez, 2024; Urruela Mora, 2025).

Si bien, en los apartados anteriores, se ha hecho referencia a la peligrosidad y a los pronósticos de reincidencia o de reinserción sin apreciar diferencias, la realidad es que, en las últimas décadas, se ha producido una evolución muy importante en términos conceptuales y metodológicos por la que, en el contexto anglosajón ya en los 90, se advirtió un cambio de paradigma en la penalidad (Feeley & Simon, 1992[148]), que ha extendido su influencia a los sistemas punitivos continentales (De Giorgi, 2000). Lo que tradicionalmente se denominaba peligrosidad y se analizaba fundamentalmente a través del juicio clínico, tiende a sustituirse por el riesgo y su análisis a través de herramientas de valoración de riesgos con base en técnicas actuariales y algoritmos.

A continuación, se abordarán estos cambios, cuáles son las diferencias y los puntos en común, cómo se interpretan por unos y otras, y qué aspectos positivos y negativos pueden apreciarse de la sustitución de la idea de la peligrosidad y su análisis a través del juicio clínico a la valoración del riesgo y el uso de instrumentos algorítmicos.

3.1. PELIGROSIDAD *VERSUS* VALORACIÓN DEL RIESGO

El concepto de peligrosidad proviene del positivismo criminológico de finales del siglo XIX, que sostenía una explicación etiológica del delito y desde el que se entendía que la peligrosidad era una cualidad de determinados sujetos que los predisponía a cometer delitos, por lo que era esperable que los cometieran en el futuro (Martínez Garay, 2018). Esta concepción de la peligrosidad contenía un "claro matiz clínico", en el sentido de que se equiparaba a un "estado mental patológico" que se entendía que determinaba el comportamiento del sujeto de manera crónica (Andrés Pueyo, 2013, pp. 490-491).

Como se ha expuesto al principio de esta contribución, el positivismo tuvo influencia en España, pero se materializó en una corriente propia, el

148. En Estados Unidos, Feeley y Simon advirtieron el surgimiento de una nueva concepción punitiva, que denominaron *New Penology*, caracterizada por la emergencia de los discursos de la probabilidad y el riesgo (en sustitución del diagnóstico clínico y el enjuiciamiento retributivo), la consecución de objetivos de control efectivo de los procesos internos o sistémicos (en lugar de los tradicionales de rehabilitación y control del crimen) y el despliegue de técnicas que tratan a los sujetos como categorías, en lugar de como casos concretos (Feeley & Simon 1992, p. 452).

correccionalismo. Este tomó las ideas positivistas relativas a la explicación causal del delito sustentada en características antropológico-sociales, pero puso el acento en la posibilidad de *corregirlas* (Roldán Barbero, 1988, pp. 115-116). Por tanto, el concepto de peligrosidad en este marco no renunciaba completamente a la posibilidad del cambio y, por eso, se aprecia que el correccionalismo español no llegó, en sus consecuencias, a extremos de eliminación del delincuente, como sí lo hizo el positivismo italiano (Quintero Olivares, 1999, p. 28). De hecho, en la Ley de Vagos y Maleantes de 1933, no se hacía referencia a sujetos peligrosos sino a "estados peligrosos", lo que denotaba su posible variación (Andrés Pueyo & Redondo Illescas, 2007, p. 163).

En cualquier caso, el concepto de peligrosidad ha sido mirado por la doctrina penal siempre con recelo (Martínez Garay, 2018), y la criminología crítica ha señalado reiteradamente su carácter selectivo, discriminatorio y estigmatizante (Baratta, 2004). Por su parte, además, desde hace algunas décadas, otras corrientes criminológicas y psicológicas se han unido a la crítica al concepto de peligrosidad, considerándolo un concepto "obsoleto", susceptible de numerosas críticas y que, en consecuencia, debe superarse y sustituirse por otro: la valoración del riesgo (Andrés Pueyo 2013, p. 484).

Esta sustitución de la peligrosidad por la valoración del riesgo tiene su origen en el ámbito anglosajón en la década de los 90 (*risk assessment*), como consecuencia de la combinación del auge de los discursos de la probabilidad y el riesgo, el gerencialismo —que consiste en planteamientos de tipo económico y eficientistas para el diseño de las políticas públicas y el actuarialismo —que se basa en el uso de métodos estadísticos para determinar la probabilidad de que un hecho ocurra en el futuro— (Brandariz García, 2016, p. 169). Se entendió como un cambio de paradigma en la penalidad (Feeley & Simon, 1992), que dio lugar a la proposición de modelos de ejecución penal basados en el riesgo (Andrews, Bonta & Hoge, 1990), y que fue desde el principio objeto de fuertes críticas (Simon, 1988; O'Malley, 2004; Harcourt, 2007).

Con la sustitución de la peligrosidad por la valoración del riesgo se producen una serie de cambios muy relevantes, pero también se da continuidad a algunas de las ideas que subyacen al concepto de peligrosidad.

Por un lado, en cuanto a los cambios, con la valoración del riesgo, ya no se trata de averiguar si un sujeto tiene o no una determinada cualidad que lo convierte en peligroso —se niega la existencia de una "personalidad crimi-

nal" (Andrés Pueyo, 2013, p. 492 y ss.[149]), sino que se examina si en el sujeto concurren una serie de factores personales y ambientales, que son cambiantes, y que, de acuerdo con los datos estadísticos disponibles, correlacionan con la mayor probabilidad de cometer delitos, de tal forma que permiten elaborar pronósticos de reincidencia que, además, no son dicotómicos, sino que admiten graduaciones (Martínez Garay, 2018).

Asimismo, lo anterior se traduce en un cambio de tipo metodológico: se sustituye el juicio clínico, considerado carente de base empírica suficiente y de poca utilidad pronóstica (Andrés Pueyo, 2013), por métodos estructurados de estimación del riesgo en los que se utilizan técnicas actuariales que, con base en la premisa de que "el futuro es una repetición del pasado", predicen la probabilidad de que un hecho suceda en el futuro teniendo en cuenta "la combinación ponderada de los factores que determinaron su aparición en el pasado" (Andrés Pueyo & Redondo Illescas, 2007, p. 168). A través de la observación empírica de grandes grupos de sujetos, se buscan "las causas que explican su conducta, ya que entre ellas aparecen los factores de riesgo" (Andrés Pueyo & Redondo Illescas, 2007, p. 164), que luego serán los que se tengan en cuenta para determinar el riesgo de reincidencia de otros sujetos. Esta nueva metodología predictiva trae consigo el uso de nuevas herramientas: los instrumentos de valoración y gestión de riesgos, con los que se pretende obtener predicciones del comportamiento delictivo, mediante la introducción de información relativa a factores de riesgo —determinados a partir de datos estadísticos de delincuencia— en una fórmula algorítmica que, al aplicarse, da como resultado niveles de riesgo.

Los defensores de estos instrumentos de valoración del riesgo sostienen que constituyen una opción mejor para continuar realizando las predicciones necesarias en el ámbito punitivo, partiendo de información similar, pero con una técnica mejorada, que las dota de una mayor objetividad y fiabilidad (Andrés Pueyo & Redondo Illescas, 2007).

Sin embargo, desde perspectivas críticas, se apunta, en primer lugar, que los errores en los pronósticos siguen siendo muy elevados (Tonry, 2019; Martínez Garay, 2020). En segundo lugar, se cuestiona que realmente se consiga una valoración más objetiva (Urruela Mora, 2025), teniendo en cuenta que —como se reconoce— la información con la que se toman las decisiones no

149. "La peligrosidad es un atributo personal, el riesgo es un estado en el que se combinan, necesariamente, factores personales y situacionales que constituyen el 'riesgo' como variable transitiva y no la peligrosidad como entidad constitutiva de la persona" (Andrés Pueyo 2013, p. 501).

es tan diferente (Andrés Pueyo & Redondo Illescas, 2007[150]) y sigue conteniendo elementos subjetivos y ambiguos. En tercer lugar, se critica que se utilicen variables que resulten discriminatorias (Tonry, 2019) que, además, quedan invisibilizadas. Es decir, con el uso de estas herramientas se legitima un tratamiento punitivo desigual de los sujetos en función del riesgo (De Giorgi, 2000) y, de esta forma, queda oculto que dicho riesgo se sustenta en factores personales y sociales que constituyen el reflejo de desigualdades estructurales. En este sentido, se criticó también, desde el principio, el efecto ideológico que podía implicar el paso de la peligrosidad, como cualidad del sujeto individual, a la valoración del riesgo, como situación en la que se encuentra el individuo pero que no integra su subjetividad. Simon apuntó que frente al señalamiento de subjetividades que —a pesar de sus consecuencias negativas— hizo que surgieran movimientos de defensa de los delincuentes, los locos o los homosexuales como identidades estigmatizadas, la práctica actuarial suprime la posibilidad del empoderamiento político, invisibilizando la clasificación de las personas en categorías por la que subsisten los patrones de discriminación estructurales (Simon, 1988, pp. 790-794).

En definitiva, con el paradigma de la valoración del riesgo se solucionan algunas cuestiones problemáticas relativas al concepto de peligrosidad, pero otras se mantienen (Guzmán Fluja, 2022[151]) y, además, surgen nuevos problemas frente a los que desde el Derecho penal debemos mantenernos alerta (Martínez Garay, 2019).

Finalmente, es fundamental tener en cuenta que, en última instancia, tanto la peligrosidad como la valoración del riesgo hacen referencia al futuro y constituyen formas de predicción del comportamiento humano. Tanto en la peligrosidad como en la valoración del riesgo subyace la vocación de predecir cómo se comportará una persona en el futuro (si delinquirá o no) y, en esa medida, son susceptibles de cuestionamientos similares, tanto en términos generales como, especialmente, desde la perspectiva de un sistema punitivo basado en la responsabilidad por el hecho —pasado, no futuro—.

150. "la información utilizada para tomar decisiones sobre peligrosidad y sobre el riesgo no es muy distinta. Lo que sí es diferente es la organización y determinación de qué información es necesaria para evaluar el riesgo de violencia (en cada tipo variará de forma concreta), la ponderación de cada factor de riesgo y las normas de relación entre las valoraciones realizadas que definen los resultados de las mismas" (Andrés Pueyo/Redondo Illescas 2007, p. 166).

151. "Esta transición desde la peligrosidad a la evaluación de riesgos no comporta la eliminación del primer concepto sino su integración en los segundos" (Guzmán Fluja 2022, p. 56).

Toda predicción resulta en alguna medida incierta y puede resultar, además, equivocada. Si bien puede considerarse legítimo e, incluso, recomendable que determinadas políticas públicas y/o consecuencias jurídicas se sustenten en predicciones, el castigo penal, de acuerdo con los principios de legalidad y culpabilidad por el hecho, debe fundamentarse y determinarse en función de hechos ciertos.

Por su parte, quienes, desde la psicología y la criminología, defienden el empleo de estas predicciones en el ámbito punitivo, explican que en todas las tradiciones culturales han existido prácticas de predicción del futuro, y que estas han evolucionado desde las "artes adivinatorias", como los horóscopos, las cartas o los posos del café —consideradas el "modo «artesanal" de predicción"—, pasando por los juicios clínicos de expertos del comportamiento, hasta llegar a las técnicas metodológicas de valoración del riesgo, que se consideran análogas a las que se utilizan para la predicción de fenómenos de la naturaleza, como los pronósticos meteorológicos (Andrés Pueyo & Redondo Illescas, 2007, p. 158[152]).

Si bien en esta contribución no se aspira a discutir esta cuestión en profundidad, consideramos muy cuestionable que este tipo de predicciones puedan equipararse en algún sentido. En primer lugar, porque, como Martínez Garay, entendemos que el comportamiento humano es "esencialmente diferente" de los fenómenos físicos (2020, p. 250), puesto que no atiende a iguales regularidades, entraña una mayor complejidad y tiene siempre un grado de imprevisibilidad irreductible (Tonry, 2013), y, por todo ello, no puede equipararse ni en términos metodológicos ni de resultados. Y, en todo caso —y, especialmente—, porque, en el Derecho penal, rige el principio básico de la culpabilidad por el hecho, por el que debemos negar la posibilidad de que el castigo penal dependa o se gradúe en función de predicciones de eventos futuros.

Sin embargo, como se ha expuesto en epígrafes anteriores, la realidad es que en la actualidad las propias normas penales abren espacios para la elaboración de predicciones. Y lo hacen, además, utilizando no solo la idea de la peligrosidad, sino también los pronósticos de reincidencia o reinserción, o la probabilidad de cometer delitos, dando lugar a una concepción "más operativa" de cara a la valoración del riesgo (Andrés Pueyo, 2013, p. 491). Por tanto, teniendo en cuenta lo anterior, procede llevar a cabo un análisis de los aspectos positivos y negativos que puede implicar el uso de

152. "Predecir la conducta no es metodológicamente distinto de pronosticar si lloverá el fin de semana" (Andrés Pueyo/Redondo Illescas 2007, p. 158).

los instrumentos de valoración de riesgos basados en algoritmos frente al tradicional juicio clínico (Varona Gómez, 2024).

3.2. ASPECTOS POSITIVOS Y NEGATIVOS DEL USO DE INSTRUMENTOS ALGORÍTMICOS FRENTE AL JUICIO CLÍNICO EN LA EJECUCIÓN PENITENCIARIA

El análisis de los aspectos positivos y negativos del uso de instrumentos de valoración de riesgos basados en algoritmos frente al método del juicio clínico se concreta en el ámbito de la ejecución penitenciaria porque, a pesar de que —como se ha visto— son pocas las previsiones penitenciarias que incorporan la peligrosidad o los pronósticos, es principalmente en este ámbito en el que se ha introducido el uso de dichos instrumentos.

No parece casualidad que sea así, principalmente, por dos motivos. En primer lugar, porque, en Estados Unidos —donde la valoración del riesgo tiene un recorrido más largo y consolidado—, el sistema penitenciario fue donde se empezaron a introducir los instrumentos actuariales. Su presencia se puede rastrear desde las primeras décadas del siglo XX, en las que se realizaron numerosos estudios en relación con la posibilidad de determinar factores a partir de datos estadísticos para predecir el éxito o el fracaso de la concesión de la *parole* (Allen, 1942), y fue en la década de los 90 cuando tuvo lugar su eclosión y llegaron a extenderse y asentarse completamente (Feeley & Simon, 1992; O'Malley, 2004). Y, en segundo lugar, porque en la fase de ejecución penitenciaria el margen de discrecionalidad administrativa es muy amplio (Alemán Aróstegui, 2025) y es, en ese espacio, donde ha sido posible introducir los instrumentos de valoración del riesgo, que no se encuentran regulados en ninguna norma legal penal ni tan siquiera en normas administrativas. Además, muchas de las apreciaciones concretas sobre el uso de este tipo de herramientas se referirán al caso de RisCanvi, puesto que se trata del instrumento algorítmico más desarrollado y que se encuentra ya desde hace años completamente implantado en las cárceles catalanas (Alemán Aróstegui, 2023).

Una primera ventaja del uso de instrumentos algorítmicos en el ámbito penitenciario es la posibilidad que ofrecen de recopilar y procesar grandes cantidades de información, sistematizar su recogida, ordenarla y generar un registro que permita tanto agilizar la toma de decisiones y dotarla de mayor seguridad como incrementar la transparencia y controlar mejor el trabajo de los y las profesionales por parte de la Administración (Castro Liñares, 2019). Desde la perspectiva de la Administración penitenciaria, no hay duda

de que supone considerables mejoras en la ordenación y fiscalización del trabajo y, además, permite rendir cuentas de la gestión penitenciaria que se lleva a cabo con mayor facilidad.

Sin embargo, desde la perspectiva de los y las profesionales de la Administración penitenciaria, se advierte que la incorporación de herramientas como RisCanvi genera una enorme carga de trabajo *de despacho*, que, si no se acompaña de un aumento de los recursos personales, reduce el tiempo del que se dispone para la intervención directa y, además, genera déficits en su aplicación porque no se dispone del tiempo necesario para elaborarlo correctamente (Alemán Aróstegui, 2023), y ello —obviamente— disminuye la fiabilidad de las predicciones (Martínez Garay, 2019[153]). Por otra parte, desde la perspectiva de las personas privadas de libertad, se duda de que el uso de estas herramientas suponga una mejora en términos de individualización y de transparencia con respecto al juicio clínico. Como apuntan Andrés Pueyo & Redondo Illescas, las "evaluaciones actuariales no son, en el fondo, evaluaciones individuales sino generalizaciones grupales aplicadas a los individuos y esta es quizás la limitación más importante del procedimiento" (2007, p. 168), de tal forma que su uso para la toma de decisiones penitenciarias puede implicar una merma en la individualización (Cervelló Donderis 2019), e incluso, dar lugar a la "despersonalización" (Urruela Mora 2025, p. 521). Y, en cuanto a la transparencia, si bien el acceso a toda la información en la que se fundamentan y que generan estas herramientas puede contribuir a que la actividad que realiza la Administración penitenciaria en su conjunto sea más accesible y se pueda fiscalizar, en relación con las concretas decisiones penitenciarias que se toman respecto de personas privadas de libertad, los motivos por los que se considera que es probable que delincan quedan ocultos tras el resultado de riesgo (Martínez Garay, 2024). Por tanto, la mejora en términos de transparencia no se extiende a las personas privadas de libertad para la defensa de sus intereses en concreto, y menos aún si —como se ha comprobado respecto de algunas herramientas— ni siquiera se les informa ni de su aplicación ni de sus resultados (Alemán Aróstegui, 2023).

