4

INTELIGENCIA ARTIFICIAL, POLICÍA PREDICTIVA Y PREVENCIÓN DE LA VIOLENCIA DE GÉNERO

Miguel Ángel Presno Linera

Professor Titular Acreditado como Catedrático de Derecho Constitucional da Universidad de Oviedo, Vicepresidencia Primera del Gobierno y Ministerio de la Presidencia 31/10/2005 2, Asesor del Secretario de Estado de Relaciones con las Cortes 05/05/2004, Magistrado Suplente del Tribunal Superior de Justicia de Asturias, Licenciado en Derecho Rama General Entidad que expide el título: Universidad de Oviedo, Doutor pela Universidade de Oviedo (Cuestiones de Derecho Histórico y Actual).

doi.org/10.58725/rivjr.v1i2.39

RESUMEN

La violencia contra las mujeres es un fenómeno cada vez más frecuente en todo el mundo y tiene un enorme impacto en la vida de las víctimas, sus familias y la sociedad. En este breve estudio nos centraremos en el recurso tanto a la inteligencia artificial como los meros algoritmos predictivos para prevenir la reiteración de esa violencia adoptando, en su caso, las medidas cautelares necesarias. El estudio comienza con una breve exposición del origen y evolución de la inteligencia artificial y su progresiva incidencia en el ámbito del Derecho; a continuación, analizamos las fortalezas y debilidades de la inteligencia artificial policial y judicial y, en tercer lugar, nos ocupamos con detalle de la importante experiencia española de aplicación de algoritmos predictivos frente a la violencia de género a través del sistema VioGén.

Palabras clave: Violencia de género, inteligencia artificial, policía predictiva, algoritmos predictivos, sistema Viogén.

ABSTRACT

Violence against women is an increasingly prevalent phenomenon around the world and has an enormous impact on the lives of victims, their families and society. In this brief study we will focus on the use of both artificial intelligence and mere predictive algorithms to prevent the repetition of this violence by adopting, where appropriate, the necessary precautionary measures. The study begins with a brief exposition of the origin and evolution of artificial intelligence and its progressive incidence in the field of Law. Next, we analyze the strengths and weaknesses of police and judicial artificial intelligence and, thirdly, we deal in detail with the important Spanish experience of applying predictive algorithms against gender violence through the VioGén system.

Keywords: Gender violence, artificial intelligence, predictive policing, predictive algorithms, Viogén system

RESUMO

A violência contra as mulheres é um fenômeno cada vez mais prevalente em todo o mundo e tem um enorme impacto na vida das vítimas, das suas famílias e na sociedade. Neste breve estudo, enfocaremos o uso da inteligência artificial e dos meros algoritmos preditivos para evitar a reiteração dessa violência e para que se adote, quando oportuno, as medidas cautelares necessárias. O estudo inicia-se com uma breve exposição sobre a origem e evolução da inteligência artificial e sua progressiva incidência no campo do Direito. Em seguida, analisamos os pontos fortes e fracos da inteligência

artificial policial e judicial e, em terceiro lugar, tratamos em detalhes da importante experiência espanhola de aplicação de algoritmos preditivos contra a violência de gênero através do sistema VioGén.
Palavras-chave: Violência de gênero, inteligência artificial, policiamento preditivo, algoritmos preditivos, sistema VioGén.

RÉSUMÉ

La violence contre les femmes est un phénomène de plus en plus fréquent dans le monde entier et a un impact énorme sur la vie des victimes, de leurs familles et de la société. Dans cette brève étude, nous nous concentrons sur l'utilisation à la fois de l'intelligence artificielle et de simples algorithmes prédictifs pour prévenir la récidive de cette violence en adoptant, le cas échéant, les mesures préventives nécessaires. L'étude commence par une brève exposition de l'origine et de l'évolution de l'intelligence artificielle et de son incidence progressive dans le domaine du droit ; ensuite, nous analysons les forces et les faiblesses de l'intelligence artificielle policière et judiciaire, et en troisième lieu, nous examinons en détail l'expérience espagnole importante d'application d'algorithmes prédictifs contre la violence de genre à travers le système VioGén.
Mots-clés : Violence de genre, intelligence artificielle, police prédictive, algorithmes prédictifs, système VioGén.

INTRODUCCIÓN: LA INTELIGENCIA ARTIFICIAL Y EL DERECHO

No siempre está claro de qué se habla cuando se habla de inteligencia artificial (IA en lo sucesivo): en la corta historia de esta disciplina se han proporcionado distintas definiciones que, en general, aluden al desarrollo de sistemas que imitan o reproducen el pensamiento y obrar humanos, actuando racionalmente - en el sentido de hacer lo "correcto" en función de su conocimiento - e interactuando con el medio. La IA pretende sintetizar o reproducir los procesos cognitivos humanos, tales como la percepción, la creatividad, la comprensión, el lenguaje o el aprendizaje (RUSSELL Y NORVIG, 2008, 1 y ss.). Para ello, utiliza todas las herramientas a su alcance, entre ellas las

proporcionadas por la computación, incluidos los algoritmos, aunque los sistemas de IA no usan cualquier algoritmo sino solo los que "aprenden" a base del procesamiento de datos[1].

Por otro lado, en ocasiones se habla de IA cuando en realidad estamos hablando de un subcampo, el aprendizaje automático (o *machine learning* en inglés, AA en lo sucesivo). El AA trata de encontrar patrones en datos para construir sistemas predictivos o explicativos; por tanto, puede considerarse una rama de la IA ya que a partir de la experiencia (los datos) toma decisiones o detecta patrones significativos y eso es una característica fundamental de la inteligencia humana. Es importante resaltar que para que un sistema de AA tenga éxito es tan necesario utilizar los algoritmos adecuados como realizar una correcta gestión y tratamiento de los datos utilizados para desarrollar el sistema.

Existe acuerdo en ubicar el nacimiento del nombre IA en un taller científico que, en el verano de 1956, reunió, entre otros, a John McCarthy, Marvin Minsky, Claude Shannon, Herbert Simon, Allan Nevell… en el Dartmouth College y en que esa denominación la propuso John McCarthy; también se coincide en que en esos primeros momentos cundió el optimismo sobre la IA y su impacto: Herbert Simon predijo que "en veinte años las máquinas serán capaces de hacer el trabajo de una persona" y Marvin Minsky declaró en 1970 a la revista *Life* que "dentro de tres a ocho años tendremos una máquina con la inteligencia general de un ser humano".

Como estas optimistas previsiones no se cumplieron, entre otras razones por la existencia de pocos datos y la escasa capacidad de la computación del momento, a principios de los años setenta se enfriaron las expectativas, que volvieron a coger auge y financiación durante los años ochenta pero que decayeron de nuevo en los noventa hasta que, en el presente siglo, el acceso a cantidades ingentes de datos — *Big Data* —, la disponibilidad de procesadores muy potentes a bajo coste y el desarrollo de redes neuronales profundas y complejas consolidaron de-

finitivamente la IA (OLIVER, 36 y ss.) y han despejado las dudas sobre su decisiva importancia en los próximos años, lo que, como es obvio, no quiere decir que todo lo que hoy se presume que puede alcanzar la IA llegue a conseguirse en las próximas décadas. Una vez más, no toda ficción llega a ser ciencia.

Puesto que ya hemos llegado a un punto avanzado de desarrollo científico y de aplicación práctica de la IA es imprescindible regularlos jurídicamente, tarea sobre la que viene llamando atención de manera especialmente intensa la Unión Europea, que, en teoría, está en estos momentos, mayo de 2023, en la fase final de aprobación de una "Ley de inteligencia artificial" en la que se define tal cosa como "el *software* que se desarrolla empleando una o varias de técnicas y estrategias que figuran en el Anexo I y que puede, para un conjunto determinado de objetivos definidos por seres humanos, generar información de salida como contenidos, predicciones, recomendaciones o decisiones que influyan en los entornos con los que interactúa" (artículo 3 de la Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión, de 21 de abril de 2021).

