

// CLAIRE MACHAN, JOÃO GOMES & ÁLVARO GARCÍA R.

Socio-technical AI and the reform of recidivism prediction

IA sociotécnica y reforma de la predicción de la reincidencia

Risk assessment tools are used across multiple stages of the criminal justice process, for example, diversion from prosecution, pre-trial bail or remand decisions, classification and management of individuals in custody, and informing probation or parole supervision. Within this landscape, these tools can support the work of various judicial staff, from prosecutors and lawyers to court researchers, and even probation officers. However, this article centres on their use in courts, examining specifically how judges engage with these tools as part of their decision-making process.

Within criminal justice systems, judges are entrusted with making a series of intuitive calculations to predict the probabilities of risk for future reoffending, and how an individual will behave in the community (DeMichele, Comfort, Barrick et al., 2021). These high-stakes decisions, however, are made under conditions that are rarely ideal: information can be incomplete, record-keeping fragmented across agencies (such as police, probation, social or health services), or historically shaped by inequalities that reproduce themselves in the data utilised (systematic biases in police surveillance, arrests and convictions), whether related to decisions concerning pre-trial conditions, sentencing, rehabilitation or release.

Risk assessment is employed at various stages of judicial case management to predict outcomes and in turn support more “reliable” and informed decision-making, however, the approach to this assessment has a direct implication for fairness and transparency in judicial systems. Early practices relied on unstructured professional judgement, what is frequently referred to as first-generation assessment, were primarily based on personal views or professional “gut feelings” that have been shaped by education and experience (Bonta, 1996). As a result, that approach lacks standard protocols and is considered highly susceptible to bias. On the other hand, actuarial approaches, also referred to as second-generation risk assessments, classify risks utilising structured statistical models, which can result in greater consistency and standardisation in decision-making (Barbaree et al., 2006; Harris et al., 2003).

More recent developments in risk assessment include third-generation structured risk/needs instruments, which integrate both clinical judgment and actuarial methods. Fourth-generation tools build on this approach by more effectively accounting for individual responsiveness to interventions. Together, these instruments support treatment planning and service delivery, in addition to informing case management decisions and determining the appropriate levels of supervision throughout the course of a case (Burman, Armstrong, Batchelor et al., 2007).

While these tools are generally regarded as objective and authoritative tools in judicial and correctional contexts, the scores they produce can still misrepresent risks in ways that undermine their perceived legitimacy (Esthappan, 2024); misinterpreting relevant risk factors, socio-demographic and gender disparities, as well as long-term trends. A growing body of research highlights that when judicial professionals rely on datasets that encode incomplete data and systemic biases, any assessment of risk built on that data can mirror and reinforce those inequalities. That is, unless design, governance, and use are explicitly oriented toward fairness and transparency (Arowosegbe, 2023; Dancy & Zalnieriute, 2025; FRA, 2020).

Las herramientas de evaluación del riesgo se usan en varias etapas del proceso penal, por ejemplo, para la derivación antes de la acusación, las decisiones sobre libertad provisional o prisión preventiva, la clasificación y gestión de personas bajo custodia y la supervisión de penas alternativas o de libertad anticipada condicional. En este contexto, estas herramientas pueden apoyar el trabajo de distintos actores judiciales, desde fiscales y abogados hasta personal de apoyo a los tribunales e incluso profesionales de supervisión comunitaria. Sin embargo, este artículo se centra en su uso en los tribunales y, en particular, en cómo los jueces recurren a estas herramientas dentro de su toma de decisiones.

En los sistemas de justicia penal, los jueces deben hacer una serie de cálculos intuitivos para prever la probabilidad de reincidencia futura y cómo se comportará una persona en la comunidad (DeMichele, Comfort, Barrick et al., 2021). Sin embargo, estas decisiones de alto impacto se adoptan en condiciones que rara vez son ideales: la información puede estar incompleta, los registros fragmentados entre organismos—como policía, supervisión comunitaria o servicios sociales y de salud—o históricamente marcados por desigualdades que se reproducen en los datos utilizados (sesgos sistemáticos en vigilancia policial, detenciones y condenas), ya sea en decisiones sobre medidas cautelares, sentencia, rehabilitación o puesta en libertad.