Un segundo aspecto positivo que debe destacarse del uso de herramientas algorítmicas frente al juicio clínico es que permite corregir sesgos discri-

153. "Existe evidencia que indica que cuando los operadores del sistema jurídico aplican estos instrumentos, no siempre respetan las instrucciones y los requisitos marcados en sus protocolos de aplicación. (…) Por ello, decir que una decisión está basada en la evidencia porque incorpora como criterio el nivel de riesgo 'científicamente' estimado de acuerdo con alguna de estas herramientas estructuradas, puede no ser una descripción muy fiel de lo que realmente está sucediendo" (Martínez Garay 2019, p. 170).

minatorios basados en estereotipos. Por ejemplo, en el caso de RisCanvi, en contra del estereotipo xenófobo que conduciría a otorgar un mayor riesgo de reincidencia a una persona extranjera frente a una nacional, esta herramienta otorga mayor riesgo a las nacionales, dado que los datos muestran dicha realidad (Martínez Garay, 2024). Sin embargo, ¿qué ocurre cuando el prejuicio discriminatorio tiene un respaldo estadístico? Ocurre así en el caso, por ejemplo, de las personas con problemática de salud mental, drogodependencia o bajo nivel socioeconómico. Entonces, efectivamente, la herramienta otorga un mayor nivel de riesgo como consecuencia de tales circunstancias puesto que los datos avalan dicha correlación. Y, de hecho, desde la perspectiva de la lógica interna de la herramienta, es necesario tener en cuenta dicha información para que las predicciones sean fiables. Sin embargo, ¿es legítimo que este tipo de condiciones o situaciones puedan dar lugar a peores pronósticos de reincidencia y, en consecuencia, mayores restricciones de derechos? ¿El hecho de que los datos estadísticos revelen que la mayoría de las personas que cometen delitos son pobres o tienen escasos recursos legitima que esa información se utilice para determinar aspectos relacionados con la intensidad del castigo penal?

No solo deben cuestionarse los efectos de los datos estadísticos disponibles, sino que cabe el cuestionamiento de la propia información estadística, teniendo en cuenta que se trata de datos del pasado en los que están presentes las discriminaciones estructurales que, como hace décadas que se advierte desde la criminología crítica, hacen que el sistema penal opere de forma *selectiva* (Baratta, 2004). En consecuencia, ¿se predice un mayor riesgo de reincidencia en una persona pobre porque objetivamente esta circunstancia hace más probable que delinca? ¿O, quizás, es que el mayor riesgo de reincidencia se nutre de que, proporcionalmente, se condena más a personas pobres porque determinados barrios se vigilan más que otros y/o porque cometen delitos cuya investigación es más sencilla?

Es obvio que las desigualdades sociales y las discriminaciones mencionadas no se originan por el uso de instrumentos de valoración del riesgo, y que están presentes también en el juicio clínico. Sin embargo, en la valoración del riesgo, al convertirlas en factores de riesgo, no es posible obviarlas (Brandariz García, 2016), y ello conduce a exacerbarlas, reproducirlas y perpetuarlas, contribuyendo a la estigmatización de las personas más desfavorecidas y, finalmente, castigando la miseria (Martínez Garay, 2024[154]). Por otra parte, al mismo tiempo que tienen influencia como factores de riesgo, sin embargo,

154. Según indica Martínez Garay (2024): "parece que estas prácticas generan al menos en alguna medida un efecto de profecía autocumplida" (p. 176).

quedan ocultas tras el algoritmo. De esta forma, la ciencia y la tecnología permiten *blanquear* el carácter discriminatorio de las predicciones: las formas de discriminación permanecen, pero se invisibilizan y, así, el sistema discriminatorio queda legitimado tras haber pasado por el filtro de lo científico (Martínez Garay, 2024[155]).

Por último, otro aspecto positivo del uso de herramientas de valoración de riesgos frente al juicio clínico es que, tratándose en ambos casos de predicciones, debe admitirse que el análisis del pasado en el que se sustenta la valoración del riesgo tiene una base empírica más sólida y es científicamente más fiable. Es decir, teniendo en cuenta que toda predicción se basa en información sobre el pasado, desde una perspectiva científica, la metodología actuarial ofrece más y mejor información para la elaboración de las predicciones (Hogan & Corăbian, 2025).

Sin embargo, hay varias consideraciones críticas que ponen en cuestión esta aparente ventaja. La mejora de la calidad científica del método predictivo hace que la valoración del riesgo quede revestida de unas notas de objetividad, neutralidad e, incluso, infalibilidad, que no son atribuibles a ningún tipo de predicción —ni a las realizadas más artesanalmente ni a estas—. Es decir, estas herramientas ofrecen una apariencia de cientificidad y, con ello, de mayor objetividad y neutralidad que hacen que se genere un optimismo "desbordante" y "exagerado" respecto de sus posibilidades (Martínez Garay, 2024, p. 168; Urruela Mora, 2025, p. 515).

En primer lugar, hay que tener en cuenta que muchos de los factores de riesgo que se utilizan en las herramientas predictivas, lejos de consistir en valoraciones objetivas, incorporan una buena dosis de subjetividad y, con ello, también de ambigüedad. En este sentido, por ejemplo, resulta revelador conocer cuáles son los factores de riesgo que tiene en cuenta RisCanvi, cómo se definen y qué indicaciones se dan a los/as profesionales para valorarlos[156]. Algunos de los factores con mayor carga subjetiva en esta herramienta son

155. "los algoritmos corren el riesgo de 'blanquear' tecnológicamente (*tech-washing*) los sesgos existentes (…). De este modo, los sesgos permanecen, pero se hacen invisibles, lo que por un lado legitima el sistema (ocultando que al castigar más severamente a determinados colectivos los discrimina injustamente), y por otro puede reforzar los sesgos: si, por ejemplo, el algoritmo otorga mayor riesgo a los pobres, éstos recibirán un trato más severo, lo que probablemente reforzará los factores que determinarán que, en la siguiente ocasión, su riesgo vuelva a ser alto o incluso superior" (Martínez Garay 2024, p. 112).

156. El listado completo de ítems que se tienen en cuenta en la escala completa de RisCanvi, su definición y las indicaciones para su interpretación y apreciación, disponible en:

la "ausencia de planes viables de futuro"[157], las "actitudes pro-criminales o valores antisociales"[158] o la "pertenencia a grupos sociales de riesgo delictivo, diferentes de una banda delictiva"[159]. Esto permite cuestionar la supuesta mayor objetividad de las valoraciones algorítmicas frente a los juicios clínicos de la peligrosidad.

En segundo lugar, como ha señalado Martínez Garay en varias de sus publicaciones, el número de errores en estas predicciones es muy elevado y no disminuye de forma significativa respecto de los que se generan a través del juicio clínico, especialmente, en relación con los falsos positivos (Martínez Garay, 2014, 2020 y 2024).

Esta subjetividad y estos errores estaban presentes en las predicciones anteriores (Martínez Garay, 2024[160]), pero, mientras el punto de partida antes

---

https://repositori.justicia.gencat.cat/bitstream/handle/20.500.14226/1324/escala-valoracio-riscanvi-completa.pdf [último acceso: 18/12/2025].

157. Para determinar si concurre o no este factor de riesgo, el protocolo ofrece la siguiente explicación: "Incapacidad o falta de voluntad para plantearse y llevar a cabo, a medio y largo plazo, planes de ocupación o integración psicosocial realistas (trabajo, familia, desarrollo personal, ocio, etc.). Tiende a vivir exclusivamente al día y cambia sus planes frecuentemente. No da importancia ni se preocupa por el futuro. No está interesado en tener un trabajo estable. Puede llevar una vida nómada y describirse a sí mismo como un 'trotamundos'. En ocasiones puede expresar objetivos de futuro poco realistas, como por ejemplos ser abogado, escritor, cirujano, trabajador social, psicólogo, piloto de avión, etc." (traducción propia del catalán al castellano).

158. Para determinar si concurre o no este factor de riesgo, el protocolo ofrece la siguiente explicación: "Para codificar este factor de riesgo se tiene que determinar en qué medida las actitudes,valores, creencias y pensamientos predominantes del individuo son antisociales o prosociales.
Se tienen que explorar las actitudes del sujeto hacia las otras personas, las instituciones yservicios sociales, la ley y las autoridades. Se tienen que evaluar sus actitudes respeto los actosdelictivos cometidos en el pasado y si el individuo transmite señales de arrepentimiento sincero o, sipor el contrario, no siente remordimientos, es insensible y le falta empatía hacia posiblesvíctimas" (traducción propia del catalán al castellano).

159. Para determinar si concurre o no este factor de riesgo, el protocolo ofrece la siguiente explicación: "Pertenece agrupos sociales (diferentes a una banda delictiva, definida al ítem 25) que pueden estar en riesgo de cometer actos delictivos (por ejemplo, se trata de un joven marginal, no tiene domicilio conocido, es traficante o distribuidor de drogas, participa en actividades relacionadas con la prostitución, etc.)" (traducción propia del catalán al castellano).

160. De acuerdo con Martínez Garay (2024): "probablemente dirían, además, que criticar los algoritmos por estos motivos es hipócrita, ya que la misma información (aunque sin llamarla 'factores de riesgo') es la que siempre se ha tenido en cuenta para hacer evaluaciones de peligrosidad" (p. 115).

era la desconfianza hacia la eficacia de las predicciones, con estos avances metodológicos en el análisis del pasado, prima el optimismo. A pesar de que lo que tenemos entre manos siguen siendo solo predicciones basadas en la toma en consideración de aspectos relacionados con la personalidad o el modo de vida de los sujetos —y que, por tanto, siempre van a resultar en alguna medida inciertas—, se confía más que nunca en su acierto. Dicho optimismo puede influir en los procesos decisorios, de tal forma que el agente que debe tomar determinadas decisiones, aunque no esté obligado a decidir en el sentido marcado por la herramienta, tienda a pegarse a su criterio y lo haga sin incorporar otra motivación, generándose así un sesgo de automatización e, incluso, de desresponsabilización (MacKay, 2020; Varona Gómez, 2024).

En última instancia, desde la perspectiva jurídica, el problema más grave que se deriva de que las predicciones de reincidencia pasen de elaborarse a través del juicio clínico a realizarse por medio de la valoración del riesgo es la mayor dificultad para cuestionarlas, tanto técnica como jurídicamente, y así poder defender frente a ellas los derechos de las personas privadas de libertad.

Desde una preocupación genuina por esta cuestión —por haber experimentado en primera persona la dificultad de defender los intereses de una persona privada de libertad frente a predicciones de riesgo[161]—, en 2022, realicé un estudio sobre el uso de RisCanvi en la toma de decisiones penitenciarias (Alemán Aróstegui, 2023). El interrogante que más me interesaba despejar en aquel momento era si los resultados de RisCanvi podían estar resultando determinantes del sentido de las decisiones. Y, además, quería

161. Hace algunos años, antes de dedicarme al mundo académico, ejercía como abogada en Catalunya, y me encontré con RisCanvi en una resolución judicial. Era un caso en el que la Junta de Tratamiento de un centro penitenciario había informado favorablemente de la concesión de un primer permiso de salida a un preso y, sin embargo, el Juzgado de Vigilancia Penitenciaria no había autorizado el permiso por la gravedad del delito (homicidio) y la lejanía de las tres cuartas partes (típicos argumentos estereotipados que se utilizan como automatismos para denegar posibilidades de mayor contacto con la libertad, que la normativa penitenciaria no contempla y que la Audiencia Provincial competente en este caso rechazaba). Recurrí la resolución con base en los múltiples argumentos que ofrecían los informes del equipo, que destacaban la buena evolución del sujeto y que estaba preparado para el inicio del contacto con la vida en libertad. Mi sorpresa llegó cuando la Audiencia denegó el permiso porque, a pesar de —cito textualmente— "la progresión tan favorable que ha tenido el interno", como había alegado el Ministerio Fiscal, la escala RisCanvi arrojaba niveles de riesgo medios y altos en reincidencia delictiva y quebrantamiento de condena.

entender mejor cómo funcionaba y cómo se estaba utilizando la herramienta para saber cómo podían discutirse, técnica y jurídicamente, sus resultados.

En el estudio se evidenció que RisCanvi no determina el sentido de las decisiones de los/as profesionales de la Administración, dado que existe una directriz de la Dirección General clara en este sentido y, además, porque lo prohíbe el art. 6.1 RP[162]. Sin embargo, los/as profesionales entrevistados/as explicaban que, cuando no están de acuerdo con los niveles de riesgo —a pesar de que se trata de personas expertas en ciencias del comportamiento y que conocen a las personas presas a las que se evalúa—, tienen reticencias a la hora de corregirlos (el protocolo indica que dichas correcciones deben realizarse, pero se hacen en muy pocos casos) porque no saben cómo se han generado ni se entienden legitimados para negar un dato obtenido a través de un procedimiento científico que ha valorado numerosas variables y una cantidad ingente de información. Por otra parte, en el caso de los operadores jurídicos, el estudio sirvió para mostrar que el escenario que se dibuja es muy diferente. En el ámbito judicial, los operadores jurídicos no reciben formación —en algunos casos, ni siquiera, información— sobre RisCanvi y no operan sobre ellos las directrices de la Administración. De manera que, cuando el resultado de RisCanvi sale de la Administración —donde funciona dentro de una lógica y un modelo— y llega al ámbito judicial, adquiere una entidad propia y descontextualizada, y tiende a apreciarse aisladamente, de tal forma que se erige como un dato aparentemente objetivo y científico que difícilmente se puede ignorar ni cuestionar, y que, entonces sí, puede llegar a resultar determinante del sentido de las decisiones.

Pero el mayor problema es en realidad si —más allá de que pueda utilizarse para la toma de decisiones— es posible discutirlo jurídicamente. ¿Cómo puede un/a abogado/a cuestionar ese resultado de riesgo que, en última instancia, constituye una mera predicción? Si dicha predicción se realiza a través del juicio clínico, al contenido del informe experto podría oponerse el informe de otra persona experta. Sin embargo, ¿qué puede oponerse al resultado de riesgo que arroja un algoritmo? Si bien los juicios clínicos pueden ser objeto de numerosas críticas, al menos, los motivos por los que determinan la probabilidad de delinquir se relatan, quedan expuestos y son susceptibles de ser discutidos (MacKay, 2020). Sin embargo, en el caso de estas herramientas algorítmicas, la opacidad del algoritmo mantiene ocultos

162. Art. 6.1 CP: "Ninguna decisión de la Administración penitenciaria que implique la apreciación del comportamiento humano de los reclusos podrá fundamentarse, exclusivamente, en un tratamiento automatizado de datos o informaciones que ofrezcan una definición del perfil o de la personalidad del interno".

los motivos concretos en los que se basa la predicción, en detrimento del principio de contradicción. Ello hace que "el derecho de defensa sea posiblemente el derecho que más gravemente puede verse afectado por el uso de sistemas algorítmicos" porque "el desconocimiento del funcionamiento del algoritmo impida la adecuada impugnación de éste" (De Luis García, 2023, p. 6-7).

4. CONCLUSIONES

El uso de instrumentos algorítmicos en el ámbito punitivo todavía es muy reducido y se circunscribe, casi exclusivamente, al sistema penitenciario, pero previsiblemente su uso se extenderá en el futuro (Urruela Mora, 2025). Por eso, resulta fundamental estudiarlos, analizarlos desde perspectivas críticas y elaborar propuestas que puedan disminuir algunos de los peligros que traen consigo.

Como se ha visto, en la actualidad, la normativa penal y penitenciaria prevé la toma de decisiones con base en predicciones, y en esa medida, estos instrumentos encuentran espacios donde pueden ser utilizados. No solo eso, desde una perspectiva técnica, puede resultar recomendable utilizarlos, sustituyendo otros métodos científicamente peor respaldados. Siendo así, la primera conclusión es que, mientras se utilicen, debe hacerse siguiendo las recomendaciones de los organismos internacionales[163] y deben tratar de solventarse, en la medida de lo posible, las carencias que presentan. En primer lugar, resulta necesario que la implementación de este tipo de herramientas se vea acompañada del incremento de recursos personales suficientes para que se lleven a cabo correctamente y ello no vaya en detrimento de la intervención psicosocial que resulta esencial en las cárceles. En segundo lugar, deben aprobarse normas legales que expresamente establezcan que los resultados de riesgo no pueden determinar automáticamente el sentido de las decisiones y que impidan que los órganos judiciales puedan basar sus decisiones únicamente en sus resultados. Asimismo, debe ofrecerse información y formación a los operadores jurídicos sobre la valoración del riesgo y sus herramientas (De Luis García, 2023, p. 8). En tercer lugar, debe mejorarse la explicabilidad de los resultados de riesgo[164] para que,

163. En este sentido, especialmente, vid. la Recommendation CM/Rec(2024)5 of the Committee of Ministers to member States regarding the ethical and organisational aspects of the use of artificial intelligence and related digital technologies by prison and probation services.

