# IA que sugiere soluciones. El sistema español "VioGén 2" y su impacto en la libertad de decisión

por *Leonor Sanz Gallardo*

El artículo analiza de forma crítica el uso de los sistemas de inteligencia artificial con funciones de recomendación en contextos jurídicos y de seguridad, centrándose en sus efectos indirectos, pero significativos, sobre la autonomía decisoria y la protección de los derechos fundamentales. Partiendo de la distinción conceptual entre sistemas de apoyo a la toma de decisiones y la decisión automatizada, el artículo examina el sistema español de evaluación del riesgo de violencia de género "VioGén", prestando especial atención a su evolución en la versión "VioGén 2". El análisis pone de manifiesto cómo, en la práctica operativa, las recomendaciones algorítmicas tienden a producir efectos vinculantes de hecho, a través de mecanismos de anclaje cognitivo, estandarización procedimental y opacidad metodológica, lo que reduce la eficacia de la supervisión humana. Estas dinámicas inciden tanto en la protección de las víctimas como en el entorno informativo que orienta la decisión judicial. El trabajo enmarca estas cuestiones críticas a la luz del derecho europeo e internacional en materia de inteligencia artificial y derechos humanos, defendiendo la necesidad de garantías de transparencia, verificabilidad y control humano.



## 1. Introducción: qué significa "IA que sugiere soluciones" en el ámbito judicial

Un sistema que formula recomendaciones algorítmicas es un tipo de inteligencia artificial que procesa datos mediante un modelo matemático o estadístico para generar una salida orientativa, como puede ser puntuación, categoría, predicción, sugerencia. Esta salida no tiene carácter vinculante y no sustituye la decisión humana; su finalidad declarada es únicamente informar a quien debe adoptar la decisión final. Este tipo de sistemas se caracterizan por no reemplazar el juicio humano, transformar los datos en una interpretación estructurada de la realidad y producir resultados estandarizados y repetibles que permiten organizar la información sin ejecutar por sí mismos acción alguna ni adoptar decisiones autónomas. A diferencia de los sistemas de decisión automatizada, no desencadenan medidas por sí solos, sino que aportan elementos que deben ser valorados por una persona.

En el ámbito judicial, donde la resolución debe ser personal, motivada y fruto de una valoración crítica, esta distinción resulta esencial.

Aunque los sistemas de recomendación algorítmica no adoptan decisiones ni poseen autoridad vinculante, su funcionamiento produce efectos psicológicos,

operativos e institucionales que pueden influir, orientar o incluso determinar *de facto* la decisión humana, con especial incidencia en la esfera de los derechos fundamentales. Al procesar datos y generar recomendaciones, estos sistemas pueden introducir errores, sesgos o distorsiones que afectan a derechos como la igualdad, la no discriminación, la privacidad o el acceso a la justicia, tal y como han señalado tanto el Alto Comisionado de Naciones Unidas para los Derechos Humanos como la Agencia Europea de Derechos Fundamentales (FRA)[1].

La convergencia entre los principales principios internacionales, entre los que se encuentran el *Convenio Marco del Consejo de Europa sobre Inteligencia Artificial* (2024)[2], de la *Recomendación de la UNESCO sobre la Ética de la IA* (2021), de las *Directrices Éticas de la Unión Interparlamentaria sobre autonomía humana y supervisión*, así como de la doctrina emergente del derecho al control humano, muestra que la regulación de los sistemas de IA debe partir del respeto y garantía de los derechos humanos, imponiendo límites y salvaguardas claras. Todos estos instrumentos coinciden en que, cuando estos sistemas se utilizan en contextos públicos sensibles, deben operar con transparencia, auditabilidad, comprensibilidad metodológica y posibilidad real de corrección humana, evitando que sus recomendaciones se conviertan en automatismos no controlados que condicionen decisiones institucionales.

Una vez establecida esta arquitectura conceptual y jurídica que obliga a preservar el control humano efectivo, resulta especialmente útil analizar un caso concreto donde estas tensiones se manifiestan con claridad: el sistema español de evaluación del riesgo de violencia de género conocido como "VioGén", en funcionamiento desde 2007 y reformado en 2025 bajo la denominación "VioGén 2".