La evaluación del riesgo se emplea en distintas fases de la gestión judicial de los casos para predecir resultados y así respaldar decisiones más «fiables» e informadas. Sin embargo, el modo en que se realiza esa evaluación tiene implicaciones directas para la equidad y la transparencia de los sistemas judiciales.

Las primeras prácticas se basaban en el juicio profesional no estructurado, lo que suele denominarse evaluación de primera generación, y descansaban sobre todo en opiniones personales o «corazonadas» profesionales moldeadas por la formación y la experiencia (Bonta, 1996). Por ello, ese enfoque carece de protocolos estándar y se considera especialmente vulnerable al sesgo. En cambio, los enfoques actuariales, también llamados evaluaciones de riesgo de segunda generación, clasifican el riesgo mediante modelos estadísticos estructurados, lo que puede aportar mayor coherencia y estandarización a la toma de decisiones (Barbaree et al., 2006; Harris et al., 2003).

Los avances más recientes incluyen instrumentos estructurados de riesgo y necesidades de tercera generación, que integran el juicio clínico y los métodos actuariales. Las herramientas de cuarta generación desarrollan este enfoque al tener mejor en cuenta la capacidad de respuesta individual a las intervenciones. En conjunto, estos instrumentos apoyan la planificación del tratamiento y la prestación de servicios, además de orientar la gestión del caso y determinar los niveles adecuados de supervisión a lo largo de este (Burman, Armstrong, Batchelor et al., 2007).

Aunque estas herramientas suelen considerarse objetivas y autorizadas en contextos judiciales y penitenciarios, las puntuaciones que generan pueden seguir representando mal el riesgo y debilitar su legitimidad percibida (Esthappan, 2024), al interpretar de forma errónea factores relevantes, disparidades sociodemográficas y de género, así como tendencias de largo plazo.



Risk assessment in the era of AI

In the age of machine learning, risk assessments are entering an era of AI-based prediction, wherein we are seeing an increase in the use of algorithms trained on large datasets to produce risk profiles. This shift brings both potential for improved accuracy and heightened concern about opacity and discrimination.

The widely discussed *State v Loomis* (881 N.W.2d 749) case illustrates these challenges: a defendant challenged the use of a proprietary risk-assessment score in sentencing because neither he nor the court could scrutinise the factors and weights used to generate the score (*State v. Loomis*, 2016). In terms of more specified discrimination, subsequent investigations, such as analyses of a proprietary risk-assessment system on defendants in Broward County, Florida, revealed that although overall accuracy was similar across racial groups, false-positive and false-negative error rates differed markedly with black defendants who did not offend nearly twice as likely to be misclassified as high risk (Larson, et al., 2016). Such disparities are a mathematical consequence of balancing competing fairness metrics in the presence of different base rates of recidivism (Chouldchova, 2017).

Principles to address these issues are now reflected in federal guidance, such as the United States Department of Justice's analysis of AI in criminal justice (U.S. DoJ, 2024), which highlights the need for transparency, accountability, and safeguards against bias in predictive tools and risk assessments. Similarly, the OMB's Memorandum M-24-10 ("Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence"; Office of Management and Budget, 2024)), and the National Institute of Standards and Technology (NIST) AI Risk Management Framework (NIST, 2024) reinforces expectations of explainability, fairness, and rights protective deployment across high impact public sector applications, including correctional settings.

In parallel, inconsistent definitions of recidivism and heterogeneous recording practices complicate cross-case and cross-jurisdiction comparisons; the very context in which judges and policy-makers seek coherence (Yukhnenko, Farouki, & Fazel, 2023). Institutions such as the European Union Agency for Fundamental Rights (FRA) have further warned that AI-assisted tools used by public authorities can affect fundamental rights where datasets or models are opaque or unrepresentative. This underscores an obligation to ensure non-discrimination, transparency, and meaningful human oversight (FRA, 2020).