164. Así se recomendó en la auditoría externa que se realizó sobre RisCanvi: "Explicabilidad. El uso de este tipo de algoritmos permite también entender cómo se ha obte-

en su caso, sea posible cuestionarlos jurídicamente por abogados/as y/o órganos judiciales (Martínez Garay, 2024; Cueto Santa Eugenia, 2025). Por último, para cumplir con las exigencias de transparencia y con el derecho de acceso a datos personales y evitar su indefensión, debe informarse a las personas privadas de libertad de que son objeto de evaluación a través de herramientas algorítmicas —las cuales tienen en cuenta rasgos de su personalidad y cuestiones relativas a su vida privada y familiar— y de cuáles son sus resultados y posibles consecuencias[165].

La segunda conclusión tras esta aproximación crítica al uso de instrumentos algorítmicos en la ejecución penitenciaria es que la valoración del

---

nido el valor final: qué factores han influido a asignar un nivel de riesgo y con qué grado. El algoritmo determina automáticamente cuáles son los factores importantes para calcular la probabilidad de reincidencia. Si se incluyera esta información en el sistema RisCanvi, ello permitiría dar una explicación del riesgo predicho con la herramienta al equipo evaluador. Los algoritmos, igual que las personas, se pueden equivocar. Por eso, es crucial que el personal que toma decisiones según las predicciones de los algoritmos RisCanvi tenga una intuición de cómo se ha generado la predicción. Acompañar el nivel de riesgo con una explicación de los factores más relevantes fomentaría el espíritu crítico de los profesionales, que podrían decidir de manera fundamentada si están de acuerdo con el nivel predicho por el algoritmo o no. De este modo, el algoritmo dejaría de ser una caja negra en la cual confiar ciegamente. A pesar de que el artículo 22 del GDPR no se aplica en este caso de uso, puesto que siempre habrá una validación manual por un humano, es recomendable contar con herramientas para justificar la toma de decisiones, no solo para el personal evaluador sino también a internos o jueces que lo pudieran pedir" (Dribia Data & Direcció General d'Afers Penitenciaris, 2024, pp. 47 y ss.) (traducción propia del catalán al castellano).

165. En este sentido, en el informe coordinado por Martínez Garay, se apunta que en el acceso a dicha información debe regir igual normativa que la relativa al acceso al expediente penitenciario. En consecuencia, se lleva a cabo un análisis detallado de la evolución en la problemática del acceso al expediente penitenciario que resulta de gran interés porque tradicionalmente había sido muy limitado y, en los últimos años, ha sido ampliado (2024, pp. 120-124). De hecho, en los recientemente publicados *Criterios de Actuación, Conclusiones y Acuerdos Aprobados por los Jueces de Vigilancia Penitenciaria en sus XXXIII Reuniones celebradas entre 1981 y 2025*, se incluye el siguiente criterio: "53.- Información del interno: acceso al expediente. Hay que partir del contenido del art. 15 bis de la LOGP, introducido por la L.O. 7/2021 de 26 de mayo, teniendo en consideración la STC 164/2021 de 4 de octubre. Conforme a lo establecido en el apartado 1 del art. 15 bis de la LOGP los internos admitidos en un establecimiento penitenciario tienen derecho a acceder al expediente personal abierto al efecto, relativo a su situación personal y penitenciaria, que solo podrá ser limitado por razones de seguridad y tratamiento individualmente valoradas (párrafo aprobado en la reunión de Valladolid 2024 y reiterado en la de Almería 2025)".

riesgo tiene en común con la peligrosidad que en ambos casos se trata de predicciones de comportamientos delictivos que se tienen en cuenta para la toma de decisiones en el ámbito punitivo. Como se ha visto, en términos técnico-científicos, la valoración del riesgo integra metodologías más fiables de análisis del pasado y, en esa medida, su utilización genera ventajas respecto del juicio clínico. Sin embargo, siguen siendo predicciones y, en esa medida, siempre incorporan un grado de incertidumbre[166]. Y, además, también dan lugar a otros problemas, como un exceso de optimismo respecto de su fiabilidad que contribuye al refuerzo del uso de predicciones para la toma de decisiones. Es decir, favorece la proliferación de decisiones punitivas basadas en pronósticos de futuro, en lugar de atenerse a tomar como referencia el hecho delictivo.

Así ha ocurrido en Estados Unidos donde, en el marco de la mejora y el perfeccionamiento de los instrumentos de valoración de riesgos, estos han pasado de utilizarse únicamente en la *probation* y la *parole*, para llegar ya hace unos años al ámbito de la determinación de la pena de sujetos imputables a través de la *evidenced-based sentencing* o *smart sentencing* (Martínez Garay, 2018, p. 488). Como apunta Urruela Mora, "con base en la pretendida fiabilidad de dichos juicios de pronóstico, las valoraciones de riesgo pasan a ocupar un rol central en el sistema de justicia penal" (2025, p. 521). Por tanto, aunque en nuestro entorno todavía no se ha llegado a estos extremos, debe prestarse atención a la posibilidad de que este tipo de instrumentos se extiendan a otros ámbitos del sistema penal, afectando a la determinación del castigo, no solo en cuanto a su intensidad, sino también en cuanto a su extensión. El riesgo de expansión del uso de predicciones delictivas para la toma de otras decisiones en el ámbito punitivo es considerable. De hecho, ya contamos con una medida de seguridad aplicable a imputables que permite ampliar el control penal más allá de la medida de la culpabilidad en función del riesgo de reincidencia (la libertad vigilada), y con una pena de prisión de duración indeterminada que somete la posibilidad de definir su duración a un juicio pronóstico de reincidencia[167].

166. Urruela Mora (2025) va más allá y señala que "la ciencia está oscureciendo lo que, en realidad, es una cuestión de arte, en el sentido de que la predicción de los comportamientos antisociales futuros de una persona requiere de una fuerte dosis de imaginación. En consecuencia, emplear el calificativo de test científicos para referirse a los instrumentos actuariales de predicción de la peligrosidad contradice la realidad de las cosas haciendo olvidar la intrínseca incompetencia actual para predecir la conducta humana" (p. 521).

167. De hecho, ya se han elaborado propuestas sobre cómo utilizar estas herramientas en el sistema de revisión de la prisión permanente revisable (Martínez Garay, 2023).

La cuestión más problemática de la valoración del riesgo es la proliferación del uso de predicciones porque choca con el principio de culpabilidad (Varona Gómez, 2024) a través de una cada vez mayor "erosión del principio del hecho" (Martínez Garay, 2019, p. 116). Con el uso de predicciones, las decisiones punitivas se alejan de responder a lo que el sujeto *ha hecho* y, en su lugar, lo que se valora es lo que el sujeto *es*, su forma de vida o las circunstancias en las que le ha tocado vivir[168]. Y, de esta forma, se abandona "el marco de una referencia segura y retrospectiva (el injusto culpable cometido) para reemplazarla —al menos en parte— por una variable prospectiva y absolutamente incierta" (Urruela Mora 2025, p. 519). De hecho, la propuesta es la posibilidad de desconexión total con el hecho, planteándose que lo que se pretende es detectar a "aquellos delincuentes que están en el circuito de la ejecución penal por haber cometido un delito no violento, pero que pueden reincidir en un nuevo delito y éste sí ser violento" (Andrés Pueyo, 2013, p. 487). Desde la perspectiva jurídico-penal, este tipo de pretensiones resultan absolutamente rechazables.

Todo esto, en definitiva, a lo que conduce es al favorecimiento del retorno del Derecho penal de la peligrosidad, del que se viene advirtiendo en la doctrina penal desde hace décadas (Hassemer, 1992; Donini, 2003; Portilla Contreras, 2007; Del Rosal Blasco, 2009; Massaro, 2023).

También en el siglo XIX el positivismo criminológico se defendió como la manera de introducir en la ciencia penal el conocimiento empírico, de tal forma que el Derecho penal tuviera una base más científica, que lo hiciera compatible con el avance tecnológico del momento, aun en detrimento de los axiomas penales clásicos (Quintero Olivares, 1999). Como apunta Sotomayor Acosta, "hoy se llama 'clásico' al Derecho penal liberal en el mismo sentido peyorativo que lo hacía Ferri respecto del Derecho penal de Carrara" (2008, p. 153).

Si bien —como se ha visto—, en la actualidad, la peligrosidad ha quedado denostada desde la perspectiva científica, a través del concepto de riesgo, vuelve a reivindicarse la necesidad de que el Derecho penal se adelante al hecho delictivo, adopte una estrategia predominantemente preventiva, y adapte sus respuestas a predicciones basadas en características personales y

168. Además, como apunta Urruela Mora (2025): "un gran número de ítems que determinan mayor peligrosidad criminal del sujeto tales como la concurrencia de patología mental, la minoría de edad o la dependencia patológica de sustancias constituyen factores que, de concurrir en el momento de comisión del hecho delictivo, aminoran la culpabilidad y, en consecuencia, la responsabilidad penal" (p. 520).

circunstancias familiares, sociales, etc., destacando de nuevo el valor científico de estas. Así, a través de las nuevas tecnologías y con la formulación de nuevos conceptos y el uso de un lenguaje innovador, se elaboran propuestas por las que se produce, en realidad, la vuelta a viejos esquemas (Donini, 2003; Massaro, 2023). Frente a la idea de una propensión natural al crimen, hoy en día, se plantea una propensión estadística a cometer delitos[169]. Se trata, en cualquier caso, de una propensión que permite conformar una predicción y, en definitiva, se trata de tomar el futuro como referente fundamental del castigo penal. Y, así, se asientan las bases para la indeterminación de los castigos penales e, incluso, para la legitimación de la represión sin delito (Quintero Olivares, 2017).

Por tanto, la invitación es a hacer memoria para evitar, como dijo Garland, "perder todo sentido de los procesos históricos de los cuales emergieron nuestras estructuras actuales" (2001), y para impedir que detrás de la innovación se escondan tendencias regresivas siempre amenazadoras de los derechos conseguidos (Zaffaroni, 2017[170]). En este sentido, es necesario defender la "función irrenunciable" del Derecho penal de la culpabilidad,

169. No obstante, también se constata un resurgimiento de aquella propensión natural sobre la base de los avances neurocientíficos. En este sentido, Guzmán Fluja (2022) señala lo siguiente: "otra tendencia incipiente, relacionada con la, cada vez mayor, interdependencia entre distintas disciplinas, tiene que ver con la posibilidad de detección de trastornos mentales en una persona, y la posibilidad de determinar sus posibles tendencias delictivas iniciales o de reincidencia (…). Hablo de lo que se conoce como 'neuropredicción'" (p. 48). Este autor plantea que "la neurocriminología puede servir como base para el diseño de instrumentos de evaluación de riesgos de comisión de hechos delictivos, con una vertiente predictiva y una vertiente preventiva", sobre la base del "reconocimiento de la importancia que tienen los factores biológicos en la determinación de la conducta criminal presente y futura, partiendo de la mayor o menor tendencia a la conducta violenta o al comportamiento agresivo" (Guzmán Fluja, 2022, pp. 60-61). Asimismo, Guzmán Fluja (2022) aprecia que "la neurociencia adquiere, con sus luces y sus sombras, un protagonismo en el ámbito jurídico penal y procesal penal que pocos hubieran imaginado a principios de este siglo XXI" (p. 60). En cambio, me atrevo a pensar que quizás sí hubiera podido ser perfectamente imaginado desde la concepción positivista del siglo XIX.

170. Zaffaroni (2017) apunta la regresión es una constante en el Derecho penal, en el que se observa una "dialéctica" permanente entre el Estado de Derecho y el Estado de policía que tantas veces se ha decantado en favor del segundo. Y, en este sentido, reclama la importancia de analizar los distintos "discursos del peligrosismo" que existieron en el pasado, por ejemplo, en los textos inquisitoriales, en el positivismo racista o en el Derecho penal nazi, porque considera que "han sido los más finamente elaborados y teorizados con las más insólitas perversiones racionalizantes, que le dieron increíble eficacia letal" (p. 157).

que consiste en servir "de frontera frente a otras concepciones del derecho penal conocidas en la historia en las que el respeto a los derechos del individuo cedía en nombre de principios defensistas o de interés del pueblo o del Estado" (Quintero Olivares, 1999, p. 200).

Es fundamental tener en cuenta que el ámbito punitivo no es cualquier ámbito, sino que es aquel en el que el Estado tiene la posibilidad de imponer las máximas restricciones de derechos y libertades. Lo que está en juego es el castigo penal y, por tanto, los derechos y libertades. Y, como sabemos quienes nos dedicamos al Derecho penal, por eso, aunque el castigo penal se legitime a través de fines preventivos, existen unos límites jurídicos que son infranqueables, y que en la actualidad se cifran en los derechos fundamentales y los principios constitucionales. En este marco, nuestro Derecho penal es y debe ser un Derecho de la culpabilidad por el hecho en el que debe respetarse el principio de igualdad y el libre desarrollo de la personalidad.

La última conclusión, con la que cierro esta contribución crítica al uso de algoritmos predictivos en el sistema penal, es de tipo propositivo. Los límites jurídicos que deben establecerse al uso de predicciones algorítmicas en el ámbito punitivo no niegan la posibilidad —y la necesidad— de continuar desarrollando este tipo de herramientas y tomar muy en cuenta la información que ofrecen. El análisis de riesgos a través de algoritmos constituye, sin duda, el resultado de avances científicos muy importantes y trae consigo una mejora en el fundamento empírico de las predicciones, esto es: una mejora considerable de la información de la que disponemos sobre el pasado, al estar basada en parámetros científicos más sólidos. La información que se obtiene a través de estos algoritmos permite tener un conocimiento más consistente de cómo *ha sido* hasta ahora la delincuencia en nuestra sociedad, y lo que revela es el gran peso que tienen las desigualdades estructurales en la conducta delictiva (Martínez Garay, 2019[171] y 2024). Tomemos esta información como referencia, no para tratar de predecir cómo se comportará en el futuro un concreto individuo y, en su caso, restringir sus derechos, sino para desarrollar e implementar políticas sociales transformadoras, que traten de cambiar los factores estructurales en los que sabemos que se asienta el fenómeno delictivo, de tal forma que se evite así el efecto performativo de la profecía autocumplida, y existan mayores probabilidades

171. Siguiendo a Martínez Garay (2019): "En ellas aparece negro sobre blanco que el nivel de peligrosidad que se asigna a los sujetos depende en buena medida de un conjunto de factores de tipo étnico, socioeconómico, educativo, familiar y laboral, y dicen abiertamente que la marginalidad, la pobreza y la miseria incrementan el riesgo" (pp. 186-187).

de evitar la comisión de delitos en el futuro. En definitiva, la propuesta es: ¿y si ponemos esta tecnología al servicio del desarrollo de políticas sociales transformadoras que verdaderamente ofrezcan posibilidades de prevenir los delitos?

5. REFERENCIAS

Alemán Aróstegui, L. (2025): *La ejecución de la pena de prisión desde la perspectiva de los límites al ius puniendi,* Aranzadi, Madrid.

Alemán Aróstegui, L. (2023): "El uso de RisCanvi en la toma de decisiones penitenciarias", en *Estudios Penales y Criminológicos*, vol. 44 *Extra*, pp. 1-43.

Allen, R.M. (1942): "A review of parole prediction literature", en *Journal of Criminal Law and Criminology*, vol. 32, pp. 548-554.

Andrews, D. & Bonta, J. & Hoge, R.D. (1990): "Classification for Effective Rehabilitation. Rediscovering Psychology", en *Criminal and Justice Behavior*, vol. 17, n. 1, pp. 19-52.

Andrés Pueyo, A. (2013): "Peligrosidad criminal: análisis crítico de un concepto polisémico", en Demetrio Crespo, E. (dir.), *Neurociencias y Derecho Penal. Nuevas perspectivas en el ámbito de la culpabilidad y tratamiento jurídico-penal de la peligrosidad*, Edisofer-BdF, Montevideo-Buenos Aires, 483-503.

Andrés Pueyo, A. & Redondo Illescas, S. (2007): "Predicción de la violencia: entre la peligrosidad y la valoración del riesgo de violencia", en *Papeles del Psicólogo*, vol. 28, n. 3, pp. 157-173.

Baratta, A. (2004): *Criminología crítica y crítica del derecho penal. Introducción a la sociología jurídico-penal*, Siglo XXI, Buenos Aires.

Beccaria, C. [1764] (1982): *De los delitos y de las penas,* 3ª ed., Alianza, Madrid, 1982.

Brandariz García, J.Á. (2016): *El modelo gerencial-actuarial de la penalidad: eficacia, riesgo y sistema penal,* Dykinson, Madrid.

Castro Liñares, D. (2019): *Los instrumentos de valoración y gestión de riesgos en el modelo de penalidad español*, Reus, Madrid.

Cervelló Donderis, V. (2019): "Individualización garantista en el ejercicio de la discrecionalidad penitenciario", en *Anuario de Derecho Penal y Ciencias Penales*, vol. LXXII, pp. 217-264.

Cobo del Rosal, M. & Vives Antón, TS (1999): *Derecho Penal. Parte General*, 5ª ed., Tirant lo Blanch, Valencia.