En la reciente Resolución del Parlamento Europeo, de 3 de mayo de 2022, sobre la inteligencia artificial en la era digital se recuerda que hay una diferencia significativa entre la IA simbólica, que constituye el principal enfoque de la IA entre los años cincuenta y los años noventa, y la IA basada en datos y aprendizaje automático, que domina desde el año 2000: durante la primera oleada, la IA se desarrolló codificando los conocimientos y la experiencia de los expertos en un conjunto de reglas que luego ejecutaba una máquina; en la segunda oleada, los procesos de aprendizaje automatizados de algoritmos basados en el procesamiento de grandes cantidades de datos, la capacidad de reunir datos procedentes de múltiples fuentes diferentes y de elaborar representaciones complejas de un entorno dado, y la determinación de patrones convirtieron a los sistemas de IA en sistemas más comple-

jos, autónomos y opacos, lo que puede hacer que los resultados sean menos explicables; en consecuencia, la IA actual puede clasificarse en muchos subcampos y técnicas diferentes.

Y siguiendo en el ámbito de la Unión Europea, en el primer párrafo del Libro Blanco sobre la inteligencia artificial de la Comisión, de 19 de febrero de 2020, se dice que "la IA se está desarrollando rápido. Cambiará nuestras vidas, pues mejorará la atención sanitaria (por ejemplo, incrementando la precisión de los diagnósticos y permitiendo una mejor prevención de las enfermedades), aumentará la eficiencia de la agricultura, contribuirá a la mitigación del cambio climático y a la correspondiente adaptación, mejorará la eficiencia de los sistemas de producción a través de un mantenimiento predictivo, aumentará la seguridad de los europeos y nos aportará otros muchos cambios que de momento solo podemos intuir. Al mismo tiempo, la IA conlleva una serie de riesgos potenciales, como la opacidad en la toma de decisiones, la discriminación de género o de otro tipo, la intromisión en nuestras vidas privadas o su uso con fines delictivos".

Así pues, la Comisión Europea asume algo de todo punto inevitable: que la IA va a cambiar – es seguro que ya lo está haciendo – nuestras vidas y, en consecuencia, esa transformación afectará, según el trabajo de investigación del Consejo de Europa sobre algoritmos y derechos humanos, a un gran número, sino a la práctica totalidad, de nuestros derechos fundamentales[2]; así, al derecho a la libertad personal y, muy relacionado con él, al derecho a un juicio justo y a la tutela de los tribunales; en segundo lugar, a los derechos de las personas en su dimensión más privada, como el derecho a la intimidad y a la protección de datos; en tercer lugar, a los derechos vinculados a la dimensión pública y relacional de las personas, como las libertades de expresión, información, creación artística e investigación pero también a las libertades de reunión y asociación, tanto en el plano meramente ciudadano como en lo que se refiere, por ejemplo, al ámbito laboral (libertad sindical, derecho de huelga); en cuarto lugar, y a su vez vinculado a muchos otros derechos, al de no sufrir discriminación por raza, género, edad, orientación

sexual…; en quinto lugar, a los derechos dependientes del acceso a los servicios públicos (educación, sanidad…) y, en general, a los derechos sociales (prestaciones por desempleo, enfermedad, jubilación…); finalmente, y por no extendernos mucho más, al derecho a intervenir en procesos participativos de índole política (elecciones, referendos, iniciativas legislativas populares…) y en, general, a las libertades en el ámbito ideológico (de pensamiento, conciencia y religión)[3].

1. FORTALEZAS Y DEBILIDADES DE LA INTELIGENCIA ARTIFICIAL POLICIAL Y JUDICIAL

Es bien conocido que los sistemas de IA ya se están aplicando en el ámbito de las investigaciones policiales para tratar de anticiparse a la comisión de posibles delitos y, en su caso, adoptar medidas preventivas limitativas de la libertad personal, bien sea atendiendo a criterios geográficos (*PredPol, CompStat*[4]…), sistemas muy frecuentes en Estados Unidos (FERGUSON, 2017), o a ciertas circunstancias personales, familiares…, como el español *VioGén*[5], sobre el que hablaremos más adelante. Y es que, como señala MIRÓ LLINARES (2019, 100), "hoy, y en parte gracias a las expectativas que parece dar la IA, la sociedad no espera sólo que la policía reaccione a los accidentes de tráfico, a los hurtos en los lugares turísticos o a los altercados y agresiones violentas relacionadas con manifestaciones deportivas o políticas, sino que no sucedan, que se intervenga incluso antes de que acontezcan… en parte esto se debe al *hype*, en el sentido de altísima esperanza, en lo que se denomina el *Predictive policing* que, a su vez, nace de la fusión entre las técnicas criminológicas del análisis delictivo, las herramientas actuariales de valoración del riesgo y la IA".

El problema surge cuando estos sistemas se apoyan en datos que pueden reflejar, de manera intencionada o no, sesgos en función de cómo se registran los delitos, qué delitos se seleccionan para ser incluidos en el análisis o qué herramientas analíticas se utilizan, pudiendo

generar una retroalimentación en la que, al menos en no pocas ciudades de Estados Unidos, la geografía -las zonas donde se concentra la vigilancia policial para prevenir delitos o reaccionar rápidamente ante ellos- puede operar, en palabras de O'NEIL (2018, 110), como "un valor sustitutivo altamente eficaz para la raza".

La Resolución del Parlamento Europeo, de 6 de octubre de 2021, sobre la inteligencia artificial en el Derecho penal y su utilización por las autoridades policiales y judiciales en asuntos penales (2020/2016(INI)), concluyó que los sesgos pueden ser inherentes a los conjuntos de datos subyacentes, especialmente cuando se emplean datos históricos, introducidos por los desarrolladores de los algoritmos o generados cuando los sistemas se aplican en entornos del mundo real y señaló que los resultados de las aplicaciones de inteligencia artificial dependen necesariamente de la calidad de los datos utilizados y que estos sesgos inherentes tienden a aumentar gradualmente y, por tanto, perpetúan y amplifican la discriminación existente, en particular con respecto a las personas pertenecientes a determinados grupos étnicos o comunidades racializadas.

Se destaca, igualmente, que las predicciones de IA basadas en las características de un grupo específico de personas acaban amplificando y reproduciendo formas de discriminación existentes; considera que deben hacerse grandes esfuerzos para evitar discriminaciones y prejuicios automatizados y pide que se establezcan salvaguardias adicionales sólidas en caso de que los sistemas de IA de las autoridades policiales y judiciales se utilicen en relación con menores (párrafos 8 y 9).

En segundo lugar, y muy relacionado con lo dicho, está el recurso a la IA en el ámbito de justicia -IA judicial- para, por ejemplo, apoyar la toma de decisiones sobre prisión provisional o libertad condicional. A este respecto, la citada Resolución del Parlamento Europeo considera (párrafos 3 y 4), habida cuenta del papel y la responsabilidad de las autoridades policiales y judiciales y del impacto de las decisiones que adoptan con fines de prevención, investigación, detección o enjuicia-

miento de infracciones penales o de ejecución de sanciones penales, que el uso de aplicaciones de IA debe clasificarse como de alto riesgo en los casos en que tienen potencial para afectar significativamente a la vida de las personas y que toda herramienta de IA desarrollada o utilizada por las autoridades policiales o judiciales debe, como mínimo, ser segura, robusta, fiable y apta para su finalidad, así como respetar los principios de minimización de datos, rendición de cuentas, transparencia, no discriminación y *explicabilidad*[6] y su desarrollo, despliegue y uso deben estar sujetos a una evaluación de riesgos y a una estricta comprobación de los criterios de necesidad y proporcionalidad, debiendo guardar proporción las salvaguardas con los riesgos identificados (ORTIZ DE ZÁRATE, 2022, 333). La confianza de los ciudadanos en el uso de la IA desarrollada y utilizada en la Unión está supeditada al pleno cumplimiento de estos criterios.