The Council of Europe similarly stresses, in its 'Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment', that although digital tools, including AI in courts, are becoming more prevalent, human oversight remains essential (CEPEJ, 2018). This aligns with the principle of "human control", which holds that the

Un número creciente de estudios muestra que, cuando los profesionales de la justicia se apoyan en conjuntos de datos incompletos y sesgados por el sistema, toda evaluación del riesgo construida sobre ellos puede reflejar y reforzar esas desigualdades. Esto solo puede evitarse si el diseño, la gobernanza y el uso se orientan explícitamente hacia la equidad y la transparencia (Arowosegbe, 2023; Dancy & Zalnieriute, 2025; FRA, 2020).

La evaluación del riesgo en la era de la IA

En la era del *machine learning*, la evaluación del riesgo está entrando en una etapa de predicción basada en IA, en la que aumenta el uso de algoritmos entrenados con grandes volúmenes de datos para generar perfiles de riesgo. Este cambio trae consigo tanto la posibilidad de mejorar la precisión como una mayor preocupación por la opacidad y la discriminación.

El conocido caso *State v Loomis* (881 N.W.2d 749) ilustra bien estos retos: un acusado cuestionó el uso de una puntuación de riesgo propietaria en la sentencia porque ni él ni el tribunal podían examinar los factores y ponderaciones utilizados para generarla (*State v. Loomis*, 2016). En cuanto a formas más concretas de discriminación, investigaciones posteriores —como los análisis de un sistema propietario de evaluación del riesgo aplicado a acusados en el condado de Broward, Florida— mostraron que, aunque la precisión global era similar entre grupos raciales, las tasas de falsos positivos y falsos negativos diferían notablemente: los acusados negros que no reincidían tenían casi el doble de probabilidades de ser clasificados erróneamente como de alto riesgo (Larson et al., 2016). Estas disparidades son una consecuencia matemática de equilibrar métricas de equidad que compiten entre sí cuando existen distintas tasas base de reincidencia (Chouldchova, 2017).

Los principios para abordar estos problemas ya se reflejan en orientaciones federales, como el análisis del Departamento de Justicia de Estados Unidos sobre la IA en la justicia penal (U.S. DoJ, 2024), que subraya la necesidad de transparencia, rendición de cuentas y salvaguardias frente al sesgo en las herramientas predictivas y las evaluaciones del riesgo. Del mismo modo, el Memorandum M-24-10 de la OMB ("Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence"; Office of Management and Budget, 2024) y el Marco de Gestión de Riesgos de IA del NIST (NIST, 2024) refuerzan las exigencias de explicabilidad, equidad y protección de derechos en aplicaciones públicas de alto impacto, también en el ámbito penitenciario.

Al mismo tiempo, las definiciones inconsistentes de reincidencia y las prácticas heterogéneas de registro dificultan las comparaciones entre casos y jurisdicciones, precisamente el contexto en el que jueces y responsables públicos buscan coherencia (Yukhnenko, Farouki, & Fazel, 2023). Instituciones como la Agencia de los Derechos Fundamentales de la Unión Europea (FRA) también han advertido de que las herramientas asistidas por IA utilizadas por autoridades públicas pueden afectar a derechos fundamentales cuando los conjuntos de datos o los modelos son opacos o no representativos. Esto refuerza la obligación de garantizar la no discriminación, la transparencia y una supervisión humana significativa (FRA, 2020).

El Consejo de Europa subraya igualmente, en su 'Carta ética sobre el uso de la inteligencia artificial en los sistemas judiciales y su entorno', que, aunque las herramientas digitales —incluida la IA en los tribunales— son cada vez más frecuentes, la supervisión humana sigue siendo esencial (CEPEJ, 2018). Esto coincide con el principio de "control humano", según el cual la responsabilidad última y la autoridad para decidir deben seguir en manos humanas, a fin de

ultimate responsibility and decision-making authority must rest with human decision-makers to safeguard professional autonomy and citizens' rights (Democritis University of Thrace, 2025).