Cueto Santa Eugenia, E. (2025): "Herramientas de predicción criminal y su impacto en nuestro proceso penal", en *Ius et Scientia*, vol. 11, n. 1, pp. 14-31.

De Giorgi, A. (2000): *Zero Tolleranza. Strategie e pratiche della società di controllo*, DeriveApprodi, Bolonia.

De Luis García, E. (2023): "Justicia, inteligencia artificial y derecho de defensa", en *Revista d'Internet, Dret i Política (IDP)*, n. 39, pp. 1-12.

Del Rosal Blasco, B. (2009): "¿Hacia un Derecho penal de la postmodernidad?", en *Revista Electrónica de Ciencias Penales y Criminológicas*, n. 11, pp. 1-64.

Díez Ripollés, J.L. (1997): "El bien jurídico protegido en un Derecho penal garantista", en *Jueces para la Democracia*, n. 30, pp. 10-19.

Donini, M. (2019): *Populismo e ragione publica. Il post-illuminismo penale tra lex e ius*, Mucchi Editore, Módena.

Donini, M. (2011): "El problema del método penal: de Arturo Rocco al europeísmo judicial", en *Nuevo Foro Penal*, vol. 7, n. 76, pp. 49-100.

Donini, M. (2003): "¿Una nueva edad media penal? Lo viejo y lo nuevo en la expansión del derecho penal económico", en *Nuevo Foro Penal*, n. 65, pp. 92-119.

Dribia Data & Direcció General d'Afers Penitenciaris (2024): *Informe Tiresias: Auditoria de l'algorisme RisCanvi,* Generalitat de Catalunya, Departament de Justícia, Drets i Memòria, Barcelona.

Feeley, M.M. & Simon, J. (1992): "The New Penology: Notes on the Emerging Strategy of Corrections and its Implications", en *Criminology*, n. 30, pp. 449-474.

Ferrajoli, L. (1986): "El Derecho penal mínimo", en *Poder y control*, n. 0, pp. 25-48.

Gargallo Vaamonde, L. (2011): *El sistema penitenciario de la Segunda República: antes y después de Victoria Kent (1931-1936). Premio Nacional Victoria Kent, 2010*. Ministerio del Interior, Madrid.

Guzmán Fluja, V.C. (2022): "Evaluación de riesgos de reincidencia delictiva en personas con trastorno mental a la luz de recientes tendencias", en

Flores Prada, I. (dir.)/Montoro Sánchez, J.A./Sánchez Rubio, A. (coords.), *Discapacidad y riesgo de los sujetos pasivos con trastorno mental en la justicia penal. Estudios sobre el Anteproyecto de Ley de Enjuiciamiento Criminal de 2020*, Aranzadi, Pamplona, 45-74.

Grupo de Estudios de Política Criminal (2016): *Revisión y actualización de las propuestas alternativas a la regulación vigente. 25 años de trabajo del Grupo de Estudios de Política Criminal*, Tirant lo Blanch, Valencia.

Harcourt, B.E. (2007): *Against Prediction. Profiling, Policing, and Punishing in an Actuarial Age*, The University of Chicago Press, Chicago.

Hassemer, W. (1992): "Rasgos y crisis del Derecho Penal moderno", en *Anuario de Derecho Penal y Ciencias Penales*, vol. XLV, n. 1, pp. 235-250.

Hassemer, W. & Muñoz Conde, F. (1989): *Introducción a la criminología y al derecho penal*, Tirant lo Blanch, Valencia.

Hogan N.R. & Corăbian, G. (2025): "Generative Artificial Intelligence in Violence Risk Assessment: Emerging Technology and the Ethics of the Inevitable", en *Behavioral Sciences & the Law*, vol. 43, n. 5, pp. 1-10.

Lascuraín Sánchez, J.A. (2021): *Principios penales democráticos*, Iustel, Madrid.

Luzón Peña, D-M. (2025): *Lecciones de Derecho Penal. Parte General*, 4ª ed., Tirant lo Blanch, Valencia.

MacKay, C. (2020): "Predicting risk in criminal procedure: actuarial tools, algorithms, AI and judicial decision-making", en *Current Issues in Criminal Justice*, vol. 32, n. 1, pp. 22-39.

Martínez Garay, L. (coord.) (2024): *Three predictive policing approaches in Spain: VIOGÉN, RISCANVI and VERIPOL. Assessment from a human rights perspective*, Universitat de València, Valencia.

Martínez Garay, L. (2023): "Revisión con riesgo bajo, y también con riesgo alto: razones para que las valoraciones de riesgo no impidan la revisión de la prisión permanente", en *Revista General de Derecho Penal*, n. 39, pp. 1-32.

Martínez Garay, L. (2020): "Evidence-based sentencing y evidencia científica. A la vez, algunas consideraciones sobre las "políticas basadas en la evidencia" y el derecho penal", en *Teoría y Derecho: revista de pensamiento jurídico (TEODER)*, n. 28, pp. 238-277.

Martínez Garay, L. (2019): "La relación entre culpabilidad y peligrosidad", en Maraver, M./Pozuelo, L. (coords.), *La culpabilidad*, BdeF, Montevideo-Buenos Aires, 115-199.

Martínez Garay, L. (2018): "Peligrosidad, algoritmos y *due process*: el caso *State v Loomis*", en *Revista de Derecho Penal y Criminología*, n. 20, pp. 485-502.

Martínez Garay, L. (2015): "La incertidumbre de los pronósticos de peligrosidad: consecuencias para la dogmática de las medidas de seguridad", en Orts Berenguer, E. (dir.)/Alonso Rimo, A./Roig Torres, M. (coords.), *Derecho Penal de la peligrosidad y prevención de la reincidencia*, Tirant lo Blanch, Valencia, 17-103.

Mir Puig, S. (2015): *Derecho Penal. Parte General*, 10ª ed., Reppertor, Barcelona.

Mir Puig, S. (2003): *Introducción a las bases del Derecho penal*, 2ª ed., BdeF, Buenos Aires.

Miró Llinares, F. & Fuentes Osorio, J.L. (dirs.) (2021): *El Derecho penal ante lo "empírico". Sobre el acercamiento del Derecho penal y la Política Criminal a la realidad empírica*, Marcial Pons, Madrid.

O'Malley, P. (2004): *Risk, Uncertainty and Government*, Glass House Press, Londres.

Otero González, P. (2015): *La libertad vigilada aplicada a ¿imputables? Presente y futuro*, Dykinson, Madrid.

Pérez Manzano, M. (1990): *Culpabilidad y prevención: las teorías de la prevención general positiva en la fundamentación de la imputación subjetiva*, Ediciones de la Universidad Autónoma, Madrid.

Portilla Contreras, G. (2007): *El Derecho Penal entre el cosmopolitismo universalista y el relativismo posmodernista*, Tirant lo Blanch, Valencia.

Quintero Olivares, G. (2017): "Culpabilidad, imputabilidad y capacidad procesal en el derecho penal y en el enjuiciamiento de personas con trastorno mental", en Flores Prada, I. (dir.)/Sánchez Rubio, A. (coord.), *Trastornos mentales y justicia penal. Garantías del sujeto pasivo con trastorno mental en el proceso penal*, Aranzadi, Pamplona, pp. 33-67.

Quintero Olivares, G. (1999): *Locos y culpables*, Aranzadi, Pamplona.

Rivera Beiras, I. (1997): *La devaluación de los derechos fundamentales de los reclusos*, Bosch, Barcelona.

Roldán Barbero, H. (1988): *Historia de la Prisión en España*, PPU, Barcelona.

Roxin, C. (1997): *Derecho Penal. Parte General*, Thompson Civitas, Madrid.

Salat Paisal, M. (2015): *La respuesta jurídico-penal a los delincuentes imputables peligrosos: especial referencia a la libertad vigilada*, Aranzadi, Pamplona.

Simon, J. (1988): "The ideological effects of actuarial practices", en *Law & Society Review*, vol. 22, n. 4, pp. 771-800.

Sotomayor Acosta, J.O. (2008): "¿El Derecho penal garantista en retirada?", en *Revista Penal*, n. 21, pp. 148-164.

Tonry, M. (2019): "Predictions of Dangerousness in Sentencing: Déjà Vu All Over Again", en *Crime and Justice*, vol. 48, n. 1, pp. 436-482.

Tonry, M. (2013): "Sentencing in America, 1975-2025", en *Crime and Justice*, vol. 42, n. 1, pp. 141-198.

Urruela Mora, A. (2025): "Evaluación actuarial de la peligrosidad criminal e Inteligencia Artificial. Implicaciones para el sistema penal", en *De la Ejecución a la Historia del Derecho Procesal y de sus protagonistas. Liber Amicorum en homenaje al Profesor Manuel-Jesús Chacón Cadenas*, Atelier, Barcelona, pp. 513-526.

Varona Gómez, D. (2024): "Algoritmos e inteligencia artificial para en el sistema penal", en *InDret*, n. 4, pp. xii-xv.

Zaffaroni, E.R. (2017): *Derecho penal humano. La doctrina de los juristas y el poder en el siglo XXI*, Hammurabi, Buenos Aires.

Capítulo 11

Instrumentos de predicción de la peligrosidad y ejecución penitenciaria: el avance del derecho penitenciario del enemigo

Ana I. Pérez Machío

Prof.ª Titular (Catedrática acredt.)
Derecho Penal UPV/EHU
Investigadora del Instituto Vasco
de Criminología/Kriminologiaren Euskal Institutua

SUMARIO: 1. REFERENCIAS.

La implementación de instrumentos de predicción de la peligrosidad en el ámbito de la ejecución penitenciaria plantea no sólo ciertos desafíos, sino serias dudas sobre las decisiones que afectan a la restricción de la libertad, en la medida en que las mismas se pueden ver afectadas por las tendencias político-criminales de las últimas décadas.

En efecto, la actual progresión de la Política Criminal de los Estados occidentales hacia el "Populismo punitivo", se caracteriza por el endurecimiento del Derecho Penal y la utilización de esta materia con fines electoralistas, sobre la base del modelo socioeconómico neoliberal, vinculado a la sociedad del riesgo, a la inseguridad ciudadana, al Derecho Penal del enemigo, al Derecho Penal simbólico y, en última instancia, a las dinámicas actuariales en las que se está embarcando el actual Derecho Penal (Pérez Machío/Berasaluze Gerrikagoitia, 2022: 117).

Estas mismas tendencias se vienen observando, en los últimos años, en el ámbito de la ejecución penitenciaria, permitiendo que emerja lo que algunos autores han convenido en denominar como "Derecho Penitenciario del enemigo", que es la plasmación, en el ámbito penitenciario de la

cultura de la emergencia y la excepcionalidad, como vías de regreso hacia el autoritarismo (Alemán Aróstegui, 2023: 8). Este fenómeno opera de la siguiente manera: a los sujetos que han cometido determinados delitos —enemigos— se les atribuye automáticamente una especie de "presunción de peligrosidad" incuestionable (Acale Sánchez, 2016: 164) y, en consecuencia, se introducen obstáculos, a las posibilidades de progresión hacia la libertad que ofrece la normativa penitenciaria, impidiendo la individualización y haciendo prevalecer un enfoque retributivo en esta fase de la pena (como la incorporación desde el 2003 del "período de seguridad"; la utilización de criterios como la gravedad del delito o la lejanía del cumplimiento de las ¾ partes de la condena, o la incorporación desde el 2015 de la prisión permanente revisable) (Cervelló Donderis, 2019: 224).

La consecuencia última, son las condenas cumplidas íntegramente en prisión, sin un contacto progresivo con la libertad, lo que además de resultar contraproducente en términos de reincidencia delictiva, es absolutamente contrario al mandato constitucional de reeducación y reinserción social (Alemán Aróstegui, 2023: 9).

Directamente vinculado a esta tendencia, se advierte la emergencia del "Derecho Penitenciario del riesgo", que constituye una manifestación, en el ámbito penitenciario, de las corrientes actuales que se basan en la introducción de racionalidades propias de la gestión de riesgos: el gerencialismo que consiste en planteamientos de tipo económico y eficientistas para el diseño de política públicas, junto al actuarialismo punitivo (Brandariz García, 2016: 169).

Los modelos actuariales en Derecho Penal usan predicciones estadísticas sustentadas en las tasas de criminalidad de ciertos grupos o de ciertas características grupales, para determinar las medidas jurídico-penales aplicables a concretos individuos de esos grupos que se categorizan como "sujetos de riesgo" (Harcourt, 2013: 85), sobre una base genealógica, en términos familiares, y/o sobre una base factorial, en términos de combinaciones de factores estadísticos, vinculados a una condición, sobre los que hay que intervenir, habida cuenta de la peligrosidad que, para la paz pública, implican. Se trata de estrategias orientadas a identificar los grupos sociales más problemáticos, a fin de incapacitarles, hasta el punto de aislarlos en condiciones de libertad, abandonando la retórica de la reinserción social.

El recurso a las dinámicas actuariales basadas en análisis estadísticos afecta al principio de igualdad y de no discriminación ante el castigo, al generar automáticamente la consideración de "factor de riesgo", sobre aque-

llos sujetos que responden al perfil estadístico elaborado. De este modo, se menoscaban, de forma muy notable, las condiciones de vida de determinados grupos poblacionales y se restringen las posibilidades de reinserción de los individuos pertenecientes a ellos, puesto que el objetivo de la Criminología actuarial no reside en eliminar el delito, sino en minimizar su riesgo (Brandariz García/Fernández Bessa, 2010: 273).

Así, el pensamiento actuarial, orientado, en el sentido manifestado, a la gestión y a la redistribución de los riesgos derivados de la criminalidad, sostiene la necesidad de identificar grupos peligrosos y de concentrar específicamente en tales segmentos sociales los recursos de vigilancia y control, convirtiéndolos en destinatarios prioritarios del sistema penal (Pérez Machío, 2016: 429).

El enfoque de riesgo en el ámbito penitenciario se traduce en la identificación de la persona privada de libertad como "fuente de peligro" a la que hay que controlar para evitar que genere riesgos futuros. Para ello, se recurre a "instrumentos de valoración y gestión de riesgos", cuyo uso se ha justificado por la necesidad de reducir la discrecionalidad y la subjetividad inherente al juicio clínico (Rueda Soriano/Navarro Blasco, 2019: 61). Así, se señala que estos métodos ofrecen criterios científicos amparados por un respaldo empírico —basado en datos estadísticos— y, por tanto, que dotan de objetividad a sus valoraciones (Castro Liñares, 2019: 106). Junto a ello, también se indica que recopilan la información que es relevante para la predicción del riesgo y desechan aquella que no lo es, permitiendo atenuar la lógica dicotómica de la afirmación o negación de peligrosidad, gracias a los grados de probabilidad (Martínez Garay, 2014: 9). Finalmente, también se considera que dan estabilidad al sistema y permiten tanto agilizar el trabajo de los y las profesionales, como reforzar su seguridad en la toma de decisiones.

Ahora bien, el hecho de que en este tipo de instrumentos se utilice el método científico-estadístico y se parta de la evidencia empírica para llevar a cabo las predicciones, puede hacer que parezcan infalibles y, en consecuencia, incuestionables. Sin embargo, varias son las críticas que han de advertirse.

En primer lugar, que la evidencia científica sobre una determinada cuestión no es unívoca y, por otro lado, que los fenómenos sociales y el comportamiento humano difieren de los fenómenos físicos y no pueden ser equiparados en términos metodológicos ni de resultados. En definitiva, que debe evitarse una concepción simplista y reduccionista del comportamiento

humano, partiendo de la premisa de su complejidad y de la imposibilidad de predecirlo (Alemán Aróstegui, 2023: 13).

En segundo lugar, que el número de errores en estas predicciones es muy elevado y no disminuye de forma significativa, respecto del que generan los juicios clínicos (Martínez Garay, 2020, 251). Esto no implica que se trate de malos instrumentos. Todo lo contrario, debe reconocerse que su configuración constituye una muestra de importantes avances científicos. De hecho, por mucho que se mejoren las técnicas predictivas, deben mantenerse todas las cautelas, teniendo en cuenta que los errores son inherentes a cualquier pronóstico respecto del comportamiento de una persona; que tratar de predecir un fenómeno poco frecuente incrementa el número de errores y, finalmente, que no sólo son aspectos técnicos los que condicionan el acierto de la herramienta, sino también cuestiones valorativas, como la fijación del punto de corte, que constituye, en realidad, una decisión de política-criminal (Martínez Garay/Montes Suay, 2018: 35).

En tercer lugar, estos sistemas de vigilancia y gestión de riesgos presentan, *a priori*, una apariencia de objetividad, debido a que son capaces de procesar un gran volumen de datos y cabría pensar que podrían llegar a superar las limitaciones cognitivas del ser humano. Sin embargo, su implementación suscita debates sobre la fiabilidad de sus predicciones, en la medida en que, lejos de eliminar la discriminación, los algoritmos pueden perpetuar las desigualdades existentes si se entrenan con datos históricos sesgados o si sus criterios de evaluación no son adecuadamente supervisados (Cueto Santa Eugenia, 2025: 15). Así, si los operadores jurídicos han trabajado históricamente con prejuicios raciales, socioeconómicos o de otro tipo, los algoritmos que alimentan estas herramientas tienden a replicar y amplificar esas desigualdades, lo que puede traducirse en una vigilancia desproporcionada sobre determinados grupos, reforzando estereotipos y perpetuando la criminalización de sectores vulnerables, como minorías étnicas, personas de bajos recursos o individuos con trastornos mentales (Cueto Santa Eugenia, 2025: 19).