Esa Resolución insiste en que el enfoque adoptado en algunos países no pertenecientes a la Unión en relación con el desarrollo, el despliegue y el uso de tecnologías de vigilancia masiva interfiere de manera desproporcionada con los derechos fundamentales y, por lo tanto, no debe ser seguido por la Unión; destaca, por tanto, que también deben regularse de manera uniforme en toda la Unión las salvaguardias contra el uso indebido de las tecnologías de IA por parte de las autoridades policiales y judiciales, y subraya el impacto del uso de herramientas de IA en los derechos de defensa de los sospechosos, la dificultad para obtener información significativa sobre su funcionamiento y la consiguiente dificultad para impugnar sus resultados ante los tribunales, en particular por parte de las personas investigadas (párrafos 7 y 10).

En suma, en la Resolución se considera esencial, tanto para la eficacia del ejercicio del derecho de defensa como para la transparencia de los sistemas nacionales de justicia penal, que un marco jurídico específico, claro y preciso regule las condiciones, las modalidades y las consecuencias del uso de herramientas de IA en el ámbito de las actuaciones policiales y judiciales, así como los derechos de las personas afectadas

y procedimientos eficaces y fácilmente accesibles de reclamación y recurso, incluidos los recursos judiciales. Subraya, además, el derecho de las partes en un procedimiento penal a tener acceso al proceso de recopilación de datos y a las evaluaciones conexas realizadas u obtenidas mediante el uso de aplicaciones de IA; destaca la necesidad de que las autoridades de ejecución participantes en la cooperación judicial, al decidir sobre una solicitud de extradición (o entrega) a otro Estado miembro o a un tercer país, evalúen si el uso de herramientas de IA en el país solicitante podría manifiestamente comprometer el derecho fundamental a un juicio justo; pide a la Comisión que elabore directrices sobre cómo llevar a cabo dicha evaluación en el contexto de la cooperación judicial en materia penal; insiste en que los Estados miembros, de conformidad con la legislación aplicable, deben velar por la información de las personas que sean objeto de aplicaciones de IA utilizadas por parte de las autoridades policiales o judiciales (párrafo 14).

Por lo que respecta a las decisiones judiciales, la IA "ya está ahí" pero, sobre todo, va a estarlo de manera cada vez más relevante pues, no en vano, las posibilidades que se abren en este ámbito son verdaderamente enormes: en ejecución de deudas, en asuntos como la elección de recursos en los países cuyos tribunales supremos dispongan del llamado *certiorari*, y que es una selección de asuntos en función de criterios de relevancia de la decisión, fundamentalmente para la formación de jurisprudencia; en materia de admisión de las pruebas, sobre todo en el proceso civil, donde los asuntos muchas veces hacen previsible que las únicas relevantes sean la pericial y la documental (NIEVA FENOLL, 2018; 2021, 153-172; 2022, 53-68).

La cuestión esencial no es, por tanto, la presencia de la IA relacionada con el derecho de acceso a la justicia sino en cómo está articulada dicha presencia y, en particular, en qué aspectos de los procesos penales cabe acudir a ella para que no resulten menoscabados derechos como el de defensa y el de presunción de inocencia; en particular, de las personas más vulnerables.

A este respecto, y como ya se ha dicho, la Resolución del Parlamento Europeo, de 6 de octubre de 2021, recuerda que, en virtud del Derecho de la Unión, una persona tiene derecho a no ser objeto de una decisión que produzca efectos jurídicos que la conciernan o la afecte significativamente y que se base únicamente en el tratamiento automatizado de datos y pide a la Comisión que prohíba el uso de la IA y las tecnologías conexas para proponer decisiones judiciales y, como ya anticipamos, en dicha Resolución toda herramienta de IA desarrollada o utilizada por las autoridades policiales o judiciales debe, como mínimo, respetar los principios de rendición de cuentas, transparencia, no discriminación y *explicabilidad* (párrafo 4).

Y en la Propuesta de Reglamento por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial) se postula (p. 33) que se consideren de alto riesgo ciertos sistemas de IA destinados a la administración de justicia y los procesos democráticos, dado que pueden tener efectos potencialmente importantes para la democracia, el Estado de Derecho, las libertades individuales y el derecho a la tutela judicial efectiva y a un juez imparcial (COTINO Y OTROS, 2021). En particular, a fin de evitar el riesgo de posibles sesgos, errores y opacidades, procede considerar de alto riesgo aquellos sistemas de IA cuyo objetivo es ayudar a las autoridades judiciales a investigar e interpretar los hechos y el Derecho y a aplicar la ley a unos hechos concretos.

En suma, este ámbito, ni se trata de confiar todo a la IA algorítmica ni de rechazar radicalmente lo que puede aportar si bien aquí la *explicabilidad* resulta, si cabe, más irrenunciable, "tanto porque el sistema de justicia penal está basado en la argumentación y la justificación, como porque el constructo esencial configurador de responsabilidad en este ámbito es la peligrosidad que ello obliga a individualizar y no objetivar y generalizar factores y variables, por lo que resulta esencial que todos los algoritmos que aporten información de pronósticos para tomar decisiones que afecten a derechos se construyan como herramientas complementarias y

de apoyo, y eviten caer en el «cum hoc ergo propter hoc» y se acerquen muchos más a modelos explicativos y argumentativos a partir de inferencias causales" (MIRÓ LLINARES/CASTRO TOLEDO, 2022, 524).

2. LA APLICACIÓN DE ALGORITMOS PREDICTIVOS EN ESPAÑA FRENTE A LA VIOLENCIA DE GÉNERO: EL SISTEMA *VIOGÉN*

En España, y en el marco establecido en la Ley Orgánica 1/2004, de 28 de diciembre, de Medidas de Protección Integral contra la Violencia de Género[7], el Gobierno aprobó un conjunto de medidas urgentes para luchar contra esa violencia, entre las que cabe destacar la elaboración de un Protocolo de valoración de riesgo de la mujer víctima para su uso por parte de las Fuerzas y Cuerpos de Seguridad[8].

Como resultado, el Ministerio del Interior creó y puso en marcha en julio de 2007 el Sistema de Seguimiento Integral de los casos de Violencia de Género (Sistema *VioGén*), dotándolo de formularios informatizados para practicar y administrar las evaluaciones de riesgo de la mujer víctima, así como de las funcionalidades precisas para llevar a cabo el seguimiento de dichos casos y la implementación de las medidas de seguridad y protección policial acordes con los niveles de riesgo resultantes[9]. La última actualización se ha llevado a cabo a través de la "Instrucción número 4/2019, de la Secretaría de Estado de Seguridad, por la que se establece un nuevo protocolo para la valoración policial del nivel de riesgo de violencia de género, la gestión de la seguridad de las víctimas y seguimiento de los casos a través del sistema de seguimiento integral de los casos de violencia de género"[10].