Legal consideration must also be given to the EU AI Act (Regulation (EU) 2024/1689) and Recommendation CM/Rec(2024)5 on the ethical and organisational aspects of the use of AI and related digital technologies by prison and probation services. Both emphasise the necessity of transparency, traceability, equality and non-discrimination in order to presuppose high-integrity data and explainable decision support (Entcheva & Mazilescu, 2024). In this regard, the Commission provides further guidelines on prohibited artificial intelligence practices, however, they note a specific exception for supporting human-led risk assessment and the prediction of criminal offences; provided such assessments are based on objective and verifiable facts directly linked to a criminal behaviour (Article 5(1)(d) AI Act, p. 66).

To uphold these standards, judges and decision-makers can benefit from the use of AI tools for recidivism prediction that support them in recognising related criminogenic and non-criminogenic patterns, identify blind spots, and reason with clearer, more coherent and robust evidence based on historical data – AI tools should therefore be enhancing, rather than replacing, judicial autonomy and accountability.

Achieving fair and transparent assessments: the FAIR-PreSONS Solution

In order to support judges distinguish lower-risk individuals who may not need custody, and focus attention and support to higher risk individuals where it can most realistically reduce harm and improve reintegration outcomes, AI-based risk prediction tools should meet two basic conditions: transparency and bias control. Transparency upholds that the factors informing an assessment, and the limits of what it can conclude, are clear enough for judges and practitioners to scrutinise and challenge. Bias control infers checking for, and reducing where possible, uneven error patterns or disparate impacts across groups rather than relying on overall accuracy alone.

This logic underpins the FAIR-PreSONS solution being developed by an international partnership of experts in risk assessment, AI developers and regulators and justice-sector staff.

The solution is built around four core priorities. The first is improving data quality through a systematic methodology for the collection, digitisation and semantic enrichment of data from offender management systems in Greece, Portugal and Bulgaria. The second is fairness, including integrating gender-sensitive variables and identifying disparities. The third is explainability, so that participants can interpret an assessment rather than defer to it. The fourth is capacity-building, because no justice-sector tool is responsible unless the people using it are trained to understand both its value and its limits.

These priorities align closely with the strict requirements governing the processing of personal data for criminal justice purposes in the European Union. Article 10 of the GDPR restricts the processing of data relating to criminal convictions and offences to circumstances under the control of an official authority or pursuant to Union or Member State law that provides appropriate safeguards. To ensure GDPR alignment, FAIR PreSONS incorporates a privacy by design approach in its development: sensitive identifiers are minimised; access is restricted; and synthetic (Portugal) or anonymised (Greece and Bulgaria) datasets are used for development and testing wherever possible. Documented data flows and robust audit trails further ensure compliance and accountability.

proteger la autonomía profesional y los derechos de la ciudadanía (Democritis University of Thrace, 2025).

También deben tenerse en cuenta el Reglamento de IA de la UE (Reglamento (UE) 2024/1689) y la Recomendación CM/Rec(2024)5 sobre los aspectos éticos y organizativos del uso de la IA y de tecnologías digitales relacionadas por parte de los servicios penitenciarios y de supervisión comunitaria. Ambos destacan la necesidad de transparencia, trazabilidad, igualdad y no discriminación como base para exigir datos de alta integridad y un apoyo a la decisión explicable (Entcheva & Mazilescu, 2024). A este respecto, la Comisión ofrece orientaciones adicionales sobre las prácticas de inteligencia artificial prohibidas, aunque prevé una excepción específica para apoyar evaluaciones del riesgo dirigidas por personas y la predicción de delitos, siempre que se basen en hechos objetivos y verificables directamente vinculados a una conducta delictiva (artículo 5(1)(d) del AI Act, p. 66).