## 2. Antecedentes: "VioGén" y su evolución a "VioGén 2"

Antes de analizar en profundidad el sistema "VioGén" y explicar en qué medida constituye un ejemplo paradigmático del impacto que los sistemas de IA pueden tener en la libertad de decisión en el ámbito jurídico, conviene recordar que se trata de un instrumento diseñado para evaluar el riesgo en casos de violencia de género, cuya operatividad puede, en la práctica, condicionar la actuación de los agentes responsables de la protección de las víctimas y, con ello, afectar a los derechos fundamentales en juego.

VioGén fue creado por el Ministerio del Interior y puesto en funcionamiento el 26 de julio de 2007, en cumplimiento de la *Ley Orgánica 1/2004, de 28 de diciembre, de Medidas de Protección Integral contra la Violencia de Género*. Su fundamento normativo inmediato fue la Instrucción 10/2007 de la Secretaría de Estado de Seguridad, que aprobó el primer protocolo de valoración policial del riesgo, posteriormente modificado mediante diversas instrucciones administrativas.

De este modo, VioGén fue concebido con una serie de objetivos institucionales: agrupar a todas las instituciones públicas competentes, integrar en un sistema único la información relevante sobre casos activos, realizar predicciones de riesgo basadas en factores empíricamente validados, asignar medidas de protección y seguimiento en función del nivel de riesgo, emitir avisos o alarmas automáticas ante incidencias relevantes y crear una red estatal de protección coordinada y homogénea disponible en todo el territorio nacional[3].

El sistema alimenta su base de datos mediante denuncias policiales, entrevistas a las víctimas, informes forenses, órdenes de protección, antecedentes policiales y judiciales, así como información personal y social tanto de la víctima como del agresor. Con toda esta información, VioGén aplica instrumentos actuariales automatizados, en particular la *Valoración Policial del Riesgo (VPR)* y la *Valoración Policial de la Evolución del Riesgo (VPER)*, basados en modelos estadísticos que siguen metodologías similares a los enfoques epidemiológicos utilizados en salud pública[4].

Para ello, el sistema establece cinco niveles de riesgo (no apreciado, bajo, medio, alto y extremo), cada uno de los cuales lleva asociadas medidas de protección concretas que pueden ir desde un seguimiento básico hasta vigilancia intensiva o la solicitud de una orden judicial restrictiva. Aunque el agente mantiene formalmente la facultad de adaptación, en la práctica la recomendación algorítmica delimita el conjunto inicial de medidas que deben considerarse.

1. Informe FRA *Getting the Future Right – Artificial Intelligence and Fundamental Rights* (2021).
2. https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence.
3. https://interior.gob.es/opencms/en/servicios-al-ciudadano/violencia-contra-la-mujer/sistema-viogen-2/.
4. https://s088672f9df8c57fe.jimcontent.com/download/version/1709741880/module/12490599698/name/SISTEMA%20DE%20SEGUIMIENTO%20INTEGRAL%20EN%20CASOS%20DE%20VG.pdf.

Desde 2007, VioGén ha registrado casi tres millones de valoraciones de riesgo, convirtiéndose en una plataforma nacional de coordinación entre instituciones policiales, judiciales, penitenciarias y sociales. El propio Ministerio del Interior reconoce que el sistema es un instrumento "en evolución constante", sometido a revisiones periódicas mediante instrucciones que actualizan su lógica actuarial y sus funcionalidades.

Una vez establecido el marco conceptual general y explicado por qué los sistemas algorítmicos pueden condicionar la decisión humana, es posible mostrar cómo estos riesgos se han manifestado de manera concreta en el caso de VioGén. Este sistema resulta paradigmático en la medida en que permite observar cómo un modelo de valoración automatizada puede intervenir directamente en decisiones vinculadas a la protección de derechos fundamentales como la vida, la integridad física o el acceso a la justicia.

La experiencia acumulada durante casi dos décadas muestra que VioGén presenta fallos estructurales que han tenido impacto real en la protección de las víctimas. Entre los más significativos, un análisis citado por el Human Rights Research Center[5], basado en datos publicados por *The New York Times*, reveló que 55 de 247 mujeres asesinadas tras una evaluación policial habían sido clasificadas como de riesgo "bajo" o "despreciable". Este resultado muestra que la herramienta no solo falla en casos marginales, sino en situaciones críticas donde la valoración resultó insuficiente para activar medidas adecuadas.