De esta manera se da también cumplimiento al mandato del artículo 282 de la Ley de Enjuiciamiento Criminal, donde se dispone que "cuando las víctimas entren en contacto con la Policía Judicial, [ésta] cumplirá con los deberes de información que prevé la legislación vigente. Asimismo, llevarán a cabo una valoración de las circunstancias

particulares de las víctimas para determinar provisionalmente qué medidas de protección deben ser adoptadas para garantizarles una protección adecuada, sin perjuicio de la decisión final que corresponderá adoptar al Juez o Tribunal".

De forma general, el Sistema *VioGén* se dirige a:

a. Aglutinar a las diferentes instituciones públicas que tienen competencias en materia de violencia de género.
b. Integrar toda la información de interés que se considere necesaria, propiciando su intercambio ágil.
c. Facilitar la valoración del riesgo de que se produzca nueva violencia.
d. d) Atendiendo al nivel de riesgo, proporcionar el seguimiento y, si es preciso, la protección a las víctimas, en todo el territorio nacional.
e. Ayudar a la víctima a que elabore un "plan de seguridad personalizado", con medidas de autoprotección pertinentes y a su alcance.
f. Facilitar la labor preventiva, emitiendo avisos, alertas y alarmas, a través de un subsistema de notificaciones automatizadas, cuando se detecte alguna incidencia o acontecimiento que pueda poner en peligro la integridad de la víctima.

Pues bien, desde la entrada en funcionamiento del Sistema de Seguimiento Integral de los Casos de Violencia de Género en julio de 2007 y hasta finales de marzo de 2023 se han evaluado 726.064 casos de violencia de género y se ha proporcionado un plan de seguridad personalizado para 647.0014 mujeres o menores víctimas de violencia. Si una víctima lo es más de un agresor se computarán tantos casos como agresores. Del total de casos registrados había, el 31 de marzo de 2023, 76.404 activos, es decir, con seguimiento policial, y 649.670 inactivos. De los activos, 30.864 sin riesgo apreciado, 33.119 con riesgo bajo, 11.253 con riesgo medio, 1.147 con riesgo alto y 21 con riesgo extremo[11]. En conjunto estamos hablando del mayor sistema del mundo en ese ámbito[12].

Esta herramienta predictiva no es, en rigor, "inteligencia artificial", pues no usa algoritmos que "aprenden" (*machine learning*) a base del procesamiento de datos, sino que es "un sistema actuarial que utiliza modelos estadísticos para inferir el riesgo que puede correr una víctima (tanto de agresión como de homicidio) así como su evolución en base a un conjunto de indicadores que han sido determinados y posteriormente evaluados por un grupo de expertos"[13]. No obstante, podría considerarse un sistema de IA en un sentido "impropio" y no está descartada la incorporación de un algoritmo de autoaprendizaje.

En la actualidad funciona a través de dos formularios (Protocolo Dual): Valoración Policial del Riesgo (VPR) y Valoración Policial de la Evolución del Riesgo (VPER). El formulario VPR realiza la primera valoración del riesgo en el momento de la denuncia de la agresión a la policía, mientras que el formulario VPER realiza el seguimiento de la evolución del riesgo de violencia de género. Estos protocolos de valoración son revisados y corregidos por un equipo multidisciplinar de expertos. La quinta versión, la más actualizada, se publicó en marzo de 2019.

Explican GONZÁLEZ ÁLVAREZ, LÓPEZ OSSORIO y MUÑOZ RIVAS (2018, 55 y 56) que

> "el protocolo español es único a nivel internacional debido a que se encuentra implantado a nivel nacional, cuenta con dos formularios (uno para establecer el nivel de riesgo de partida y su aparejamiento con medidas de protección policial concretas para cada nivel de riesgo y otro para reevaluarlo conforme pasa el tiempo) y está desarrollado en un sistema informático "on line y multiagencia", al que se conectan miles de usuarios de forma simultánea.
> El empleo de dos formularios de valoración de riesgo distingue este procedimiento español de valoración del riesgo del resto de protocolos conocidos en el mundo, que solo utilizan uno…
> Este protocolo va más allá de la mera valoración del ries-

> go, puesto que conlleva la activación y puesta en práctica de una serie de medidas de protección policial, tasadas y proporcionadas a cada nivel de riesgo resultante. Es importante señalar que, en todos los casos, la estimación del riesgo no descansa en una mera máquina, sino que el Sistema permite que los agentes policiales, que son los que mejor conocen los casos por haberlos investigado en profundidad, puedan corregir el resultado automático del protocolo de valoración de riesgo cuando cuenten con información que así lo aconseje.
> De este modo, debe subrayarse que el Sistema es una herramienta desarrollada para facilitar el trabajo diario a los agentes, asumiendo la importancia que tiene la experiencia profesional, como en cualquier profesión. Así, al final de cada valoración policial de riesgo el Sistema VioGén resume las respuestas señaladas y pregunta por la conformidad del agente con el resultado automático (que suele ser muy alta, del orden del 95%), permitiendo que el usuario manifieste su desacuerdo y asigne el nivel de riesgo que él considera más apropiado, facilitando sus razones, permitiendo así el perfeccionamiento del Sistema".

De acuerdo con el protocolo de actuación, cuando una mujer presenta una denuncia se rellena el formulario VPR5.0-H, que se cumplimentará por los agentes policiales actuantes, nunca por la víctima ni otras personas implicadas y sólo cuando se haya recopilado información suficiente y contrastada de todas las fuentes disponibles, sobre el supuesto concreto. Según el protocolo, en ningún caso una víctima abandonará las dependencias policiales sin haber sido valorada ni se le hayan asignado las medidas policiales de protección que correspondan conforme al nivel de riesgo resultante.

Según el mismo protocolo, durante el proceso de valoración no se realizarán preguntas directas a la víctima, salvo en supuestos muy concretos y siempre que falte algún dato muy específico que sólo pueda recabarse por esta vía. En estos supuestos, se prestará especial cuidado en la formulación de las preguntas imprescindibles, todo ello a fin de evitar doble victimización en el momento de recabar información muy

sensible y personal de la víctima o su agresor y también para evitar sugerencias que conduzcan a desviaciones o sesgos en las respuestas.[14]

El formulario incluye 5 dominios con 35 indicadores de riesgo. Cada ítem se valora como "presente" y "no presente". De este modo, la recogida de información está estandarizada en todo el país.