Para mantener estos estándares, los jueces y responsables de decisión pueden beneficiarse de herramientas de IA para la predicción de la reincidencia que les ayuden a reconocer patrones criminógenos y no criminógenos, identificar puntos ciegos y razonar con pruebas más claras, coherentes y sólidas basadas en datos históricos. Estas herramientas deberían reforzar, y no sustituir, la autonomía y la responsabilidad judiciales.

Lograr evaluaciones justas y transparentes: la solución FAIR-PreSONS

Para ayudar a los jueces a distinguir a las personas de menor riesgo que quizá no necesiten internamiento y a concentrar la atención y el apoyo en quienes presentan mayor riesgo, allí donde esto pueda reducir el daño de forma más realista y mejorar la reintegración, las herramientas de predicción del riesgo basadas en IA deben cumplir dos condiciones básicas: transparencia y control del sesgo. La transparencia exige que los factores que informan una evaluación, y los límites de lo que esta puede concluir, sean lo bastante claros para que jueces y profesionales puedan examinarlos y cuestionarlos. El control del sesgo implica detectar y, cuando sea posible, reducir patrones de error desiguales o impactos dispares entre grupos, en lugar de fijarse solo en la precisión global.

Esta lógica sustenta la solución FAIR-PreSONS que está siendo desarrollada por una alianza internacional de especialistas en evaluación del riesgo, desarrolladores y reguladores de IA, y personal del sector de la justicia.

La solución se articula en torno a cuatro prioridades. La primera es mejorar la calidad de los datos mediante una metodología sistemática para la recogida, digitalización y enriquecimiento semántico de datos procedentes de sistemas de gestión de personas condenadas en Grecia, Portugal y Bulgaria. La segunda es la equidad, incluida la incorporación de variables sensibles al género y la identificación de disparidades. La tercera es la explicabilidad, para que quienes participan puedan interpretar una evaluación en lugar de limitarse a acatarla. La cuarta es el fortalecimiento de capacidades, porque ninguna herramienta del sector justicia es responsable si quienes la usan no están formados para comprender tanto su valor como sus límites.

Estas prioridades se ajustan estrechamente a los estrictos requisitos que regulan el tratamiento de datos personales con fines de justicia penal en la Unión Europea. El artículo 10 del RGPD restringe el tratamiento de datos relativos a condenas e infracciones penales a supuestos bajo el control de una autoridad oficial o con arreglo al Derecho de la Unión o de los Estados miembros que establezca

Building a real-world logic and explainable system

An additional central innovation of the solution is the Reoffending Risk Assessment Ontology (RRAO), which organises information about people, decisions, sanctions¹ and sentence execution, and institutional pathways in a consistent way. A shared structure as the backbone of the solutions' data portal further helps to bring those records into the same frame, which ensures analysis is more coherent and easier to scrutinise. The RRAO captures the complexity of the criminal justice lifecycle in semantic classes and logic, linking data under correctional and judicial institutions, legal outcomes (including court decisions), human subjects and historic profiles, and sentence-execution information.

This semantic enrichment in data collection and treatment is a first step toward bias mitigation. Based on a Human-Centered Ontology Engineering (HCME) methodology, the model blends human expertise with language-model assistance for scaffolding. By incorporating variables like occupation and education level, socio-economic factors linked to recidivism, into a transparent knowledge graph², the framework distinguishes between a propensity for crime and the systemic effects of poverty (FAIR-PreSONS, 2026). Maintaining this mapping at the forefront reduces the risk of a "black box" analysis built on unknown or mis-specified variables such as race or gender. The system also uses an adversarial debiasing approach to detect whether protected characteristics can still be inferred from the prediction or whether the system's outcomes fall unevenly across groups. The reasoning is adjusted to reduce that dependency. It then combines this semantic structure with a hybrid modelling approach that joins the ability to represent complex relationships in the data with knowledge graphs and the ability to learn patterns of neural architectures. This "hybrid AI" approach overcomes the limitations of symbolic AI (rigidity) and neural AI (lack of explainability).