En relación con posibles sesgos preexistentes, así como los riesgos que los mismos implican, hay que destacar el tenor literal del párrafo 42 del Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, cuando indica lo siguiente: "*En consonancia con la presunción de inocencia, las personas físicas de la Unión siempre deben ser juzgadas basándose en su comportamiento real. Las personas físicas nunca deben ser juzgadas a partir de comportamientos predichos por una IA basados únicamente en la elaboración*

de sus perfiles, en los rasgos o características de su personalidad, como la nacionalidad, el lugar de nacimiento, el lugar de residencia, el número de hijos, el nivel de endeudamiento o el tipo de vehículo, sin una valoración humana y sin que exista una sospecha razonable, basada en hechos objetivos comprobables, de que dicha persona está implicada en una actividad delictiva. Por lo tanto, deben prohibirse las evaluaciones de riesgos realizadas con respecto a personas físicas para evaluar la probabilidad de que cometan un delito o para predecir la comisión de un delito real o potencial basándose únicamente en la elaboración de perfiles de esas personas físicas o la evaluación de los rasgos y características de su personalidad […]".

A tenor del mismo, se establece, de forma directa, la prohibición de evaluar el riesgo sobre características y rasgos de la personalidad, en la medida en que se proyecta una imagen social de las personas internas en centros penitenciarios identificada con la "alteridad", con lo peligroso, en definitiva, con una "fuente de riesgo social". En efecto, las evaluaciones de riesgo realizadas sobre determinados rasgos de la personalidad, identifican, discursivamente, a las personas privadas de libertad con "grupos de riesgo", que suscitan sentimientos de inseguridad, entre aquellos sujetos considerados como ciudadanos por el sistema, como consecuencia, únicamente, de su posición económica, social, cultural, ideológica y política. Este sentimiento subjetivo de inseguridad, asociado a la posible puesta en peligro del Estado de bienestar, permite, desde la perspectiva de una parte importante de la ciudadanía, que se legitime el discurso securitario sobre determinados sujetos (los privados de libertad) en beneficio de la parte competitiva de la población (Pérez Machío, 2016: 431) confirmándose, así un Derecho Penitenciario del enemigo, que prioriza las intervenciones inocuizadoras, actuariales y contrarias a la reinserción social.

La lógica actuarial en los sistemas de evaluación del riesgo no sólo deforma, así, la concepción de la pena justa (Harcourt, 2013: 126) —en cuanto desatiende la finalidad rehabilitadora de la misma—, sino que configura un régimen penitenciario alejado del principio de igualdad ante la ley y evidencia un tipo de control penal que se proyecta sobre un sujeto que está modulado por predicciones de riesgo, construidas a partir de rasgos grupales, es decir, de datos de otras personas o grupos de personas que han reincidido, adoleciendo estas predicciones de un déficit de individualización que contradice la orientación resocializadora de la ejecución de la pena de prisión (Alemán Aróstegui, 2023: 15).

Finalmente, la última crítica está relacionada con el uso que se hace de estos instrumentos. Concretamente, es relevante en qué ámbito se utilizan y

qué derechos hay en juego, así como qué incidencia tienen en dicho ámbito: si son, o pueden llegar a ser decisivos. El medio penitenciario se caracteriza precisamente por la restricción del derecho fundamental a la libertad y, en tanto así, constituye un espacio particularmente sensible, en el que la introducción de cualquier elemento debe ser cuidadosamente analizada. Además, existen determinados incidentes que son susceptibles de generar una afectación directa en el nivel de restricción de la libertad, como son las decisiones sobre la concesión de permisos de salida, la clasificación en tercer grado o la concesión de la libertad condicional, entre otros. La existencia de derechos fundamentales en juego debería precipitar el cuestionamiento de si es aceptable que los resultados que ofrecen estas herramientas puedan, si quiera contemplarse en estos procesos decisorios (Alemán Aróstegui, 2023: 15).

En definitiva, estos instrumentos pueden proporcionar información muy valiosa en la práctica para la adopción de determinadas decisiones, pero no se puede perder de vista las limitaciones de los mismos. En efecto, si bien los niveles de precisión son satisfactorios en la estimación del riesgo, la evidencia empírica sigue demostrando que son mucho más imprecisos en la estimación del riesgo absoluto y que producen una notable y sistemática sobreestimación de la peligrosidad, especialmente, cuando se aplican a la predicción de fenómenos con bajas tasas de prevalencia como lo es la delincuencia violenta (Martínez Garay, 2016: 25).

Por lo tanto, en el sentido mencionado anteriormente, en la medida en que lo que está en juego es la libertad de un colectivo de la ciudadanía, el recurso a los mismos para la toma de decisiones penitenciarias, lejos de presentarse como un instrumento exclusivo y determinante en las mismas, debería acoger una función auxiliar, de apoyo, un instrumento más que, junto a la valoración de otros criterios, facilite la adopción de las mismas.

1. REFERENCIAS

Acale Sánchez, M., "Apuntes sobre la inconstitucionalidad de la pena de prisión permanente revisable desde la perspectiva del Derecho Penitenciario", en *Contra la cadena perpetua.* Editado por Arroyo Zapatero, Lascuraín Sánchez, Pérez Manzano, coordinado por Rodríguez Yagüe. Edit. UCLM, Cuenca, 2016, pp. 163 A 169.

Alemán Aróstegui, L. "El uso del RISCANVI en la toma de decisiones penitenciarias", en *Revista de Estudios Penales y Criminológicos*, núm. 44, 2023, pp. 1 a 43.

Brandariz García, J.A., *El modelo gerencial-actuarial de la penalidad: eficia, riesgo y sistema penal.* Edit. Dykinson, Madrid, 2016.

Brandariz García, J.A./Fernández Bessa, C., "La construcción de los inmigrantes como categoría de riesgo: fundamento, funcionalidad y consecuencias para el sistema penal español", en *Criminalización racista de los migrantes en Europa,* dirigido por Palidda y Brandariz y coordinado por Iglesias Skulj y Ramos Vázquez. Edit. Comares, Granada, 2010, pp. 271 a 289.

Castro Liñares, D., *Los instrumentos de valoración y gestión de riesgos en el modelo de penalidad español.* Edit. Reus, Madrid, 2019.

Cervelló Donderis, V., "Individualización garantista en el ejercicio de la discrecionalidad penitenciaria", en *Anuario de Derecho Penal y Ciencias Penales,* Vol. LXXII, 2019, pp. 217 a 264.

Cueto Santa Egugenia, E., "Herramientas de predicción criminal y su impacto en nuestro proceso penal", en *Revista Ius et Scientia,* vol. 11, 2025, pp. 14 a 31.

Harcourt, B.E., *Política Criminal y gestión de riesgos. Genealogía y crítica.* Edit. Ad-Hoc, Buenos Aires, 2013.

Martínez Garay, L., "La incertidumbre de los pronósticos de peligrosidad: consecuencias para la dogmática de las medidas de seguridad", en *InDret,* núm. 2, 2014, pp. 1 a 77.

Martínez Garay, L., "Errores conceptuales en la estimación de riego de reincidencia. La importancia de diferenciar sensibilidad y valor predictivo y estimaciones de riesgo absolutas y relativas", en *Revista española de investigación Criminológica,* núm. 14, 2016, pp. 1 a 27.

Martínez Garay, L., "Evidence-based sentencing y evidencia científica. A la vez, algunas consideraciones sobre las "políticas basadas en la evidencia" y el Derecho Penal", en *Teoría y Derecho. Revista de pensamiento jurídico (TEODER),* núm. 28, 2020, pp. 238 a 277.

Martínez Garay, L. /MONTES SUAY, F., "El uso de valoraciones del riesgo de violencia en Derecho Penal: algunas cautelas necesarias", en *InDret,* núm. 2, 2018, pp. 1 a 47.

Pérez Machío, A.I., "Trata de personas: la globalización del delito y su incidencia en la criminalización de la víctima inmigrante irregular a partir

de las dinámicas actuariales", en *Revista de Estudios Penales y Criminológicos*, vol. XXXVI, 2016, pp. 371 a 446.

Pérez Machío, A.I./Berasaluze Gerrikagoitia, L., "Tendencias actuales", en *Manual de Política Criminal,* dirigido por Ruíz Rodríguez y coordinado por Miró Llinares, Pérez Machío y González Agudelo. Edit. Atelier, Madrid, 2022, pp. 105 a 129.

Rueda Soriano, Y./Navarro Blasco, E., "Los sistemas actuariales de prevención y gestión de riesgos en el ámbito penitenciario", en *Juezas y Jueces para la Democracia,* núm. 94, 2019, pp. 59 a 68.

Capítulo 12

Perspectivas de una psicóloga en prisiones vascas sobre la evaluación del riesgo de reincidencia

Marian Moreno Mata

Psicóloga Clínica funcionario del cuerpo superior técnico de Instituciones Penitenciarias

Predecir el comportamiento futuro de personas condenadas por la comisión de delitos es una de las tareas que los equipos de tratamiento de los Centros Penitenciarios hacen de manera habitual para la realización de propuestas de permisos, terceros grados o libertad condicional, ya que la estimación de un riesgo medio bajo o bajo es uno de los elementos a tener en cuenta para realizar dichas propuestas en aras de la prevención de futuros delitos.

La predicción que hoy en día se realiza carece de rigor alguno, basándose en percepciones e intuiciones en la mayoría de los casos. La realidad es que hay factores psicológicos afectivos (estados de ánimo, emociones, etcétera) que, en muchos casos, no se pueden objetivar y considero determinantes en la predicción del comportamiento futuro de las personas. También son fundamentales los factores cognitivos. Se puede dar el caso, como es habitual, de falta de honestidad o la deseabilidad social del interno, en aras de conseguir un beneficio penitenciario, un permiso de salida o un régimen de semilibertad. Estas cuestiones también resultan difíciles de objetivar. En definitiva, los aspectos dinámicos de la persona condicionan especialmente su conducta futura y son los más difíciles de baremar para alcanzar un resultado fiable de riesgo de reincidencia.

Para llevar a cabo este pronóstico de reincidencia es fundamental el conocimiento del interno y, en particular, de la etiología delictiva. Resulta

imprescindible analizar si los factores que determinaron la comisión del delito se han trabajado a nivel psicoterapéutico o educativo y si dicho abordaje ha sido valorado por los profesionales de manera exitosa. Si bien, como ha sido mencionado, existe una gran complejidad y dificultad para poder objetivar.

Esta carga de realizar ese pronóstico recae, sobre todo, en la figura del psicólogo, quien, normalmente junto con el jurista, es el profesional que más volumen de internos tiene, en comparación con los educadores y los trabajadores sociales cuya *ratio* de internos es menor por profesional.

En este sentido, considero positivo el uso de herramientas porque va a potenciar el conocimiento del interno que no siempre es el adecuado, sobre todo, por la falta de profesionales y la elevada carga de trabajo. Asimismo, también puede resultar interesante para que todos los profesionales del equipo asuman también esta responsabilidad y no recaiga, como es habitual, en la figura del psicólogo.

Por otra parte, el uso de estas herramientas debe tener un valor informativo y no decisorio, siendo esta una premisa a tener muy en cuenta a la hora de formar a los profesionales en el uso de las mismas. Debe evitarse que el resultado de la herramienta sea considerado por algunos profesionales como el único aspecto a valorar o tener en cuenta, siendo importante, en muchos casos, asumir riesgos, quizás elevados, pero justificados en aras de motivar al interno. A modo de ejemplo, puede concederse un permiso teniendo en cuenta el esfuerzo y no tanto el resultado de la intervención como una herramienta de motivación para continuar con el trabajo terapéutico y reducir en un futuro el riesgo de reincidencia.

Aun y todo considero que, en algunos casos, puede condicionar la decisión, especialmente la judicial, ya que, si bien no es el caso, actualmente, del Juzgado de Vigilancia Penitenciaria de Bilbao, si constan casos, como el Juzgado de Vigilancia Central, que apoya algunas decisiones en el riesgo de quebrantamiento de la tabla de riesgo que es preceptivo remitir.

Considero que tanto los profesionales del medio penitenciario como los jueces de vigilancia están plenamente capacitados y son muy conscientes del carácter informativo de este tipo de herramientas, cuyo resultado no puede determinar por sí solo una decisión. No obstante, existe el riesgo, sobre todo en casos cuyos delitos generen mucha alarma social y el riesgo pronosticado por la herramienta sea elevado, que se determine, de manera sistemática, una decisión restrictiva basada en la herramienta.

Respecto a los delitos que generan alarma social, como son los delitos contra las personas, los delitos con victimas menores o los de agresión sexual, considero que ya de por si se recurre de manera habitual a argumentos denegatorios, bien de permiso o de régimen abierto, donde el interno no puede sino resignarse y aceptar, ya que no dependen de su evolución personal. En ellos, suele aludirse a "la alarma social del delito" o "la gravedad de los hechos", aspectos estáticos inherentes al delito, que son utilizados sistemáticamente y que no son factores que el interno pueda modificar con su conducta, su avance terapéutico o el cumplimiento de la pena.

De la misma manera que la utilización de estos factores son argumento válido para aquellos profesionales con un estilo más punitivo o restrictivo, pudiendo darse una vulneración del derecho al cambio de las personas (la reinserción), existe el peligro, como he citado anteriormente, del uso sistemático de un riesgo alto asociado a "la reincidencia", "la cuantía de la pena" o elementos relacionados, que no pueden ser abordados por el interno, lo que termina produciendo un estancamiento de su avance personal y penitenciario.

En mi experiencia como psicóloga considero que los datos estáticos no son muy relevantes (cuantía de la pena, número de ingresos, quebrantamientos, etcétera) a la hora de tomar decisiones, ya que es importante considerar a la persona y al momento vital en que se encuentra cuando se toman las decisiones, siendo muy relevante su evolución en los programas de tratamiento. Estos programas se deben hacer en los centros, pero apenas se realizan por falta de personal o por falta de motivación o formación del personal penitenciario.

Dichos programas, su realización y su valoración favorable, son el elemento clave en el riesgo de reincidencia: programas de toxicomanía, agresores sexuales, violencia de género, de conductas violentas, etcétera, considerando que se debe hacer énfasis en estos aspectos. Sin embargo, nos podemos encontrar que el interno no puede ser valorado de manera adecuada por el simple hecho de no haber programas o porque la realización de los programas depende de las entidades sociales, no haciéndose una valoración directa por parte de los profesionales que toman la decisión en los equipos de tratamiento.

Otra dificultad que habrá de abordar esta herramienta es la percepción de no ser considerada una herramienta útil y efectiva, sobre todo, por el tiempo que se debe dedicar a completar, actualizar y analizar la información para realizar la predicción.

En el día a día, el volumen de tareas, la falta de personal y la burocratización del sistema penitenciario, que supone hacer informes por y para todo, pueden tener, como consecuencia, la dedicación de poco tiempo para completar la información necesaria para una herramienta de estas características, especialmente si hay que hacerlo de una manera concienzuda y adecuada, sin caer en "las prisas", lo que llevaría a un escaso rigor y, por lo tanto, no aportar los datos necesarios para que pueda emitirse un resultado acorde a la realidad.

En relación con lo anterior, otro factor a tener en cuenta es el alcance de la predicción del riesgo de la reincidencia a nivel temporal: ¿el riesgo de reincidencia en qué espacio temporal se proyecta?, ¿en el riesgo en el momento en que se tome la decisión?, ¿en lo que dura la condena? La actualización de los datos ¿cada cuánto tiempo se debe realizar? ¿Cada vez que se modifique una variable?, ¿cada vez que vaya a salir de permiso?, ¿cada revisión de grado?, o ¿en otro momento?

En este sentido, es muy importante pensar en la evaluación del riesgo de reincidencia en el período de semilibertad o libertad condicional, siendo estas las etapas más complicadas para los internos. Si bien en régimen ordinario cada vez hay más personal técnico que puede cumplimentar los datos y tenerlos actualizados, en régimen abierto el personal es inferior y, sin embargo, es cuando resulta necesario hacer un seguimiento más exhaustivo del interno. Ello es así porque, al encontrarse el interno en semilibertad con menos capacidad de observación y menos sujeto a control formal, pueden darse modificaciones de factores determinantes del riesgo de reincidencia que pasen desapercibidas por los profesionales ante las dificultades ya mencionadas.

Como conclusión, considero que todas las herramientas que nos permitan pronosticar el comportamiento futuro de las personas son de utilidad y, por supuesto, son mejores que la mera intuición, si bien deben desarrollarse y aplicarse considerando sus propias limitaciones.