1.-HISTORIA DE VIOLENCIA EN LA RELACIÓN DE PAREJA	Respuestas
Indicador 1: Violencia psicológica (vejaciones, insultos y humillaciones)	SI NO N/S
1.1 Intensidad de la violencia psicológica	Leve Grave Muy grave
Indicador 2: Violencia física	SI NO N/S
2.1 Intensidad de la violencia física	Leve Grave Muy grave
Indicador 3: Sexo forzado	SI NO N/S
3.1 Intensidad de la violencia sexual	Leve Grave Muy grave
Indicador 4: Empleo de armas u objetos contra la víctima	SI NO N/S
4.1 Arma blanca 4.2. Arma de fuego 4.3. Otros objetos	
Indicador 5: Existencia de amenazas o planes dirigidos a causar daño a la víctima	SI NO N/S
5.1 Intensidad de las amenazas	Leve Grave Muy grave
5.2 Amenazas de suicidio del agresor	SI NO
5.3 Amenazas de muerte del agresor dirigidas a la víctima	SI NO
Indicador 6: En los últimos seis meses se registra un aumento de la escalada de agresiones o amenazas	SI NO N/S
2.-CARACTERÍSTICAS DEL AGRESOR	
Indicador 7: En los últimos seis meses, el agresor muestra celos exagerados o sospechas de infidelidad	SI NO N/S
Indicador 8: En los últimos seis meses, el agresor muestra conductas de control	SI NO N/S
Indicador 9: En los últimos seis meses, el agresor muestra conductas de acoso	SI NO N/S
Indicador 10: Existencia problemas en la vida del agresor en los últimos seis meses	SI NO N/S
10.1 Problemas laborales o económicos	SI NO
10.2 Problemas con el sistema de justicia	SI NO
Indicador 11: En el último año el agresor produce daños materiales	SI NO N/S
Indicador 12: En el último año se registran faltas de respeto a la autoridad o a sus agentes	SI NO N/S
Indicador 13: En el último año agrede físicamente a terceras personas y/o animales	SI NO N/S
Indicador 14: En el último año existen amenazas o desprecios a terceras personas	SI NO N/S
Indicador 15: Existen antecedentes penales y/o policiales del agresor	
Indicador 16: Existen quebrantamientos previos o actuales (cautelares o penales)	
Indicador 17: Existen antecedentes de agresiones físicas y/o sexuales	SI NO N/S
Indicador 18: Existen antecedentes de violencia de género sobre otra/s pareja/s	
Indicador 19: Presenta problemas un trastorno mental y/o psiquiátrico	SI NO N/S
Indicador 20: Presenta ideas o intentos de suicidio	SI NO N/S
Indicador 21: Presenta algún tipo de adicción o conductas de abuso de tóxicos (alcohol, drogas y fármacos)	SI NO N/S
Indicador 22: Presenta antecedentes familiares de violencia de género o doméstica	SI NO N/S
Indicador 23: El agresor tiene menos de 24 años	SI NO N/S
3.-FACTORES DE RIESGO / VULNERABILIDAD DE LA VÍCTIMA	
Indicador 24: Existencia de algún tipo de discapacidad, enfermedad física o psíquica grave	SI NO N/S
Indicador 25: Víctima con ideas o intentos de suicidio	SI NO N/S
Indicador 26: Presenta algún tipo de adicción o conductas de abuso de tóxicos (alcohol, drogas y fármacos)	SI NO N/S
Indicador 27: Carece de apoyo familiar o social favorable	SI NO N/S
Indicador 28: Víctima extranjera	SI NO
4.-CIRCUNSTANCIAS RELACIONADAS CON LOS MENORES	
Indicador 29: La víctima tiene a su cargo menores de edad	SI NO N/S
Indicador 30: Existencia de amenazas a la integridad física de los menores	SI NO N/S
Indicador 31: La víctima teme por la integridad de los menores	SI NO N/S
5.-CIRCUNSTANCIAS AGRAVANTES	
Indicador 32: La víctima ha denunciado a otros agresores en el pasado	
Indicador 33: Se han registrado episodios de violencia lateral recíproca	SI NO N/S
Indicador 34: La víctima ha expresado al agresor su intención de romper la relación hace menos de seis meses	SI NO N/S
Indicador 35: La víctima piensa que el agresor es capaz de agredirla con mucha violencia o incluso matarla	SI NO N/S

En el formulario VPR5.0 se incluyen dos escalas con algoritmos diferentes: una para estimar los riesgos de reincidencia con cinco niveles (no apreciado, bajo, medio, alto y extremo), y otra para estimar el riesgo de feminicidio con dos niveles (en bajo y alto). Con el objetivo de facilitar a los miembros de las fuerzas y cuerpos de seguridad sus decisiones en materia de protección de las víctimas se optó por programar un mecanismo dual: cuando se recibe la denuncia de un caso, los policías cumplimentan la VPR. En este momento, sin mostrar todavía el resultado, el Sistema *VioGén* aplica el primer algoritmo y calcula el riesgo de reincidencia que presenta el caso en ese momento, e inmediatamente después, calcula el riesgo de feminicidio con el segundo algoritmo. En caso de que aparezca riesgo mortal, se ha dispuesto que se incremente en un nivel el riesgo de reincidencia, que es el que se muestra finalmente a los agentes, junto con una alerta de que el caso es de especial interés, para que se pueda adecuar la protección policial a las características del caso concreto. Además, esta alerta se deja reflejada en una diligencia en el atestado policial, que se envía al Juzgado y a la Fiscalía competentes, para conocimiento de la singularidad del caso, y por si estimaran pertinente que los implicados fueran evaluados cuanto antes por psicólogos o médicos forenses, quienes podrían profundizar más en las circunstancias del caso y proponer nuevas medidas protectoras[15]. Los agentes de policía sólo pueden modificar la puntuación a un nivel de riesgo más alto, y no al revés, es decir, no se puede bajar la puntuación de riesgo calculada por el algoritmo *VioGén*. El resultado se comunicará a la Autoridad Judicial y Fiscal, en forma de Informe automatizado que genera el propio Sistema.

Cada uno de los niveles de riesgo llevará aparejadas medidas policiales para la protección y seguridad de las víctimas, que serán de aplicación obligatoria e inmediata. Así, por ejemplo, si el riesgo es "alto" y en caso de no haberse podido localizar todavía al agresor, se insistirá a la víctima, para su más efectiva protección, en la posibilidad de traslado a centro de acogida, casa de un familiar o domicilio distinto y se llevará

a cabo un control frecuente y aleatorio en el domicilio y lugar de trabajo de la víctima y, si procede en centros escolares de los hijos a la entrada y salida y contactos con personas de su entorno para mejor protección; respecto del agresor, se hará un control aleatorio de sus movimientos y contactos esporádicos con personas que frecuente o de su entorno; si el riesgo se califica como "extremo" se dará protección permanente de la víctima hasta que el mismo agresor o sus circunstancias dejen de ser una amenaza inminente y, si es procedente, se hará vigilancia en centros escolares de los hijos de la víctima a la hora de entrada y salida; respecto al agresor se hará un control intensivo de sus movimientos hasta que este deje de ser una amenaza inminente para la seguridad de la víctima. Esas medidas se adaptarán a las circunstancias concretas del caso, de manera que sean de aplicación personalizada e individual y se comunicarán a la víctima. Si tras la primera actuación judicial se acordara alguna Medida de alejamiento/Orden de protección, esta será comunicada expresamente tanto a víctima como a agresor. Dicha comunicación incluirá la explicación concreta de la medida acordada, alcance y consecuencias de su quebrantamiento[16].

Una vez explicado de manera sucinta el funcionamiento del sistema *Viogén* es importante mencionar también alguna de las críticas y objeciones que se han formulado al mismo; así, en primer lugar, y tras un análisis de esta herramienta LÓPEZ-OSSORIO, GONZÁLEZ ÁLVAREZ y ANDRÉS PUEYO (2016, 6) explicaban en 2016 que "la sensibilidad o identificación correcta del riesgo de violencia cuando existe la reincidencia fue del 85%, y la capacidad del instrumento para descartar el riesgo cuando no se dio reincidencia o especificidad fue del 53,7%"[17]. El valor predictivo negativo sería del 98.5% y el positivo del 8,6%. Con los ajustes de 2019 se estima una sensibilidad del 81% y una especificidad del 61%.

Es preciso aclarar, utilizando las palabras de MARTÍNEZ GARAY (2014, 28), que la sensibilidad es la capacidad de un instrumento de predicción para detectar a las personas que sí reincidirán; la especificidad es un valor complementario al anterior: es la capacidad del

instrumento para detectar correctamente a los que no reincidirán. Ambas categorías son complementarias en el sentido de que, cuanto mayor es una de ellas, generalmente menor es la otra: cuanto más amplios sean los criterios para clasificar a una persona como peligrosa, mayor será la sensibilidad (i.e., menos peligrosos se «escaparán» del diagnóstico), pero menor será la especificidad, porque aumenta la probabilidad de incluir como peligrosas a personas que en realidad no lo son ("falsos positivos"). Y al contrario, si son muy estrictos los criterios para clasificar a alguien como peligroso tendremos menos fallos de este segundo tipo (pocos no-peligrosos serán erróneamente considerados peligrosos), pero habrá personas que sí iban a delinquir en el futuro que se nos habrán quedado fuera del diagnóstico ("falsos negativos"). Pues bien, si las cosas son así resulta que el sistema *Viogén* fallaría bastante en especificidad, pues, con arreglo al estudio de 2015 casi la mitad de las personas (46,3%) habrían sido diagnosticadas incorrectamente como peligrosas ("falsos positivos"), cantidad que bajaría en 2019 al 39%.