Un caso emblemático es el de Lobna Hemid, asesinada en 2022 después de haber sido evaluada por VioGén como de bajo riesgo. La clasificación inicial, percibida como técnicamente neutra, condicionó la respuesta institucional y contribuyó a que no se adoptaran las medidas de protección necesarias.

Otra crítica recurrente se refiere a la aceptación automática del resultado. Según la auditoría de Eticas Foundation[6], los agentes mantienen la puntuación generada por el sistema en el 95 % de las evaluaciones, pese a que el protocolo permite realizar ajustes basados en la apreciación profesional. Como consecuencia, la supervisión humana queda reducida a una formalidad, y la recomendación algorítmica, en una decisión operativa *de facto*.

El problema se ve agravado por el hecho de que, en muchos casos, las víctimas no revelan información crítica durante las entrevistas policiales debido al miedo, la dependencia económica o el estatus migratorio, lo que genera datos incompletos o sesgados. Estos sesgos de origen se transmiten al algoritmo, que infraestima el riesgo precisamente en las víctimas más vulnerables.

A ello se suman las limitaciones inherentes al enfoque actuarial. Como explica el Ministerio del Interior, los modelos utilizados en VPR y VPER se basan en marcadores estadísticos asociados a la reincidencia violenta, pero sin relación causal directa. En consecuencia, pequeñas variaciones en las respuestas pueden desplazar el nivel de riesgo asignado, sin que los modelos capten factores cualitativos o dinámicos propios del caso concreto.

Por último, la falta de transparencia del sistema constituye una de las críticas más reiteradas. La auditoría de Eticas Foundation señaló que VioGén *"no permite conocer la fórmula exacta, los pesos asignados ni la base empírica completa del modelo"*, y que tampoco ha sido sometido a auditorías técnicas externas independientes. Esta opacidad vulnera los principales estándares internacionales sobre el uso ético y responsable de la inteligencia artificial, que exigen trazabilidad, auditabilidad y supervisión humana efectiva en sistemas que afectan a derechos fundamentales. En esta línea, el Convenio Marco del Consejo de Europa sobre Inteligencia Artificial, primer tratado internacional vinculante en la materia, impone a los Estados la obligación de garantizar la trazabilidad de los sistemas y una supervisión humana adecuada a lo largo de todo su ciclo de vida. La Recomendación de la UNESCO sobre la Ética de la IA (2021) establece igualmente que estos sistemas deben operar bajo criterios de transparencia, explicabilidad y supervisión humana, en defensa de la dignidad y los derechos fundamentales de las personas. Asimismo, las Directrices Éticas de la Unión Interparlamentaria subrayan que toda IA empleada en funciones públicas debe preservar la autonomía humana y evitar cualquier forma de delegación encubierta de decisiones, insistiendo en la necesidad de garantizar un control humano continuo y efectivo.

El conjunto de estos factores erosiona las garantías que deben regir el uso de IA en contextos de derechos fundamentales. Además, la clasificación algorítmica condiciona la intensidad de la protección prestada: según datos citados por *Politico*, solo una de cada siete mujeres que solicita protección policial la recibe finalmente, en gran parte ya que la asistencia prioritaria viene condicionada por el nivel de riesgo atribuido por el sistema[7].

5. www.humanrightsresearch.org/post/spain-s-gender-violence-detection-algorithm-reveals-failures-in-saving-domestic-violence-victims.

6. https://eticasfoundation.org/wp-content/uploads/2024/12/Eticas_Audit_of_VioGen.pdf.

7. https://www.politico.eu/newsletter/ai-decoded/spains-flawed-domestic-abuse-algorithm-ban-debate-heats-up-holding-the-police-accountable-2/.

Las críticas acumuladas durante más de una década llevaron al Ministerio del Interior a impulsar en 2025 una reforma integral del sistema que dio lugar a "VioGén 2". Esta actualización se articula en torno a tres líneas de actuación estrechamente vinculadas a los fallos detectados en el modelo anterior: en primer lugar, la eliminación de la categoría "sin riesgo", cuya utilización había dado lugar a infravaloraciones especialmente graves en la práctica; en segundo lugar, la incorporación obligatoria de información cualitativa y contextual en las entrevistas policiales, orientada a reducir la incidencia de datos incompletos derivados del miedo, la dependencia económica o la situación administrativa de muchas víctimas; y, en tercer lugar, la introducción de planes de protección personalizados y una mejor coordinación interinstitucional, destinadas a evitar respuestas excesivamente dependientes de la clasificación algorítmica. Estas líneas de reforma están recogidas en el material editorial sobre la actualización del sistema difundido en 2025.