Result: a judicial co-pilot for decision support

The system's practical model is intelligent analysis, variable weighting and visualisation, keeping the system squarely in the category of decision support rather than autonomous decision-making. It does not determine outcomes; rather, it handles data-intensive tasks, helping to identify blind spots and present evidence that can be weighed alongside legal considerations and the particulars of each case. The result is a "co-pilot" user interface, primarily for use by judges, alongside their ethical and legal reasoning. Its utility can also be extended to other justice professionals involved in decisions wherein risk of recidivism are pertinent, such as prosecutors, lawyers, court researchers, probation professionals). FAIR-PreSONS represents a pioneering effort to modernise recidivism prediction. By moving away from proprietary "black boxes" toward an open-source, semantically enriched socio-technical framework, the initiative addresses algorithmic discrimination while adhering to the rule of law. The methodologies piloted here provide a blueprint for high-risk AI under the EU AI Act, ensuring that accuracy, fairness and transparency are achieved through rigorous design and ethical governance.

garantías adecuadas. Para ajustarse al RGPD, FAIR-PreSONS incorpora un enfoque de privacidad desde el diseño en su desarrollo: se minimizan los identificadores sensibles, se restringe el acceso y, siempre que es posible, se utilizan conjuntos de datos sintéticos (Portugal) o anonimizados (Grecia y Bulgaria) para el desarrollo y las pruebas. Los flujos de datos documentados y unos registros de auditoría sólidos refuerzan además el cumplimiento y la rendición de cuentas.



Construir un sistema explicable y con lógica del mundo real

Otra innovación central de la solución es la Ontología de Evaluación del Riesgo de Reincidencia (RRAO), que organiza de forma coherente la información sobre personas, decisiones, sanciones¹ y ejecución de la pena, así como los itinerarios institucionales. Una estructura compartida, como columna vertebral del portal de datos de la solución, ayuda además a situar esos registros en un mismo marco, lo que hace que el análisis sea más coherente y más fácil de examinar. La RRAO capta la complejidad del ciclo de justicia penal mediante clases semánticas y lógica, vinculando datos de instituciones judiciales y correccionales, resultados jurídicos —incluidas decisiones judiciales—, sujetos humanos y perfiles históricos, e información sobre la ejecución de la condena.

Este enriquecimiento semántico en la recogida y el tratamiento de datos es un primer paso hacia la mitigación del sesgo. Basado en una metodología de Ingeniería Ontológica Centrada en las Personas (HCME), el modelo combina la experiencia humana con la ayuda de modelos de lenguaje para la estructuración del entorno de datos. Al incorporar variables como la ocupación y el nivel educativo, factores socioeconómicos vinculados a la reincidencia, en un grafo de conocimiento² transparente, el marco distingue entre la propensión al delito y los efectos sistémicos de la pobreza (FAIR-PreSONS, 2026). Mantener este mapeo en primer plano reduce el riesgo de un análisis de "caja negra" construido sobre variables desconocidas o mal definidas, como la raza o el género.

El sistema también utiliza un enfoque de mitigación adversarial del sesgo para detectar si el modelo puede inferirse características protegidas a partir de la predicción o si los resultados del sistema se distribuyen de manera desigual entre grupos. Después, ajusta el razonamiento para reducir esa dependencia. A continuación, combina esta estructura semántica con un enfoque de modelización híbrida que une la capacidad de los grafos de conocimiento para representar relaciones complejas en los datos con la capacidad de las arquitecturas neuronales para aprender patrones. Este enfoque de "IA híbrida" supera las limitaciones de la IA simbólica (rigidez) y de la IA neuronal (falta de explicabilidad).