El valor predictivo negativo de *Viogén* - los casos en los que no se advirtió riesgo y, efectivamente, no hubo agresiones- sería muy alto (más del 98%) y muy bajo el positivo -supuestos en los que se pronosticó agresiones y las hubo- (solo el 8,7), lo que parece lógico pues, como recuerdan MARTÍNEZ GARAY y GARCÍA ORTIZ (2022, 168), "las estimaciones de riesgo de reincidencia se diferencian de las realizadas en otros contextos en que la ocurrencia del evento estimado no es independiente del resultado de la valoración. Si predecimos que hará buen tiempo el fin de semana, puede que ocurra o no, pero nuestra predicción no habrá influido en ello. Sin embargo, cuando se estima el riesgo de violencia, se toman medidas como consecuencia de esas valoraciones (imponer o no una medida cautelar, etc.), medidas que influyen sobre la propia situación valorada. Así, si como consecuencia de una valoración de riesgo alto se adoptan medidas para minimizarlo y estas son eficaces, el evento, contrariamente a lo esperado, no se producirá. Podrá parecer que la estimación fue "equivocada", y no necesariamente es así".

Y volviendo a la muy relevante cuestión de los falsos positivos y negativos, VUKOVIĆ ET AL. (2021, 520) destacan que el impacto varía según el propósito del sistema de predicción; por ejemplo, cuando se trata de errores de predicción relacionados con el terrorismo, los falsos negativos pueden ser más costosos ya que pueden conducir a ataques y muertes que podrían haberse evitado en comparación con falsos positivos, pero, añadimos nosotros, no es trivial lo que está en juego si se producen falsos positivos, especialmente cuando implican restricciones importantes en los derechos fundamentales de las personas afectadas. A este respecto, MARTÍNEZ GARAY y GARCÍA ORTIZ (2022, 165) señalan que "éste no es un problema que resuelvan la estadística ni los algoritmos, porque es una cuestión político criminal, que presupone una decisión sobre qué es preferible: ¿restringir la libertad de muchas personas que en realidad no hubieran delinquido después, o renunciar al control penal sobre personas que van a seguir cometiendo delitos? Cuando se programa un algoritmo para ayudar a hacer predicciones, una persona física ha tomado esta decisión y ha decidido situar los umbrales de discriminación en unos puntos concretos. Y debería estar en condiciones de defender esa decisión ante los afectados por ese algoritmo y ante la opinión pública. En este punto la transparencia aparece [como veremos más adelante] como una cuestión fundamental".

En segundo lugar, y como explican GONZÁLEZ ÁLVAREZ, SANTOS HERMOSO y CAMACHO COLLADOS (2020, 34), es importante destacar que "las técnicas de policía predictiva se basan en el análisis de datos históricos, es decir, casos que llegan a conocimiento de los cuerpos policiales. Es por esto que los algoritmos que se generen serán específicos para esos casos, y permitirán predecir casos que muestren características similares o sigan un mismo patrón. El problema…es que tanto en la violencia de género, como en la violencia doméstica en general, muchos casos no llegan a denunciarse, y en consecuencia no forman parte de los registros policiales históricos. Esto plantea otra reflexión importante, y es que, con estos casos no denunciados pueden suceder dos cosas: 1) que sean similares a los casos que sí denuncian, por lo que

las herramientas podrían ser aplicables; o 2) puede que tengan una serie de características distintivas que, en parte, expliquen el por qué no se denuncia, y las herramientas de predicción no sirvan".

En las siguientes columnas se puede ver cómo, en los últimos 20 años, en el total de mujeres asesinadas víctimas de violencia de género predominan los casos en los que no había una denuncia previa y ese predominio es en una muy alta proporción:

Año	Denuncia agresor	Número de mujeres víctimas mortales
Año 2003	**No consta denuncia**	71
Año 2004	**No consta denuncia**	72
Año 2005	**No consta denuncia**	57
Año 2006	**No había denuncia**	47
Año 2006	Había denuncia	22
Año 2007	**No había denuncia**	50
Año 2007	Había denuncia	21
Año 2008	**No había denuncia**	58
Año 2008	Había denuncia	18
Año 2009	**No había denuncia**	43
Año 2009	Había denuncia	14
Año 2010	**No había denuncia**	51
Año 2010	Había denuncia	22
Año 2011	**No había denuncia**	47
Año 2011	Había denuncia	15
Año 2012	**No había denuncia**	41
Año 2012	Había denuncia	10
Año 2013	**No había denuncia**	43
Año 2013	Había denuncia	11
Año 2014	**No había denuncia**	38
Año 2014	Había denuncia	17

Año 2015	**No había denuncia**	47
Año 2015	Había denuncia	13
Año 2016	**No había denuncia**	32
Año 2016	Había denuncia	16
Año 2016	**No consta denuncia**	1
Año 2017	**No había denuncia**	38
Año 2017	Había denuncia	12
Año 2018	**No había denuncia**	38
Año 2018	Había denuncia	15
Año 2019	**No había denuncia**	45
Año 2019	Había denuncia	11
Año 2020	**No había denuncia**	42
Año 2020	Había denuncia	8
Año 2021	**No había denuncia**	39
Año 2021	Había denuncia	10
Año 2022	**No había denuncia**	29
Año 2022	Había denuncia	20
Año 2023	**No había denuncia**	10
Año 2023	Había denuncia	3

Elaboración propia a partir de los datos de http://estadisticasviolenciagenero.igualdad.mpr.gob.es/ (a 2 de mayo de 2023). La información sobre denuncias al agresor de la mujer víctima mortal por violencia de género se empieza a recoger en 2006, por lo que no consta ese dato para años anteriores.

Y esa ausencia de denuncia previa puede atribuirse, en buena medida, a la violencia que el agresor ejerce sobre la mujer y al miedo que tal situación genera; en palabras del Tribunal Supremo (sentencia nº 247, de 24 de mayo de 2018, fundamento jurídico segundo[18]):

> *"... El maltrato habitual produce un daño constante y continuado del que la víctima, o víctimas tienen la percepción de que no pueden salir de él y del acoso de quien perpetra estos actos, con la circunstancia agravante en cuanto al autor, de que éste es, nada*

> *menos, que la pareja de la víctima, lo que provoca situaciones de miedo, incluso, y una sensación de no poder denunciar.* Ello provoca que en situaciones como la presente el silencio haya sido prolongado en el tiempo hasta llegar a un punto en el que, ocurrido un hecho grave, *se decide, finalmente, a denunciar por haber llegado a un límite a partir del que la víctima ya no puede aguantar más actos de maltrato hacia ella y, en ocasiones, también, hacia sus hijos*. Sin embargo, es preciso señalar y destacar en el caso que ahora nos ocupa que *cuando esta decisión se adopta por la víctima se incrementa el riesgo de que los actos de maltrato pasen a un escenario de "incremento grave del riesgo de la vida de la víctima*", ya que si ésta decide comunicar la necesidad de una ruptura de la relación, como aquí ha ocurrido, o le denuncia por esos hechos, o el más reciente, el sentimiento de no querer aceptar esa ruptura el autor de los mismos provoca que pueda llegar a cometer un acto de mayor gravedad, como aquí ha ocurrido. Y ello requiere en estos casos medidas de detección urgente del riesgo de que estos hechos puedan ocurrir cuando se denuncian hechos de maltrato…"

En tercer lugar, este sistema parece generar un sesgo de "autoridad tecnológica"[19] o de "automatización" si, como se ha venido diciendo , hasta en el 95% de los casos los agentes mantienen la puntuación de riesgo asignada automáticamente por el algoritmo a pesar de que, como ya se ha dicho, pueden estimar que existe en riesgo superior al que predice *Viogén* y si tal cosa no se hace por una confianza casi automática en las predicciones del sistema se estará incumpliendo el propio protocolo lo que, en su caso, podría dar lugar a la atribución de diferentes tipos de responsabilidad.