Capítulo 13

La revolución de la IA en el mundo jurídico-judicial. Algunos problemas actuales y una mirada hacia el futuro

Alberto Saiz Garitaonandia[172]

Universidad del País Vasco UPV/EHU

SUMARIO: 1. INTRODUCCIÓN. LA IA EN SU MOMENTO "MANHATTAN". 2. LA IA Y SU USO EN EL ÁMBITO JURÍDICO-JUDICIAL: DE LAS PRIMERAS EXPERIENCIAS PREDICTIVAS A LA IA GENERATIVA. 2.1. *Experiencias de uso de IA en la predicción de resoluciones judiciales*. 2.2. *El salto cualitativo de la IA generativa y su capacidad de aportar textos complejos.* 3. ALGUNAS NOTAS SOBRE EL FUNCIONAMIENTO Y LOS PROBLEMAS DEL USO DE LA IA Y SU PROYECCIÓN EN EL ÁMBITO JURÍDICO. 3.1. *La aproximación probabilística como base del funcionamiento de la IA*. 3.2. *La trascendencia y limitaciones de los datos pasados como fuente del modelo de IA*. 3.3. *Los bias o sesgos algorítmicos*. 3.4. *Los sesgos humanos ocultos y su proyección algorítmica*. 3.5. *¿Y qué hacemos con los datos sesgados?* 4. LA IRONÍA DE LA AUTOMATIZACIÓN COMO RIESGO EN EL USO DE IA. 5. A MODO DE CONCLUSIÓN: HERRAMIENTAS DEL PRESENTE Y OBSERVACIONES HACIA EL FUTURO DE UNA TECNOLOGÍA EN PLENA EFERVESCENCIA. 6. REFERENCIAS.

1. INTRODUCCIÓN. LA IA EN SU MOMENTO "MANHATTAN"

Han pasado escasos tres años desde aquel 30 de noviembre de 2022, en el que la empresa OpenAI puso a disposición del público ChatGPT, su conocido modelo de IA Generativa que supuso un antes y un después en el desarrollo de esta tecnología revolucionaria. A partir de ese momento se amplió notable-

172. Profesor Agregado de Derecho Procesal. Universidad del País Vasco UPV/EHU.

mente la percepción pública sobre la IA, que ha pasado a constituirse en tema de primera plana de actualidad. Desde ese momento las empresas tecnológicas no han parado de desarrollar innovaciones en esta área, con nuevos modelos cada vez más capaces y precisos, en una carrera que ha extendido el uso de la IA a innumerables ámbitos de nuestra vida cotidiana.

La carrera por el liderazgo en la IA ha trascendido mucho más allá del ámbito privado para colocarse en el centro de la contienda geopolítica mundial. Así es, tal y como en su día sucediera con la carrera espacial o el liderazgo atómico, en la actualidad los principales actores globales luchan por encabezar esta competición tecnológica y desarrollar los modelos más potentes, entendiendo que ello servirá de palanca para reforzar su posición de liderazgo e influencia internacional. Prueba de ello la encontramos en el alumbramiento del proyecto "Stargate", esfuerzo consorciado del Gobierno de Estados Unidos y varias empresas privadas de aquel país por movilizar 500.000 millones de dólares para el desarrollo de infraestructura de IA con el objeto de garantizar que ese país desarrolle los modelos más avanzados de esta tecnología en todo el mundo[173]. Esta iniciativa recuerda el Proyecto Manhattan, impulsado por el ejército de los Estados Unidos durante la segunda guerra mundial de cara a concebir y producir la bomba atómica, en lo que constituyó una carrera contra los científicos de la Alemania nazi alineados con el mismo fin en torno al Proyecto "Uranio". Además, también como ocurrió en aquel momento, al anuncio de "Stargate" le ha acompañado un relajamiento de las normas de seguridad y transparencia en el desarrollo de la IA en aras a acelerar el desarrollo de la misma[174], lo que está teniendo un cierto impacto en otras normativas sobre IA aprobadas en distintos lugares, entre ellos la Unión Europea[175].

La revolución anterior también está impactando en el ámbito jurídico-judicial, siempre criticado por su lentitud y sus altas tasas de penden-

173. Vid. en https://stargateprojects.net/

174. Conjuntamente con la presentación del proyecto Stargate, el Presidente Trump firmó una Orden Ejecutiva por la que, entre otras medidas, revocaba la regulación aprobada por el anterior Presidente, Joe Biden, relativa a la IA segura ("Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence") con el objetivo de eliminar barreras para conseguir el liderazgo americano en el desarrollo de la IA (https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/).

175. Véase, por ejemplo, la muy reciente propuesta de estrategia digital Omnibus, que trata de simplificar la gestión de la tramitación administrativa de cara a potenciar la labor de innovación en el ámbito tecnológico europeo: https://digital-strategy.ec.europa.eu/en/library/digital-omnibus-regulation-proposal

cia. Si bien hablamos de un sector tradicionalmente conservador[176], las dos últimas décadas han visto surgir no pocas novedades tecnológicas dignas de mención en la realidad de los tribunales de justicia, como la digitalización de expedientes y la extensión de plataformas de recepción y remisión de documentos procesales (Susskind, R., 2019). Además, la pandemia de Covid-19 trajo consigo el uso generalizado de la videoconferencia para no pocas actuaciones procesales orales (Sourdin, T., 2021)[177], elemento que en la actualidad ha quedado asumido con una cierta naturalidad y ha tomado carta de naturaleza en las propias normas procesales[178]. Con todo, la IA tiene un potencial todavía mayor, pues su uso podría suponer no ya una ayuda o apoyo en ciertas tramitaciones —que también—, sino, en un escenario extremo, la toma de decisiones totalmente automatizada por parte del modelo preparado para tal fin, también en entornos tan sensibles como el jurisdiccional, lo que pondría en cuestión principios históricamente aceptados y constitucionalmente asumidos desde la génesis del Estado moderno y la consolidación de la división de poderes.

2. LA IA Y SU USO EN EL ÁMBITO JURÍDICO-JUDICIAL: DE LAS PRIMERAS EXPERIENCIAS PREDICTIVAS A LA IA GENERATIVA

2.1. EXPERIENCIAS DE USO DE IA EN LA PREDICCIÓN DE RESOLUCIONES JUDICIALES

Ya antes de la eclosión de la IA Generativa la ciencia exploró las aptitudes de los modelos previos de cara a poder predecir con acierto el sentido de las resoluciones judiciales. En el sentido anterior, no podemos encontrar una referencia clara que, a modo de "benchmark" o referencia de comparación, pudiera dar por bueno un determinado umbral de acierto en el uso de esta tecnología. Asimismo, tampoco podemos obviar que no todas las resoluciones judiciales son acertadas: en algunos casos estas son revocadas y su

176. Algunos, como Susskind (2019), lo achacan a un "status quo bias" entre los profesionales jurídicos, que hace que este colectivo suela ser reacio al cambio.

177. Sobre las innovaciones acaecidas en los últimos tiempos en el mundo judicial puede leerse la interesante aportación de Sourdin (2021).

178. En este sentido, el Real Decreto-ley 6/2023, de 19 de diciembre, por el que se aprueban medidas urgentes para la ejecución del Plan de Recuperación, Transformación y Resiliencia en materia de servicio público de justicia, función pública, régimen local y mecenazgo, introdujo distintas modificaciones en las Leyes de Enjuiciamiento Civil y Criminal que naturalizan la realización de distintas actuaciones y comparecencias telemáticas con el objetivo de lograr una mayor eficiencia y accesibilidad.

fallo modificado por medio del sistema de recursos establecido, mientras que no podemos afirmar la suerte que habrían corrido aquellas que adquirieron firmeza sin haber sido impugnadas. Este elemento de la fiabilidad de las resoluciones judiciales no encuentra fácil respuesta y, como veremos, se proyecta como una idea latente en relación con el uso de la tecnología en la toma de decisiones en el ámbito de la Administración de Justicia, ya sea de forma coadyuvante al humano o en su hipotética implementación totalmente autónoma.

Como ya expusimos en su momento, un análisis doctrinal nos permite encontrar muy diversas experiencias predictivas pre-IA Generativa realizadas, a su vez, en muy distintos lugares[179] (Saiz Garitaonandia, 2023). En todo caso, de todas ellas citaremos aquí las tres que, en nuestra opinión, tuvieron una mayor repercusión o eco científico: la de Aletras et al. (2016) en relación con las sentencias del Tribunal Europeo de Derechos Humanos, la de Katz et al. (2017) sobre predictibilidad de sentencias del Tribunal Supremo de los Estados Unidos de América y, finalmente, la experiencia de Sulea et al. (2017), que tomó como referencia para su análisis sentencias del Tribunal Supremo francés.

En el caso de las tres experiencias comentadas los modelos predictivos consiguieron unos resultados notables. Así, la experiencia de Aletras (2016) logró llegar a una predictibilidad media del 79% en relación con el fallo de las sentencias del Tribunal Europeo de Derechos Humanos. En el caso de Katz (2017), su modelo de IA consiguió un 70% de acierto a la hora de establecer el fallo de las sentencias del Tribunal Supremo de Estados Unidos. Por último, el mejor resultado fue el logrado por Sulea y su equipo en 2017, que consiguieron un extraordinario 97% de acierto a la hora de predecir el resultado del fallo de las sentencias del Tribunal Supremo de Francia. Estos datos daban muestra de las potencialidades que tenía la aplicación predictiva de la IA al ámbito de las decisiones judiciales, todavía más en un entorno caracterizado por una evolución fulgurante de esta tecnología que, con la incorporación de los LLM —Large Language Models— y la IA Generativa,

179. "Sin ánimo de ser exhaustivos, un primer análisis a algunas investigaciones publicadas nos permite contemplar estudios predictivos de diferente índole realizados sobre resoluciones judiciales en países tan diversos como, por ejemplo, China –donde se están desarrollando grandes esfuerzos en la materia-, Brasil, Estados Unidos, Singapur, Tailandia, Filipinas, Alemania o las decisiones del Tribunal Europeo de Derechos Humanos, que han centrado no pocos estudios en la materia" (Saiz Garitaonandia, 2023). Las referencias a las investigaciones en concreto pueden encontrarse en el artículo en cuestión.

ahora no solo se limita a predecir el fallo sino a generar un texto articulado que pueda servir de resolución.

2.2. EL SALTO CUALITATIVO DE LA IA GENERATIVA Y SU CAPACIDAD DE APORTAR TEXTOS COMPLEJOS

No obstante, tal y como mencionábamos inicialmente, la verdadera revolución de la IA en relación con su penetración social y profesional masiva vino constituida por la puesta a disposición pública del modelo ChatGPT a finales de noviembre de 2022. Desde ese momento los modelos de IA, tanto el de la matriz OpenAI como el de otras empresas tecnológicas, no han parado de evolucionar y presentar mejoras, en lo que se ha constituido como una verdadera carrera por liderar este segmento. El motivo de este fulgurante desarrollo y éxito podemos encontrarlo en su casi universal "usabilidad" y la utilidad en sus "outputs" o productos en distintos contextos: mientras que la IA tradicional analizaba y clasificaba información de una forma más o menos compleja, incluso aprendiendo en su propio proceso —Machine Learning—, la IA Generativa va más allá, creando y produciendo contenido original que, en forma de texto coherente y de muy buena calidad, responde a las cuestiones que le sean planteadas. Este paso de la mera clasificación a la generación o producción novedosa y de calidad hace que la utilización de esta tecnología se haya extendido exponencialmente en poco tiempo, especialmente en aquellas tareas intelectuales en las que el análisis de datos y la producción posterior de algún tipo de escrito o trabajo cobra especial relevancia (Saiz Garitaonandia, 2024).

Ya desde los inicios no faltaron expertos que pusieron a prueba la fiabilidad de estos nuevos modelos, tratando de probar su pericia ante distintos retos intelectuales. En este sentido, fueron pioneros los trabajos de Terwiesch (2023) y, por lo que al ámbito legal respecta, las importantes aportaciones de Katz y Bommaritto (2022), ambos realizados a los pocos días del lanzamiento público de ChatGPT. El primero sometió al modelo de OpenAI a cinco problemas que constituirían un examen de su materia en el MBA (Master in Business Administration) de la prestigiosa escuela de negocios de Wharton en la Universidad de Pennsylvania, consiguiendo unos resultados generales muy positivos, aunque resaltando que, en algunos extremos, las respuestas aportadas cometían imprecisiones propias de un nivel escolar. En el caso de Katz y Bommaritto, su experiencia puso a prueba a ChatGPT en dos exámenes de larga tradición y reconocida dureza en los Estados Unidos: el BAR —examen de acceso a la abogacía— y el CPA —Certified

Public Accountant—. En el caso del BAR (Katz y Bommaritto, 2022), la IA consiguió unos resultados muy positivos, pero que todavía no se acercaban a la media de los obtenidos por los candidatos humanos. Por lo que respecta al CPA (Bommarito et al., 2023), la media de acierto superó el umbral del 50%, lo que, en términos comparativos, lo situaba como un candidato competitivo, toda vez que este examen muestra unos números elevados de suspendidos entre los humanos que aspiran a certificarse. En síntesis, nos encontrábamos ante una tecnología que, en general, entendía el contexto y aportaba respuestas, tanto simples como complejas, lo que la hacía apta para poder colaborar en tareas iniciales de áreas técnicas en las que, eso sí, el conocimiento y razonamiento humano seguían siendo claramente necesarios y preminentes, toda vez que su precisión, si bien notable, no llegaba a un umbral en el que se pudiera depositar en ella una confianza total de cara a la confección de las tareas pretendidas.

La sorpresa llegó de la mano de GPT-4, la versión evolucionada del modelo anterior, que solo escasamente 4 meses después del lanzamiento del anterior ChatGPT fue dada a conocer públicamente[180]. El nuevo modelo fue sometido a numerosos *test* y exámenes, consiguiendo resultados muy positivos y sensiblemente superiores a la versión anterior. Por lo que respecta al ámbito legal, la referencia volvió a ser el examen BAR (Katz, et al., 2023), en el que deslumbró con un 75% de acierto en los *test* —superior a la media humana del 68%— y resultados muy satisfactorios en el apartado de elaboración de escritos en respuesta a casos prácticos[181]. La evolución fue tal y tan rápida —cuestión de meses— que sorprendió todavía más que el modelo anterior, y contribuyó de una manera extraordinaria a popularizar el uso de la IA Generativa en múltiples ámbitos de la vida, entre ellos el jurídico.

Pero con la extensión de su uso comenzaron, a su vez, a conocerse distintos problemas que este llevaba aparejado: su uso indiscriminado y sin control expuso con crudeza cómo ChatGPT, en ocasiones, inventaba citas y jurisprudencia en textos que, por lo demás, gozaban de una buena estructuración y fundamentación desde un punto de vista formal[182]. Los hechos

180. El modelo GPT-4 se presentó de forma oficial el 14 de marzo de 2023, en este caso para los usuarios de las cuentas de pago de ChatGPT.

181. MEE (Multistate Essay Examination).

182. La primera referencia sobre una de estas cuestiones la encontramos en mayo de 2023, al presentar un letrado un documento producido por ChatGPT ante una corte de Nueva York. Dicho documento, bien citaba sentencias de algunos tribunales de Estados Unidos que posteriormente se demostró que no existían, lo que conllevó una sanción para el letrado en cuestión. Sobre la cuestión puede leerse https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html

anteriores constituyen una proyección del fenómeno que ha venido en llamarse "alucinaciones" (Saiz Garitaonandia, 2025) que, en ocasiones, pueden generar los modelos de IA: una tendencia de los Grandes Modelos de Lenguaje que utiliza la IA Generativa de producir contenidos que, en algunos casos, no es coincidente con datos reales del pasado, aunque sí muestra una apariencia de verosimilitud igual que el resto del texto proporcionado[183]. En este sentido, en ámbitos como el del asesoramiento o defensa jurídica, en los que la precisión y veracidad de la información aportada se constituye como un elemento fundamental, la existencia de estas alucinaciones hace que la prudencia a la hora de utilizar estos sistemas deba extremarse, haciéndose necesaria una supervisión humana experta de cara a la elaboración final del documento. De hecho, lejos de constituirse en sucesos anecdóticos que ocurren con poca frecuencia, determinados estudios apuntan a una muy alta presencia de alucinaciones en los outputs de la IA Generativa, cuyos modelos son en ocasiones incapaces de discernir si la información proporcionada es acorde con la realidad o ha sido inventada en su proceso de generación tecnológico[184].

Los sucesos descritos han impulsado en no pocos lugares la implementación de regulación o guías de uso de la IA por parte de los profesionales jurídicos a la hora de realizar escritos procesales[185], una pauta que se ha ido generalizando a medida que, progresivamente, iban surgiendo y se iban conociendo casos de uso de IA Generativa con una supervisión humana insuficiente[186] o, directamente, inexistente. Además, los dos últimos años

183. Para un seguimiento periódico de las numerosas alucinaciones que van surgiendo en el panorama legal comparado –fundamentalmente en el mundo anglosajón-, puede visitarse la web https://naturalandartificiallaw.com/

184. Al ser preguntados por cuestiones específicas relativas a casos de tribunales federales en Estados Unidos, ChatGPT proporcionó algún tipo de información alucinatoria en un 69% de los casos, mientras que la cifra fue superior para otros modelos, llegando hasta el 88% en modelos como Llama 2 de la compañía Meta (https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive).