References | Referencias:
Angelis, S., Priho, J., Sykizota, A., Markov, D., Chatzistamatis, S., Spirou, S., Tsakouras, G., & Kotis, K. I. (2025, July). RRAO: An ontology for the representation of reoffending risk assessment knowledge. In 2025 26th International Conference on Information, Intelligence, Systems & Applications (IISA).
Aronowicz, J. O. (2023). Data bias, intelligent systems and criminal justice outcomes. *International Journal of Law and Information Technology*, 33(1).
Burman, M., Armstrong, S., Batchelor, S., McNeil, F., & Nicholson, J. (2007). *Research and practice in risk assessment and risk management of children and young people engaging in offending behaviour*. Scottish Centre for Crime and Justice Research.
Choudhurova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153–163.
Council of Europe, Committee of Ministers. (2024, October 9). Recommendation CM/Rec(2024)15 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.
Dancy, T., & Zalmariute, M. (2025). AI and transparency in judicial decision-making. *Oxford Journal of Legal Studies*. Advance online publication.
Democritus University of Thrace. (2025, September 22). *Respecting and strengthening human autonomy when using artificial intelligence in the administration of justice*.
DeMichele, M., Comfort, M., Barrick, K., & Baumgartner, P. (2021). The intuitive-override model: Nudging judges toward pretrial risk assessment instruments. *Fed Probation*, 85, 22.
Encheva, K., & Mazliewski, J. (2024). Artificial intelligence and digitalisation of judicial cooperation: The main provisions in recent EU legislation. *eu-crim: The European Criminal Law Association's Forum*, 19(3), 202–205.
Esthappan, S. (2024). Assessing the risks of risk assessments: Institutional tensions and data-driven judicial decision-making in US pretrial hearings. *Social Problems*, 60, 1–35.
European Commission for the Efficiency of Justice (CEPEJ). (2014). *European ethical charter on the use of artificial intelligence in judicial systems and their environment*. Council of Europe.
European Commission. (2025). *Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/2689 (AI Act)*.
European Union Agency for Fundamental Rights. (2020). *Getting the future right – Artificial intelligence and fundamental rights*.
FAIR-PreSONS. (n.d.). *Future-proofing the legal profession: Fostering AI understanding, competence, and skills across the criminal justice sector*.
Larson, J., Matu, S., Kirchner, L., & Angwin, J. (2016). *How we analyzed the COMPAS recidivism algorithm*. ProPublica.
National Institute of Standards and Technology. (2024, July 26). *Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1)*.
State v. Loomis, 881 N.W.2d 749 (Wis. 2016).
U.S. Department of Justice. (2024). *Artificial intelligence and criminal justice: Final report*.
Yakimovskiy, D., Farouki, L., & Faal, S. (2023). Criminal recidivism rates globally: A 6-year systematic review update. *Journal of Criminal Justice*, 88, Article 102115.

Resultado: un copiloto judicial para apoyar la decisión

El modelo práctico del sistema se basa en análisis inteligente, ponderación de variables y visualización, lo que lo sitúa claramente en la categoría de apoyo a la decisión y no de toma de decisiones autónoma. No determina resultados; más bien, se ocupa de tareas intensivas en datos, ayudando a identificar puntos ciegos y a presentar pruebas que pueden valorarse junto con las consideraciones jurídicas y las particularidades de cada caso.

El resultado es una interfaz de usuario tipo "copiloto", pensada principalmente para jueces, junto a su razonamiento ético y jurídico. Su utilidad también puede extenderse a otros profesionales de la justicia que intervienen en decisiones en las que el riesgo de reincidencia es relevante, como fiscales, abogados, personal de apoyo a los tribunales y profesionales de supervisión comunitaria.

FAIR-PreSONS representa un esfuerzo pionero para modernizar la predicción de la reincidencia. Al alejarse de las "cajas negras" propietarias y avanzar hacia un marco sociotécnico de código abierto y enriquecido semánticamente, la iniciativa aborda la discriminación algorítmica sin apartarse del Estado de derecho. Las metodologías aquí pilotadas ofrecen un modelo para la IA de alto riesgo conforme al Reglamento de IA de la UE, garantizando que la precisión, la equidad y la transparencia se logren mediante un diseño riguroso y una gobernanza ética.