A este respecto, la sentencia de la Audiencia Nacional 2350/2020, de 30 de septiembre, recuerda algo que tendría que ser obvio cuando se trata de valoraciones que llevan a cabo personas expertas:

> "siendo las relaciones interpersonales y la realidad cambiantes por definición, el Protocolo para la valoración

> policial del nivel de riesgo de violencia sobre la mujer prevé un sistema dinámico que permite una modificación de la valoración del riesgo, para lo que es imprescindible que la autoridad policial realice un seguimiento, serio y riguroso, de las distintas circunstancias generadas en cada caso y su evolución… la respuesta policial a la violencia contra la mujer exige que el sistema pueda prevenir la violencia y reevaluar el riesgo, esto es, más allá de la recogida de datos automatizados, la predicción y la prevención son la finalidad primordial del sistema de evaluación que exige agentes especializados en su tratamiento y sensibilización en su seguimiento" (FJ.3).

Estas exigencias se insertan en lo dicho por el Tribunal Supremo en la sentencia 371/2018, 19 de julio:

> "… Este tipo de casos evidencian la necesidad de llevar a cabo un esfuerzo en la valoración de la presencia de incremento del riesgo en las víctimas con una especial atención en su detección en las denuncias que presentan las víctimas, y que se debe acompañar en la denuncia policial al estudio que al efecto se elabore, así como en los institutos de medicina legal en la valoración forense, como consta en el Protocolo médico-forense de valoración urgente del riesgo de violencia de género del Ministerio de Justicia, donde se marcan las pautas de la detección del riesgo. Ello supone actuar desde el campo de la prevención en la evitación de la reiteración de estos hechos, y alertando a la víctima del riesgo concurrente, así como pudiendo articularse instrumentos de ayuda social y económica a las víctimas de malos tratos que así puedan entrar en ese arco de víctimas en situación de riesgo, pudiendo individualizarse las situaciones en aras a evitar la agravación de conductas que acaben con el crimen de género… tanto las Administraciones, para adoptar las medidas conducentes a dar protección a las víctimas, como estas mismas para darles información y asesoramiento sobre el riesgo de una posible decisión de reanudar la convivencia, son piezas y factores claves para potenciar la protección de

> las víctimas en la adopción de medidas preventivas que eviten desenlaces mortales incidiendo en la detección y valoración del riesgo…" (FJ 3).

Por otra parte, puede existir el riesgo de que, ante el temor a incurrir en algún tipo de responsabilidad se tienda a elevar de forma casi automática el nivel de riesgo que ha pronosticado el sistema *Viogén*.

Finalmente, hay que mencionar la falta de transparencia del sistema Viogén: aunque, como hemos visto, es público y bien conocido el formulario que incluye 5 dominios con 35 indicadores de riesgo no lo es cómo se combinan, qué relevancia tiene cada uno en el resultado final… En suma, "no se puede acceder a ningún dato o información más allá de lo producido por los expertos que participaron en la definición del sistema. Ni los auditores externos ni los grupos de mujeres tienen ningún tipo de acceso. El sistema no ha sido evaluado ni auditado de forma independiente y tampoco involucra a las destinatarias del mismo, que nunca han sido consultadas sobre el sistema, ni en su fase de diseño ni posteriormente durante las diferentes decisiones sobre cómo modificarlo"[20].

A este respecto, es imprescindible recordar que la Resolución del Parlamento Europeo, de 6 de octubre de 2021, sobre la inteligencia artificial en el Derecho penal y su utilización por las autoridades policiales y judiciales en asuntos penales (2020/2016(INI))[21], considera (párrafos 3 y 4), habida cuenta del papel y la responsabilidad de las autoridades policiales y judiciales y del impacto de las decisiones que adoptan con fines de prevención, investigación, detección o enjuiciamiento de infracciones penales o de ejecución de sanciones penales, que el uso de aplicaciones de IA debe clasificarse como de alto riesgo en los casos en que tienen potencial para afectar significativamente a la vida de las personas y que toda herramienta de IA desarrollada o utilizada por las autoridades policiales o judiciales debe, como mínimo, ser segura, robusta, fiable y apta para su finalidad, así como respetar los principios de minimización de datos, *rendición de cuentas, transparencia*, no discriminación y *explicabilidad*[22] y su

desarrollo, despliegue y uso deben estar sujetos a una evaluación de riesgos y a una estricta comprobación de los criterios de necesidad y proporcionalidad.

Para concluir, y asumiendo como propias las palabras de MARTÍNEZ GARAY y GARCÍA ORTIZ (2022, 172), la transparencia "es imprescindible para garantizar los derechos a la defensa y a la contradicción, y para poder detectar, discutir y en su caso corregir posibles sesgos o efectos discriminatorios. Pero también es un presupuesto básico para aceptar que son herramientas científicamente rigurosas, cuyo funcionamiento puede ser contrastado por terceros. Y en segundo lugar, los algoritmos deben utilizarse como apoyo para la toma de decisiones pero sin resultar vinculantes, de manera que el operador jurídico pueda tener en cuenta, además del resultado de la estimación automatizada, otros factores que le parezcan relevantes para poder dar una respuesta individualizada a cada caso concreto".

CONCLUSIONES

La inteligencia artificial e, incluso, los algoritmos predictivos que no están dotados de autoaprendizaje, pueden ser útiles para prevenir la reiteración de la violencia de género y ello como herramientas al servicio de la policía predictiva articuladas a partir de la fusión entre las técnicas criminológicas del análisis delictivo, las herramientas actuariales de valoración del riesgo y, en su caso, la propia inteligencia artificial.

A este respecto, los estudios empíricos han mostrado evidencias del elevado valor predictivo de sistemas como *Viogén* cuando diagnostican ausencia de riesgo de reiteración de la violencia de género y también su escaso valor cuando pronostican dicha reiteración pero, como ya se ha dicho, es muy importante tener en cuenta que en este último supuesto eso no implica que el sistema haya incumplido mínimamente

su labor sino que bien puede haber sido la adopción de medidas cautelares derivadas de la prognosis de un alto riesgo lo que ha servido, precisamente, para evitar que la agresión se produzca.

También se ha insistido en que ningún sistema predictivo, por muy sofisticado que sea, es inmune al fallo e, inevitablemente, generará más o menos casos de "falsos positivos" y "falsos negativos", lo que exige una reflexión previa de carácter político criminal para decidir si se opta por maximizar la prevención de delitos o por priorizar las libertades y derechos fundamentales de quienes puedan cometerlos. En el caso de la violencia de género podemos encontrar argumentos iusfundamentales que avalen la adopción de medidas cautelares en posible detrimento de los derechos del presunto agresor pues se tratará, en principio, de predicciones hechas para tener validez durante un período de tiempo no muy largo, para proteger derechos de extraordinaria relevancia como la vida y la integridad física y moral y que no necesariamente llevarán aparejada la privación de libertad del denunciado.