185. Por citar algunos de los últimos ejemplos, UNESCO, "Guidelines for the use of AI systems in Courts and Tribunals", Paris, 2025; e Ilustre colegio de la abogacía de Madrid, "Guía ICAM de buenas prácticas para el uso de la Inteligencia Artificial (IA) en la Abogacía", Madrid, 2025.

186. De hecho, la propia OpenAI ha procedido, con fecha de 29 octubre de 2025, a modificar la política de uso de sus modelos GPT para prohibir su uso de cara a realizar asesoramientos personalizados en aquellas profesiones que, como es el caso de la abogacía o la medicina, requieren una licencia profesional específica, "sin la adecuada participación de un profesional colegiado": "You cannot use our services for: … provision of tailored advice that requires a license, such as legal or medical advice,

nos han mostrado la irrupción de diferentes modelos y sus evoluciones que, si bien van consiguiendo mejoras graduales con respecto a sus predecesores, proyectan un progreso más ralentizado después de los impresionantes saltos producidos en las primeras generaciones de los Grandes Modelos de Lenguaje, en la línea descrita en los párrafos precedentes. Este momento está siendo aprovechado por muchas empresas y organizaciones para adoptar soluciones de IA adaptadas a sus necesidades concretas. En este sentido, el mundo jurídico-judicial no es una excepción, pues en él están surgiendo no pocas iniciativas de desarrollo e implementación de IA que tratan de aportar soluciones a necesidades concretas por medio del uso de modelos adaptados o diseñados específicamente para determinadas tareas.

3. ALGUNAS NOTAS SOBRE EL FUNCIONAMIENTO Y LOS PROBLEMAS DEL USO DE LA IA Y SU PROYECCIÓN EN EL ÁMBITO JURÍDICO

En los apartados precedentes hemos expuesto la acelerada forma en la que, en un muy breve espacio de tiempo, la IA se ha instalado en nuestras vidas, con una especial referencia al ámbito jurídico, objeto de nuestro estudio. De hecho, la velocidad con la que ha irrumpido explica, en nuestra opinión, que buena parte de los usuarios no tengan unos conocimientos mínimos de la forma en la que esta funciona y, de esta manera, se hayan lanzado a utilizarla sin reparar en los posibles inconvenientes o problemas que, como las mencionadas alucinaciones, su uso podía llevar aparejado. Trataremos a continuación de acercarnos a las cuestiones que, en nuestra opinión, pueden aportar más luz de cara a conocer los potenciales riesgos que conlleva el uso de esta tecnología.

3.1. LA APROXIMACIÓN PROBABILÍSTICA COMO BASE DEL FUNCIONAMIENTO DE LA IA

El primer elemento a tener en cuenta en relación con la IA y, especialmente, con los grandes modelos de lenguaje en los que se sustentan sus modernas versiones generativas, viene constituido por su funcionamiento matemático-estadístico. Efectivamente, si bien en un primer momento puede llamar la atención que mencionemos la matemática como fundamento esencial de una tecnología generadora de textos —nos centramos en estos *outputs*—, la realidad es que los algoritmos fundamentan su funcionamiento en la bús-

---

without appropriate involvement by a licensed profesional" (https://openai.com/policies/usage-policies/).

queda de patrones, de pautas que se repiten en las inmensas bases de datos con las que han sido entrenados (Coeckelberg, M., 2020). La obtención de esos patrones se realiza en base a un cálculo de probabilidades totalmente matemático, según el cual el modelo de IA busca relaciones entre los datos para proyectar su *output*. Por pura esencia, esa aproximación probabilística se configura con una naturaleza totalmente matemático-cuantitativa, lo que nos permite entender, a su vez, que la IA proporcionará resultados óptimos en escenarios en los cuales la determinación cuantitativa cobre especial relevancia (Zerilli, J., 2021)[187]. Por el contrario, cuando el ámbito de referencia esté marcado por una presencia más marcada de elementos cualitativos de difícil o imposible cuantificación —como cuestiones de matiz o detalle—, la experiencia demuestra que la IA no va a proyectar resultados tan satisfactorios (Wu, T., 2023). Ello es una muestra de la dificultad de los modelos de IA en captar el factor de relevancia de la información si ésta no encuentra reflejo, a su vez, en datos cuantificables que puedan ser reconocidos en los patrones que marcan su funcionamiento.

Trasladado al mundo jurídico-judicial, la idea anterior nos proyecta la aplicabilidad de esta tecnología en ámbitos o circunstancias en los que los elementos cualitativos no detectables no tengan un papel trascendente o, *a sensu contrario*, aquellos escenarios en los que las circunstancias cuantificables ostenten una presencia mayoritaria y preponderante sobre los posibles elementos no cuantificables. Además, la comprensión del funcionamiento matemático del algoritmo nos es especialmente útil para entender la diferente aproximación que realizamos máquinas y humanos a la hora de elaborar nuestras conclusiones. De hecho, desde el punto de vista del enjuiciamiento, el razonamiento judicial, en principio, lleva a sus conclusiones en un proceso de determinados de hechos por medio de prueba directa o presunciones, para luego pasarlos por el tamiz jurídico-normativo y, con ello, llegar a una conclusión en forma de fallo. El juez no emite su sentencia en base a un porcentaje de probabilidades, sino en un convencimiento en relación con la realización de unos hechos concretos. Esta realidad es especialmente relevante en el ámbito penal, donde el tribunal, de cara a enervar la presunción de inocencia, debe llegar a un convencimiento pleno sobre la culpabilidad del acusado, sin que pueda existir duda razonable en relación con los hechos en cuestión:

187. Como muy bien explica Zerilli, la existencia de una relación en forma de patrón no implica necesariamente una relación de causalidad: "their whole modus operandi is to look for correlations in the data. Correlations are a legitimate source of knowledge, to be sure, but as most of us learn in high school: correlation is not causation".

> "El motivo nos sitúa en el epicentro de un problema complejo sobre el que gira, en muy buena medida, el propio núcleo garantizador del principio de presunción de inocencia como regla de juicio: cuándo puede considerarse satisfecho el estándar más allá de toda duda razonable para fundar una sentencia condenatoria[188]".

Esta visión clásica de la presunción de inocencia no puede ser pasada por alto y condiciona la aplicabilidad de la IA en el ámbito judicial en sus usos más extremos —esto es, la automatización total en la toma de decisiones—, al menos en lo que a las áreas del derecho más sensibles se refiere. Obviamente, en la medida en la que hablemos de proyecciones menos decisorias y de modelos más precisos y menos "alucinatorios", las posibilidades de uso se irán expandiendo, tanto en su vertiente autónoma como en opciones colaborativas entre el personal humano-jurisdiccional/jurídico que puede verse ayudado en sus funciones por las facilidades que puede aportar esta solución tecnológica. En este sentido, considerando los progresos que se están constantemente produciendo en esta rama de la ciencia, creemos que deberemos estar muy atentos, pues lo que hoy parece impensable puede convertirse en realidad en un futuro no tan lejano.

3.2. LA TRASCENDENCIA Y LIMITACIONES DE LOS DATOS PASADOS COMO FUENTE DEL MODELO DE IA

En las líneas precedentes mencionábamos la forma matemática en la que "razona" la IA, buscando patrones entre las bases de datos con las que ha sido entrenada o a las que tiene acceso para, así, poder emitir sus respuestas. Esas operaciones, que han alcanzado una complejidad enorme de cara a poder contextualizar el lenguaje y obtener los *outputs* correspondientes, siguen en cualquier caso encontrando su base existencial en los datos con los que se ha entrenado o alimentado el modelo de IA. En este sentido, no podemos obviar que esa información que la IA toma de referencia en forma de datos proviene del pasado, esto es, se alimenta de información pretérita que proyecta hacia el futuro en forma de *outputs* que, como hemos visto, encarnan predicciones probabilísticas de un resultado, para cuya elaboración se han tenido en cuentas los mencionados datos pasados (Saiz Garitaonandia, 2023).

La cuestión anterior no es baladí, y va a constituir una nueva limitación en relación con el uso de la IA: si partimos de la base de un funcionamiento de la IA anclado y determinado por información y decisiones del pasado,

188. STS 680/2022, Penal, de 17 de febrero, F.J. 11.

parece evidente que sus respuestas van a ser un reflejo de lo que sucedió en dicho espacio temporal —una *cristalización del pasado hacia el futuro*— y, por lo tanto, la capacidad de asumir nuevas tendencias, corrientes o cambios sociales, o su aplicación en escenarios *ex novo*, va a quedar totalmente condicionada a la readaptación del modelo a esa nueva realidad por medio de nuevos datos que, de forma suficiente, puedan ayudar a reflejar dichas novedades en sus productos o resultados. Pensamos, por ejemplo, en novedades normativas que deban de ser interpretadas, para las que no exista un *corpus* de información elaborada que vaya más allá del propio texto de la norma, lo que en cualquier caso exigiría la presencia exegética y valorativa de humanos que puedan aportar las decisiones con las que, en posteriores estadios y ya con los datos suficientes, el ingenio tecnológico pueda desarrollar su labor (Sourdin, T., 2021).

3.3. LOS *BIAS* O SESGOS ALGORÍTMICOS

En general, por sesgo debemos entender la tendencia o parcialidad que, desviándose de la neutralidad, puede tomar toda información, análisis o decisión. Desde una perspectiva estadística el sesgo vendría constituido por una selección o favorecimiento de ciertos datos sobre otros dentro de un muestreo estadístico, lo que daría como resultado una falta de representatividad real del mismo sobre la realidad analizada.

En todo caso, todo sesgo discriminatorio no tiene por qué tener una connotación negativa: en determinadas ocasiones podemos encontrarnos con sesgos voluntariamente asumidos, introducidos de forma voluntaria para causar un efecto concreto, como el favorecimiento de un grupo sobre otro. Este sería el caso, por ejemplo, de las discriminaciones positivas, comúnmente aceptadas por la jurisprudencia a pesar de no seguir criterios estrictamente igualitarios. No obstante, por lo que respecta a la IA, en general, cuando se mencionan los sesgos o *bias* se desea trasladar la idea de una tendencia involuntaria o no directamente conocida, que tiene como consecuencia una discriminación contra ciertas personas o colectivos, que por culpa de dicho sesgo sufre un perjuicio por todas o algunas de las decisiones que toma el modelo de IA (Friedman, B. y Nissembaum, H.,2016).

La emisión de los *outputs* por parte de los algoritmos modernos se realiza tras un proceso en el que la información introducida —input— pasa a través de unas redes neuronales que, tras analizar sus inmensas bases de datos y realizar las complejas operaciones propias del modelo en cuestión en

busca de patrones[189], emite sus productos u *outputs*. En este sentido, parece evidente que la automatización algorítmica va a plasmar en sus resultados los posibles sesgos de que adolezca la información con la que haya sido entrenado el modelo, en un juego de conservación y perpetuación de los mismos. Sobre este particular es muy ilustrativa la dicción coloquial inglesa GIGO —"garbage in, garbage out"—, con la que quiere expresarse que si un modelo es entrenado con datos de baja calidad y con sesgos de origen, no debe de sorprender que los resultados obtenidos se encuentren, a su vez, sesgados y sean de una calidad cuestionable (Yeung, W. N., 2025). Así, el proceso de selección, cribado, etiquetado y calificación de los datos de cara al entrenamiento del modelo cobra una trascendencia esencial para obtener unos resultados finales de calidad y libres de sesgos. En todo caso, el origen de los sesgos puede encontrarse también relacionado con otros factores, como el escaso número de datos con el que se haya entrenado el modelo, la falta de un equilibrio en el origen de dichos datos o la propia configuración tecnológica del algoritmo. No obstante, el fundamento habitual y más común de los *bias* viene constituido por la existencia de sesgos en la información de entrenamiento y referencia del algoritmo, en el sentido mencionado en las líneas precedentes.

3.4. LOS SESGOS HUMANOS OCULTOS Y SU PROYECCIÓN ALGORÍTMICA

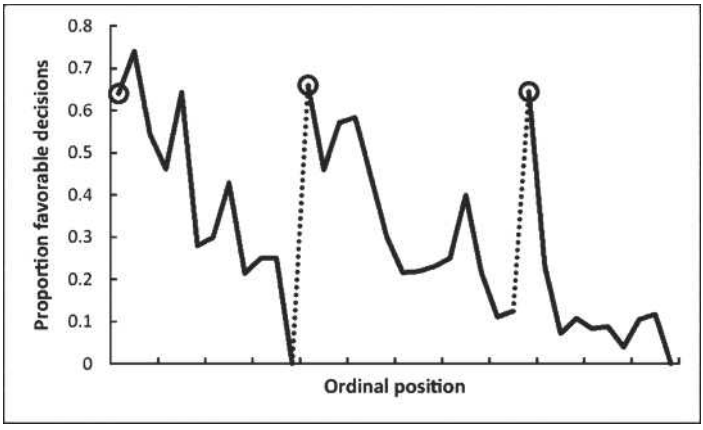
Ciertamente, como decíamos, la presencia de información tendenciosa o viciada en los datos de entrenamiento y acceso del algoritmo se presenta como el motivo principal que causa que los outputs producidos por la IA se encuentren, a su vez, sesgados.

Una primera aproximación al problema podría conducirnos a insistir en la depuración de los datos previos como medida preventiva para evitar el GIGO antes mencionado y, con ello, mejorar la calidad de los productos de la IA. Con todo, no podemos pasar por alto la dificultad que puede entrañar la detección de los sesgos, pues muchos de ellos pueden pasar inadvertidos incluso para los propios afectados (Plous, S., 2003). Trataremos de ilustrar la problemática anterior por medio de un ejemplo que nos ayude a percibir la realidad de los sesgos ocultos en la información que utilizaríamos para

189. Obviamos aquí, de forma deliberada, la problemática relativa a la falta de trazabilidad y transparencia de los modelos de IA más comunes, que no nos permiten conocer los motivos o razones concretas por las que dichos modelos han llegado a las conclusiones u *outputs* que emiten, en un fenómeno comúnmente conocido como de caja negra o *black box*. Sobre ello puede leerse a Saiz Garitaonandia, (2023).

programar los algoritmos, incluso cuando hablamos de datos de naturaleza jurisdiccional.

Danziger, Levav y Annain-Pesso (2011) analizaron más de 1000 resoluciones relativas a libertad condicional decididas por ocho jueces que dirigían *parole boards* en Israel. El estudio de estos académicos israelíes ordenó los casos resueltos tomando en consideración un parámetro que, *a priori*, no debería influir a la hora de decantar el sentido de una decisión, mucho menos si se trata de una cuestión como la libertad de una persona: las horas en las que las *parole boards* eran realizadas. El estudio muestra que cuando los jueces estaban más frescos —al inicio de la jornada y tras las dos pausas para comer y descansar que se hacían a lo largo del día—, el porcentaje de personas a las que se les concedía la libertad condicional era de entre un 65-75 %, mientras que, a medida que iba avanzando la sesión sin descansos, dicho porcentaje se iba reduciendo paulatinamente de una forma evidente hasta llegar, incluso, a desestimar todas las peticiones. Del estudio se podía inferir, por lo tanto, que las decisiones de los jueces de las citadas *parole boards* estaban sesgadas por la hora en la que cada una de ellas se tomaba, existiendo un patrón claramente identificable que señalaba un mayor éxito de las personas sometidas al criterio de esos jueces si su caso era valorado al inicio de la jornada o tras las pausas de descanso que si lo era en un momento posterior.



Porcentaje de decisiones favorables de las *parole boards* (Danziger, S., et al., 2011).

El estudio anterior nos muestra cómo, en ocasiones, las personas podemos estar influenciadas por factores ajenos a los datos que se han presentado como objeto de análisis a la hora de tomar nuestras decisiones, incluso en aquéllas con unas muy serias consecuencias tomadas por personas expertas

y con un alto grado de preparación[190]. Esta realidad no puede ser perdida de vista, pues serán datos como los anteriores los que servirán de base para que, como *dataset*, los modelos de IA puedan ser entrenados y emitan los consiguientes *outputs*.

Considerando que en no pocas situaciones podemos encontrarnos con sesgos humanos ocultos en decisiones pasadas, un buen punto de partida podría pasar por utilizar la propia IA para detectar esos dichos sesgos que, en forma de patrones, quedarían patentes con el uso de esta tecnología. Tras este ejercicio podríamos analizar los patrones encontrados, tratar de identificar el factor que perturba la objetividad en la toma de decisiones y, en su caso, tomar algún tipo de medida preventiva de cara a eliminarlo o reducir su influencia (Surden, H., 2014). En el sentido anterior, aplicado al supuesto de los jueces del *parole board*, se podría reflexionar sobre una reorganización de la jornada de trabajo o de la toma final de decisiones que tuviera en cuenta nuestro sesgo negativo a medida que va avanzando la jornada —o, tal vez, la propensión excesivamente positiva a primeras horas del día o tras las pausas—. De esa manera construiríamos un mejor escenario de cara a que la toma de decisiones sea un reflejo fiel de los elementos objetivos analizados, reduciendo las posibles penalizaciones que el paso de las horas deparaba en los resultados.

3.5. ¿Y QUÉ HACEMOS CON LOS DATOS SESGADOS?