CLAIRE MACHAN

Rehabilitation, Reintegration, and Community Portfolio Coordinator at IPS_Innovative Prison Systems. She holds a MSc in Forensic Mental Health and is upon completion of a PhD in Experimental Psychology. With over eight years of academic experience in applied forensic psychology, she has worked with prison services, LEAs, victim support, and the military delivering education and training. She is an Expert Member of the ICPSA Juvenile Justice Committee and active in global research societies including EAPL, BPS, APA, and SJDIM.

Coordinadora del equipo de Rehabilitación, Reintegración y Comunidad en IPS_Innovative Prison Systems. Tiene un Máster en Salud Mental Forense y está finalizando un doctorado en Psicología Experimental. Con más de ocho años de experiencia académica en psicología forense aplicada, ha trabajado con servicios penitenciarios, fuerzas del orden, apoyo a víctimas y el ámbito militar, impartiendo educación y formación. Es miembro del Comité de Justicia Juvenil de la ICPSA y participa activamente en sociedades internacionales de investigación como EAPL, BPS, APA y SJDIM.



JOÃO GOMES

Coordinator for the International Judicial Cooperation & Human Rights portfolio at IPS_Innovative Prison Systems, overseeing the design and implementation of projects related to judicial cooperation in criminal matters in the EU, as well as procedural and fundamental rights in the Union. He holds a master's degree in international relations, with a post-graduation in Intelligence Management and Security. Experienced in political and diplomatic work, with professional experiences in both internal and external services of the Ministry of Foreign Affairs of the Portuguese Republic.

Coordinador del equipo de Cooperación Judicial Internacional y Derechos Humanos en IPS_Innovative Prison Systems, supervisando el diseño y la ejecución de proyectos relacionados con la cooperación judicial en materia penal en la UE, así como con los derechos procesales y fundamentales en la Unión. Tiene un máster en Relaciones Internacionales y un posgrado en Gestión de Inteligencia y Seguridad. Cuenta con experiencia en trabajo político y diplomático, con trayectoria profesional tanto en los servicios internos como externos del Ministerio de Asuntos Exteriores de la República Portuguesa.



ÁLVARO GARCÍA R.

CTO/CIO specialising in the digitalisation of justice systems. He holds a Double European Master's in Distributed Computing and a PhD from Université de Rennes 1. His work spans cloud computing, privacy, and decentralised protocols, including research on decentralised private set intersection (dPSI) and abuse-detection tools such as Trollslayer. He has contributed across leading European R&D labs and releases open-source software to support reproducible research. As part of his work at IPS, he is currently participating in the development of the FAIR-PreSONS's AI tool for DG JUST.

CTO/CIO especializado en la digitalización de los sistemas de justicia y como director de equipos de tecnologías lleva publicando artículos que versan sobre el uso de la IA y ML desde 2018. Tiene un Doble Máster Europeo en Sistemas Distribuidos por KTH en Suecia y Técnico en Lisboa y un doctorado por la Université de Rennes 1 de Francia. Su trabajo abarca la computación en la nube, la privacidad y los protocolos descentralizados, y herramientas de caracterización y detección de contenido abusivo (i.e. Trollslayer) en la "WWW". Ha contribuido en destacados laboratorios nacionales en Iberia y europeos enfocados en I+D pública. Sus publicaciones en este ámbito incluyen contribuciones de software de código abierto para apoyar la investigación reproducible. Como parte de su trabajo en IPS, participa actualmente en el desarrollo de la herramienta de IA FAIR-PreSON para DG JUST.

¹Here "sanctions" refers broadly to custodial and non-custodial penalties imposed by courts. [Aquí, "sanciones" se refiere en sentido amplio a las penas privativas y no privativas de libertad impuestas por los tribunales].
²A knowledge graph is a structured network of entities and relationships represented as nodes and edges in which semantic meaning is explicit. [Un grafo de conocimiento (también conocido como red semántica) es una red estructurada de entidades y relaciones, representadas como nodos y aristas, en la que el significado semántico es explícito].