Con estas medidas, VioGén 2 no solo pretende mejorar la precisión técnica del sistema, sino, sobre todo, corregir los riesgos de automatización encubierta detectados en la práctica, ampliando los márgenes de supervisión humana y devolviendo a los operadores la capacidad de ejercer un juicio más completo y contextualizado. VioGén 2, así, constituye una respuesta institucional a los desafíos que plantea el uso de IA en la protección de derechos fundamentales.

## 3. Funcionamiento real: ¿sugerencia o decisión encubierta?

Aunque el sistema VioGén fue concebido formalmente como una herramienta de apoyo destinada a ofrecer una guía en la actuación profesional, su funcionamiento práctico revela que la recomendación algorítmica termina convirtiéndose, en una parte sustancial de los casos, en una decisión encubierta. Esta transformación no es intencional, sino que responde a mecanismos específicos que se producen de manera habitual en los sistemas de inteligencia artificial integrados en circuitos institucionales. Como se ha mostrado anteriormente, aunque el diseño oficial de VioGén establece que el algoritmo proporciona únicamente un nivel de riesgo orientativo y que la autoridad policial debe valorar el caso de manera independiente, la práctica documentada indica que los operadores siguen la recomendación algorítmica casi sin intervención humana, manteniendo el resultado automático en el 95 % de las evaluaciones según la auditoría de Eticas Foundation.

Ello ocurre porque, en la práctica, la clasificación generada por el sistema funciona como un marco cognitivo que actúa como un diagnóstico preliminar y condiciona la forma en que se interpretan la entrevista, las señales de alarma o la narrativa de la víctima. Se trata de un fenómeno típico del efecto de anclaje, en el que el primer valor recibido se convierte en referencia dominante para todo análisis posterior. A esto se suma que la carga de trabajo, la presión operativa y la necesidad de estandarización llevan a los agentes a asumir la recomendación como punto de partida operativo, desencadenando protocolos, prioridades y determinando la intensidad del seguimiento. En este contexto, la recomendación funciona como una pre-decisión, porque ya viene conectada a un repertorio de medidas policiales preconfiguradas.

A nivel psicológico, también influye el hecho de que el algoritmo no produce un texto interpretativo, sino una categoría cerrada ("bajo riesgo", "medio riesgo", "alto riesgo"), simplificando una realidad compleja en un resultado aparentemente exacto, que adquiere una autoridad técnica superior a la intuición profesional. La combinación de anclaje cognitivo, estandarización organizativa y falta de transparencia convierte la recomendación algorítmica en una referencia decisoria que condiciona la práctica policial y limita la capacidad real de revisión humana.

Estas tensiones adquieren una relevancia especial en el ámbito jurisdiccional, ya que los sistemas de IA utilizados en justicia o en seguridad están clasificados por el EU AI Act (Reglamento (UE) 2024/1689)[8] como sistemas de alto riesgo, lo que exige transparencia, trazabilidad, supervisión humana significativa y evaluaciones de impacto en derechos fundamentales, precisamente para evitar condicionamientos indebidos sobre la libertad decisoria judicial en contextos donde el juez recibe información previamente procesada por sistemas automatizados.

## 4. Impacto en la libertad decisoria judicial

La utilización de sistemas algorítmicos en el ecosistema de la justicia plantea una cuestión decisiva: cómo proteger la libertad decisoria cuando la información que llega al juez ha sido configurada previamente por un sistema automatizado? En sede judicial, dicha libertad exige la posibilidad de analizar el caso sin condicionamientos externos, evaluando los hechos y el riesgo con plena autonomía. Sin embargo, cuando la información que alcanza al juzgado procede de un sistema cuya lógica interna no es auditable,

8. https://artificialintelligenceact.eu/es/ai-act-explorer/.

se altera el entorno cognitivo en el que el juez debe deliberar. Esa configuración previa, derivada no de una intención de interferencia, sino de la manera en que los sistemas de IA organizan, sintetizan o filtran la información policial previa, puede influir en la ponderación de urgencia, proporcionalidad y riesgo en la adopción de medidas cautelares o de protección.