Centrándonos de nuevo en el ámbito jurídico-judicial, la posible existencia de información que, de origen, se encuentre sesgada y se proyecte, posteriormente, en las decisiones de nuestro modelo de IA nos conmina a hacer una reflexión en relación con dichos datos viciados y, en caso de detectarlos, la forma de obrar ante los mismos. Es este un asunto de gran trascendencia, que puede parecer de fácil respuesta inclinándonos a eliminar los datos sesgados pero que, como trataremos de mostrar a continuación, entraña una problemática compleja y pluridimensional que nos obliga a una reflexión más sosegada y profunda.

Ciertamente, los datos que alimentan un modelo de IA jurídico provienen, entre otros pero especialmente, de la jurisprudencia pasada emitida por jueces y magistrados que, con sus circunstancias y entornos, al tomar dichas decisiones han proyectado en las mismas el mundo en el que vivían y la forma en la que este era percibido por ellos. En este sentido, los sesgos

190. Sobre otros elementos ajenos que pueden impactar en los jueces en su proceso de toma de decisiones puede leerse Sourdin, T., (2021).

que estas decisiones puedan contener —en ocasiones, como hemos visto, de forma inadvertida— son un reflejo de dicho mundo, por lo que su eliminación implicaría, a su vez, una cierta mutación del pasado en el que estas fueron tomadas. Por supuesto, algunas de esas mutaciones reflejarán evoluciones sociales en temas sensibles —progresos en temas de igualdad, por ejemplo—, pero no todos los sesgos que encontremos tienen que ostentar etiquetas tan claras, siendo en ocasiones complejo discernir si su supresión se configura como un elemento positivo o no[191]. Nos hallamos, sin duda, ante una cuestión tremendamente compleja, que trasciende muy ampliamente el plano jurídico o tecnológico para adentrarse en dilemas éticos, sociológicos o antropológicos que superan con mucho el alcance de este trabajo y que apuntan a la necesidad de contar con equipos pluridisciplinares con alta capacidad y visión que puedan tomar las precitadas decisiones de exclusión de determinados datos del sistema con plena consciencia de las consecuencias multifacéticas que dicha exclusión puede comportar.

4. LA IRONÍA DE LA AUTOMATIZACIÓN COMO RIESGO EN EL USO DE IA

Pensemos, en cualquier caso, en un modelo de IA que, con un muy alto nivel de acierto y desempeño, es puesto en uso en el ámbito jurisdiccional. Ciertamente, incluso a pesar de que la IA Generativa sea capaz de emitir unos escritos correctamente motivados, en la línea con la jurisprudencia emitida en el pasado por otros órganos judiciales, no podemos perder de vista que su aproximación a la resolución del asunto —a la emisión de su *output*— es meramente matemática, en base a predicciones que tienen en cuenta elementos exclusivamente estadísticos recabados entre la información pasada que sirvió de entrenamiento para el modelo. Destacábamos, en esta línea, que la IA afronta serias dificultades para detectar elementos cualitativos, esto es, para identificar factores de relevancia que, en forma de matices o detalles, podrían aparecer en la información del caso. En nuestra opinión, estos elementos desaconsejarían un hipotético sistema que pudiera juzgar automatizadamente una generalidad de asuntos sin discriminación

191. Sobre esta cuestión expone muy acertadamente Završnik (2019): "The point is that de-biasing entails that inherently political decisions are to be made, for example as to what is merely gendered and what is sexist language that needs to be 'cleaned', or what is hate speech targeting minorities and which differential treatment should be deemed to be discriminatory. In a machine-based utopia, such decisions would thereby be relegated to the experts of the computer science elite. In this sense, de-biasing is not even desirable".

y, en caso de establecerse para alguna tipología más acotada y estandarizada, sería en todo caso necesario ofrecer a las partes algún tipo de medio impugnatorio ante órganos jurisdiccionales humanos en el que se pudieran destacar las posibles peculiaridades cualitativas del asunto en cuestión.

En el contexto actual, en todo caso, las opciones que están ganando presencia se basan en aproximaciones menos radicales, con aplicaciones asistenciales, colaborativas o de asesoramiento al personal humano en los tribunales de justicia. Dichas aplicaciones abarcan un abanico muy amplio de posibilidades, y pueden ir desde la mera ayuda en la tramitación procesal hasta la confección de borradores de escritos que el personal judicial puede tener en cuenta de cara a la confección final de sus resoluciones.

Podría parecer que la idea anterior vence cualquier inconveniente que pudiera sugerirse en relación con el uso de la IA en el ámbito judicial. Con todo, no podemos perder de vista la propensión psicológica humana a "sobreconfiar" en las innovaciones tecnológicas una vez estas alcanzan un umbral de acierto. Esta sería una de las llamadas "ironías de la automatización" (Bainbridge, L., 1983) que, como efecto pernicioso en la implementación de las nuevas tecnologías, afectaría, en nuestro caso, al personal que, judicial o no, asumiría y daría por buenas las recomendaciones emitidas por el modelo de IA sin realizar un control real o efectivo sobre el contenido del mismo. Así, cobra una especial relevancia la búsqueda de un necesario equilibrio entre, por un lado, la implementación de sistemas de IA confiables que ayuden en las distintas tareas a realizar por parte del personal al servicio de la Administración de Justicia y, por el otro, la existencia de un control efectivo (Zerilli, J., et al., 2019)-no meramente formal— de los outputs emitidos por el modelo de IA pues, como ha quedado dicho, la asunción directa de los productos emitidos por el modelo puede llevar consigo no realizar ningún tipo de cribado sobre un escrito en el que, como antes se mencionaba, se hayan podido plasmar distintas alucinaciones del modelo.

5. A MODO DE CONCLUSIÓN: HERRAMIENTAS DEL PRESENTE Y OBSERVACIONES HACIA EL FUTURO DE UNA TECNOLOGÍA EN PLENA EFERVESCENCIA

Hasta aquí hemos tratado de describir, de una forma sintética, el gran vuelco que ha supuesto la irrupción de la IA en los tres últimos años y los riesgos más destacables que su uso puede llevar aparejado. En la actualidad podemos encontrarnos con una multitud de herramientas de IA que, en el ámbito jurídico, proporcionan diferentes soluciones para distintas tareas que

les sean solicitadas. Algunas han sido desarrolladas por las grandes firmas consolidadas que, como Thomson Reuters Westlaw[192] o LexisNexis[193], no han querido quedarse fuera de este segmento en auge y se han esforzado por implementar soluciones asistidas con IA para sus servicios tradicionales. Además, no faltan ejemplos de *start-ups* jurídicas que, como Harvey[194] o, dentro de nuestras fronteras, maite.ai[195], ofrecen sus servicios y se abren paso de forma disruptiva entre las grandes empresas del sector. Tampoco las Administraciones Públicas están siendo ajenas a las posibilidades que brinda la IA de cara a mejorar el servicio público y, de hecho, el propio CGPJ[196] está poniendo a disposición de los miembros de la carrera judicial diferentes soluciones basadas en IA que facilitan resúmenes de documentos o permiten la interacción conversacional para obtener información más precisa sobre el contenido de los mismos, además de otras que ya se encontraban disponibles y que permitían, por ejemplo, la anonimización de las resoluciones. Todo ello en un escenario en el que no dejan de surgir propuestas de uso de IA en diferentes escenarios y con diferente alcance[197], muestra del panorama dinámico y en constante cambio en el que nos encontramos.

Ciertamente, la Justicia se configura como un elemento superior dentro de nuestro ordenamiento jurídico, un valor esencial que, en sus decisiones, proyecta la conjunción de un buen número de elementos como sensibilidad social, vulnerabilidad económica y otros muchos conceptos jurídicos indeterminados, todos ellos con unas componentes cualitativas de difícil percepción desde la mera aproximación algorítmico-matemática (Calaza López, S., 2025). No obstante, las posibilidades que la IA brinda pueden ayudar a agilizar el servicio público que brinda la Administración de Justicia, aquejado de una perenne lentitud y visto por no pocos ciudadanos como algo distante y costoso.

Tras solo tres años desde el surgimiento de la IA Generativa, creemos que nos ajustamos a la realidad si afirmamos que la implementación de esta

192. Vid. https://legal.thomsonreuters.com/en/products/westlaw-edge
193. Vid. https://lexislegal.ai/?utm_campaign=mcsearch&gad_source=1&gad_campaignid=21953434238&gclid=EAIaIQobChMI6efkv87RkQMVH0hBAh3RfgQwEAAYASAAEgI7w_D_BwE
194. Cfr. https://www.harvey.ai/
195. Vid. https://www.maite.ai/
196. Véase https://www.poderjudicial.es/cgpj/es/Poder-Judicial/En-Portada/El-CGPJ-pone-a-disposicion-de-la-Carrera-Judicial-nuevas-herramientas-de-Inteligencia-Artificial
197. Buena prueba de ello es, por ejemplo, la propuesta del Magistrado Javier Ercilla en relación con la admisión a trámite de los recursos de casación para la unificación de doctrina ante la Sala Cuarta del Tribunal Supremo (Ercilla García, 2025).

tecnología en el ámbito jurídico y, más específicamente, jurisdiccional no ha hecho más que empezar. La capacidad transformativa que acompaña a este fenómeno es de una naturaleza revolucionaria, absolutamente disruptiva, lo que, unido a las innovaciones y evoluciones constantes que están llevando a los nuevos modelos a ofrecer soluciones cada día más completas y ajustadas, debe abrirnos a una reflexión amplia sobre sus posibles aplicaciones futuras.

A un jurista que, en el año 2000, entrara en el nuevo milenio, le costaría creer que, en poco más de veinte años, su expediente judicial se iba a encontrar totalmente digitalizado, con documentos procesales que van y vienen de forma telemática y algunas de sus actuaciones orales realizadas por videoconferencia. Viendo las transformaciones tecnológicas que ha sufrido la Administración de Justicia en ese período, antes del advenimiento de la revolución de la IA, se nos antoja complejo describir el escenario en el que el Poder Judicial podrá moverse cuando la IA Generativa —u otras alternativas que puedan surgir— alcance un desarrollo y evolución mayor que el que tiene en la actualidad, después de escasos tres años desde su lanzamiento. Como ya expusimos anteriormente (Saiz Garitaonandia, A., 2023), desde la perspectiva anterior no debería sorprendernos que en unas décadas vivamos una realidad en la que determinados asuntos son resueltos por jueces AI insertos en una justicia con prevalencia humana, en la que los juicios presenciales tal vez quedarán limitados a cuestiones de especial sensibilidad o relevancia, mientras que el resto se ventilan en sesiones por videoconferencia o con otro tipo de presencia virtual. La sociedad y sus formas de interactuar están evolucionando de una forma fulgurante y, en ese contexto, la Justicia no puede quedarse al margen del cambio si no quiere ser percibida como un elemento disfuncional y desfasado dentro del Estado, alejado de los ciudadanos y ciudadanas que forman parte de esa sociedad.

6. REFERENCIAS

Aletras, N., Tsarapatsanis, D., Preotiuc-Pietro, D. & Lampos, V., "Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective", en *PeerJ Computer Science*, 2:e93, 2016.

Bainbridge, L., "Ironies of Automation", en *Automatica*, Vol. 19, No. 6. 1983.

Bommarito II, M. & Katz, D.M., 2022. *GPT takes the bar exam*. arXiv preprint arXiv:2212.14402.

Bommarito, J., Bommarito, M., Katz, D.M. & Katz, J., 2023. *GPT as knowledge worker: A zero-shot evaluation of (AI) CPA capabilities*. arXiv preprint arXiv:2301.04408.

Calaza López, Sonia, "*Inteligencia Artificial e Inteligencia Natural de la Justicia: Enemigas íntimas por más de diez razones y una "sinrazón"*", Diario LA LEY, N.º 99, Sección Ciberderecho, 22 de octubre de 2025.

Coeckelberg, M., "*AI Ethics*", MIT Press, Cambridge, 2020.

Danziger, S., Levav, J. & Avnaim-Pesso, L., "Extraneous factors in judicial decisions", en *Proceedings of the National Academy of Sciences*, 108 (17), Apr 2011.

Ercilla García, J., "*IA Agéntica para la admisión del RCUD: método, resultados y gobernanza*", en Diario LA LEY, N.º 99, Sección Ciberderecho, 3 de noviembre de 2025).

Friedman, B. & Nissenbaum, H., "Bias in computer systems", en *ACM Transactions on Information Systems*, Volume 14, Number 3, July 1996.

Ilustre Colegio de la Abogacía de Madrid, "*Guía ICAM de buenas prácticas para el uso de la Inteligencia Artificial (IA) en la Abogacía*", Madrid, 2025.

Katz, D.M., Bommarito, M. & Blackman, J., "A General Approach for Predicting the Behavior of the Supreme Court of the United States", *PLoS ONE* 12(4): e0174698, 2017.

Katz, D., Bommarito, M., Gao, S. & Arredondo, P., *GPT-4 Passes the Bar Exam* (March 15, 2023). SSRN: https://ssrn.com/abstract=4389233.

Plous, S., "The Psychology of Prejudice, Stereotyping, and Discrimination", en *PLOUS* (ed.), Understanding Prejudice and Discrimination, McGraw-Hill, New York, 2003.

Saiz Garitaonandia, Alberto, "*Deus ex machina al estrado: Inteligencia Artificial y resolución de conflictos. Reflexiones para una aproximación flexible a un fenómeno en constante evolución*", en Diario La Ley, 8 de septiembre de 2023.

Saiz Garitaonandia, Alberto, "*Inteligencia artificial y profesiones jurídicas. El camino hacia un destino todavía incierto*", en Estudios sobre los efectos de la aplicación de la inteligencia artificial en el sistema jurídico, Coords. Richard González y Poelemans de Lara, J.M. Bosch, 2024.

Saiz Garitaonandia, Alberto, "Prompting, alucinaciones y lo que Kahneman puede enseñarnos del debido uso slow de la Inteligencia Artificial en el ámbito jurídico", en *Estudios Jurídicos en Homenaje al Profesor Víctor Moreno Catena*, Ed. Tirant lo Blanch, 2025.

Sourdin, T., "*Judges, Technology and Artificial Intelligence. The Artificial Judge*", Edward Elgar Publishing, Cheltenham (UK), 2021.

Sulea, O.M., Zampierini, M., Vela, M. & Van Genabith, J., "Predicting the Law Area and Decisions of French Supreme Court Cases", en *Proceedings of the International Conference Recent Advances in natural Language Processing*, INCOMA Ltd., Varna, 2017.

Surden, H., "*Machine Learning and Law*", Washington Law Review, Vol. 89, 1, 2014.

Susskind, R., "*Online Courts and the Future of Justice*", Oxford University Press, Oxford, 2019.

Terwiesch, C., "*Would Chat GPT Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course*", Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania, 2023.

UNESCO, "*Guidelines for the use of AI systems in Courts and Tribunals*", Paris, 2025.

Wu, T., "Will Artificial Intelligence eat the Law? The rise of Hybrid Social-Ordering Systems", en *Columbia Law Review*, V. 119, N. 7, 2019.

Yeung, Wang Ngai, Di Clemente, Riccardo & Lambiotte, Renaud. "Garbage in garbage out? Impacts of data quality on criminal network intervention", *EPJ Data Science*, 2025, vol. 14, no 1.

Završnik, A., "Algorithmic Justice: Algorithms and Big Data in Criminal Justice Settings", en *European Journal of Criminology*, 1, 2019.

Zerilli, J., Knott, A., Maclaurin, J. & Gavaghan, C., "Algorithmic Decision-Making and the Control Problem", en *Minds & Machines*, 29, 2019.

Zerilli, J., "*A Citizen´s Guide to Artificial Intelligence*" MIT Press, Cambridge, 2021.

ESTUDIOS

A lo largo de estas páginas, con estudios jurídicos y empíricos propios, se resalta la preocupación por el modo de asegurar tanto el respeto de la normativa de protección de datos a la hora del manejo de herramientas predictivas de la reincidencia y la violencia, en el ámbito penitenciario, como, asimismo, la fiabilidad y calidad de unos datos que, basados en modelos estadísticos construidos sobre variables predefinidas de personas con perfiles similares a las evaluadas, deberían quedar libres de todo sesgo o ser fuentes de estigmatización, y ayudar en la toma de decisiones individualizadas, sin sustituir a los profesionales responsables cuya capacitación en el uso de estas herramientas debe quedar igualmente garantizada. La necesidad de establecimiento de una arquitectura clara de responsabilidades y dación de cuentas, así como la transparencia del funcionamiento interno del sistema algorítmico utilizado —que debería contribuir a la reinserción social y no dar lugar a «automatismos punitivos»—, son cuestiones clave abordadas, de forma práctica, por diferentes profesionales, desde distintas disciplinas. Precisamente esta interdisciplinariedad y contraste de ideas y métodos interesarán a académicos, estudiantes, gestores públicos, profesionales del Derecho, la Criminología, la Psicología, la Ciencia Política, la Ingeniería, el Trabajo y la Educación Social, etc.

ISBN: 978-84-1085-891-2

AENOR GESTIÓN DE LA CALIDAD ISO 9001 ER-0280/2005
IQNET RECOGNIZED CERTIFICATION
AENOR GESTIÓN AMBIENTAL ISO 14001 GA-2005/0100

ARANZADI