En el marco europeo e internacional, la autonomía decisoria del juez no se concibe únicamente como ausencia de presiones externas directas, sino como el derecho y el deber de resolver a partir de información comprensible y susceptible de verificación, sin condicionamientos procedentes de mecanismos automatizados. Así lo exige el *artículo 6 del Convenio Europeo de Derechos Humanos*, que requiere un tribunal "*independiente e imparcial*", garantía que incluye la independencia frente a influencias tecnológicas que puedan limitar la capacidad real de formar criterio. A su vez, el *artículo 47 de la Carta de Derechos Fundamentales de la Unión Europea* consagra la "*tutela judicial efectiva*", que presupone la disponibilidad de información suficiente y claramente accesible para que el juez motive su decisión sin sujeciones estructurales. A esta comprensión se suman los estándares del Consejo de Europa, recogidos por el CCJE, cuyas Opiniones (en particular la *Opinión n.º 26 (2023)* sobre tecnologías de apoyo en la justicia) insisten en que la independencia judicial incluye la capacidad de controlar la racionalidad del proceso decisorio, lo que exige que el juez pueda entender, auditar y cuestionar la información que incorpora a su motivación.

El Convenio Marco del Consejo de Europa sobre Inteligencia Artificial (2024) refuerza esta exigencia al imponer obligaciones de transparencia y supervisión humana adecuada durante todo el ciclo de vida de los sistemas algorítmicos, impidiendo que la IA opere como un filtro no controlado de la realidad que llega a los tribunales. En la misma dirección, la Recomendación de la UNESCO sobre la Ética de la IA (2021) exige la *primacía de la agencia humana* en toda toma de decisiones pública, advirtiendo de que los sistemas no transparentes pueden desplazar la capacidad de juicio profesional y debilitar la motivación individualizada de las resoluciones. Este instrumento establece que la supervisión humana debe ser real, significativa y constante, especialmente cuando la IA afecta a derechos esenciales. Por su parte, las Directrices Éticas de la Unión Interparlamentaria subrayan que la IA empleada en funciones públicas debe preservar la autonomía decisoria humana y evitar toda delegación encubierta de decisiones, insistiendo en la necesidad de un control humano permanente.

En el ámbito de la Unión Europea, el EU AI Act (Reglamento (UE) 2024/1689) califica los sistemas de IA aplicados a justicia y seguridad dentro de la categoría de "alto riesgo", sometiéndolos a los requisitos más estrictos del Reglamento: transparencia reforzada, trazabilidad documental, supervisión humana significativa, gobernanza rigurosa de datos y evaluaciones de impacto en derechos fundamentales antes de su puesta en servicio. Según la síntesis del Oxford Institute of Technology and Justice, esta categoría refleja que la IA utilizada en justicia puede influir materialmente en el debido proceso, en la igualdad de las partes y en la imparcialidad efectiva del tribunal, por lo que la UE exige salvaguardias reforzadas para evitar que la IA deforme el entorno fáctico sobre el que el juez delibera.

Desde esta perspectiva europea e internacional, la irrupción de VioGén no plantea solamente un problema técnico, sino constitucional en sentido amplio: la jurisdicción no puede operar con plena independencia cuando la realidad procesal que examina llega ya condicionada por un sistema cuya metodología, lógica interna y sesgos no resultan accesibles para el juez. La independencia judicial implica también independencia frente a infraestructuras tecnológicas no transparentes, y la tutela judicial efectiva exige que el juez pueda reconstruir críticamente el caso y motivar su resolución sin depender de una categoría estadística cerrada o de una valoración algorítmica no explicada.

Por ello, el verdadero desafío que plantean los sistemas de evaluación automatizada del riesgo no radica en una sustitución formal del juez, prohibida en el marco europeo sin excepción, sino en la merma del control epistémico que el juez debe ejercer sobre la información relevante. Preservar la libertad decisoria judicial requiere que la IA no defina los contornos de la realidad procesal, sino que se limite a aportar información comprensible, verificable y auditada, siempre subordinada al juicio humano.