Pero, precisamente para dotar de la máxima legitimidad a las decisiones anteriormente mencionadas, es imprescindible que estas herramientas sean lo más seguras y fiables que se pueda, deben haberse supervisado con minuciosidad antes de su entrada en funcionamiento, deben incluir mecanismos de rendición de cuentas, ser transparentes y explicables en la mayor medida posible y, finalmente, deben estar sujetas a una estricta comprobación de los criterios de necesidad y proporcionalidad.

REFERENCIAS

COTINO, L. Y OTROS. "Un análisis crítico constructivo de la Propuesta de Reglamento de la Unión Europea por el que se establecen normas armonizadas sobre la Inteligencia Artificial (Artificial Intelligence Act)", **Diario La Ley**, 2 de julio, 2021.

GONZÁLEZ ÁLVAREZ, J. L; LÓPEZ OSSORIO, J. J. y MUÑOZ RIVAS, M. **La valoración policial del riesgo de violencia contra la mujer pareja en España — Sistema VioGén**, Madrid: Ministerio del Interior. Gobierno de España, 2018.

GONZÁLEZ ÁLVAREZ, J. L., SANTOS HERMOSO, J. y CAMACHO COLLADOS, M. "Policía predictiva en España aplicación y retos futuros", **Behavior & Law Journal**, vol. 6, nº 1, 2020, págs. 26-41.

LÓPEZ-OSSORIO, J. J., GONZÁLEZ ÁLVAREZ, J. L. y ANDRÉS PUEYO, A. Eficacia predictiva de la valoración policial del riesgo de la violencia de género. **Psychosocial Intervention**, vol.25, n.1, 2016. pp.1-7.

MARTÍNEZ GARAY, L. La incertidumbre de los pronósticos de peligrosidad: consecuencias para la dogmática de las medidas de seguridad. Indret: **Revista para el Análisis del Derecho**, nº 2, 2014.

MARTÍNEZ GARAY, L., GARCÍA ORTIZ, A. Paradojas de los algoritmos predictivos utilizados en el sistema de justicia penal. **El Cronista del Estado Social y Democrático de Derecho**, nº 100, 2022 (Ejemplar dedicado a: Inteligencia artificial y derecho), pp. 160-173.

MIRÓ LLINARES, F. "El modelo policial que viene: Mitos y realidades del impacto de la inteligencia artificial y la ciencia de datos en la prevención policial del crimen", en MARTÍNEZ ESPASA, J. (coord.) **Libro blanco de la prevención y seguridad local valenciana**: Conclusiones y propuestas del Congreso Valenciano de Seguridad Local: la prevención del siglo XXI, pp. 98-113, 2019.

MIRÓ LLINARES, F./CASTRO TOLEDO, F. J. "¿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»", en DEMETRIO CRESPO, D. (dir) **Derecho penal y comportamiento humano**. Avances desde la neurociencia y la inteligencia artificial, Valencia: Tirant lo Blanch, 2022, pp. 507-529.

NIEVA FENOLL, J. **Inteligencia artificial y proceso judicial**. Madrid: Marcial Pons, 2018.

NIEVA FENOLL, J. Un cambio generacional en el proceso judicial: la inteligencia artificial", en GUERRA MORENO, D. (coord.) **Constitución y justicia digital**. Bogotá: Grupo Editorial Ibáñez/Universidad Libre, 2021, pp. 153-172.

NIEVA FENOLL, J. "Technology and fundamental rights in the judicial process", **Civil Procedure Review**, v. 13, núm. 1, 2022, pp. 53-68.

OLIVER, N. Inteligencia artificial, naturalmente. **Observatorio Nacional de Tecnología y Sociedad**, 2020. https://www.ontsi.es/es/publicaciones/Inteligencia-artificial%2C-naturalmente

O'NEIL, C. **Armas de destrucción matemática**. Cómo el big data aumenta la desigualdad y amenaza la democracia. Madrid: Capitán Swing, 2018.

ORTIZ DE ZÁRATE ALCARAZO, L. "Explicabilidad (de la inteligencia artificial)", Eunomía. **Revista en Cultura de la Legalidad**, núm. 22, 2022, pp. 3-28.

PRESNO LINERA, M. A. **Derechos fundamentales e inteligência artificial**. Madrid: Marcial Pons, 2022.

RUSSELL S./NORVIG, P. **Inteligencia Artificial**: un enfoque moderno. Madrid: Pearson Education, 2008.

VUKOVIĆ ET AL. "Challenges of contemporary predictive policing". Thematic conference proceedings of international significance. **International Scientific Conference "Archibald Reiss Days"**, Belgrado: University of Criminal Investigation and Police Studies, 2021.

NOTAS

1 En el glosario que incorpora la *Carta Ética Europea sobre el uso de la Inteligencia Artificial en los sistemas judiciales y su entorno*, de 4 de diciembre de 2018, se define el algoritmo como una "secuencia finita de reglas formales (operaciones lógicas e instrucciones) que permiten obtener un resultado de la entrada inicial de información. Esta secuencia puede ser parte de un proceso de ejecución automatizado y aprovechar modelos diseñados a través del aprendizaje automático".

2 *Algorithms and Human Rights. Study on the human rights dimensions of automated data processing techniques and possible regulatory implications*, Published by the Council of Europe, 2018, disponible en https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5 (a 2 de mayo de 2023).

3 Puede verse, más ampliamente, PRESNO LINERA, M. A., *Derechos fundamentales e inteligencia artificial*. Marcial Pons, Madrid, 2022.

4 <https://www.predpol.com/> y <https://www1.nyc.gov/site/nypd/stats/crime-statistics/compstat.page> (a 2 de mayo de 2023).

5 <https://www.interior.gob.es/opencms/es/servicios-al-ciudadano/violencia-contra-la-mujer/> (a 2 de mayo de 2023).

6 Según la Recomendación sobre la Ética de la Inteligencia Artificial de la UNESCO, aprobada en la reunión del 9 al 24 de noviembre de 2021, "la explicabilidad supone hacer inteligibles los resultados de los sistemas de IA y facilitar información sobre ellos. La explicabilidad de los sistemas de IA también se refiere a la inteligibilidad de la entrada, salida y funcionamiento de cada componente algorítmico y la forma en que contribuye a

los resultados de los sistemas. Así pues, la explicabilidad está estrechamente relacionada con la transparencia, ya que los resultados y los subprocesos que conducen a ellos deberían aspirar a ser comprensibles y trazables, apropiados al contexto. Los actores de la IA deberían comprometerse a velar por que los algoritmos desarrollados sean explicables. En el caso de las aplicaciones de IA cuyo impacto en el usuario final no es temporal, fácilmente reversible o de bajo riesgo, debería garantizarse que se proporcione una explicación satisfactoria con toda decisión que haya dado lugar a la acción tomada, a fin de que el resultado se considere transparente".

Conforme a las *Directrices éticas para una IA fiable* del Grupo de expertos de alto nivel sobre inteligencia artificial de la Unión Europea, "la explicabilidad es crucial para conseguir que los usuarios confíen en los sistemas de IA y para mantener dicha confianza. Esto significa que los procesos han de ser transparentes, que es preciso comunicar abiertamente las capacidades y la finalidad de los sistemas de IA y que las decisiones deben poder explicarse –en la medida de lo posible– a las partes que se vean afectadas por ellas de manera directa o indirecta. Sin esta información, no es posible impugnar adecuadamente una decisión. No siempre resulta posible explicar por qué un modelo ha generado un resultado o una decisión particular (ni qué combinación de factores contribuyeron a ello). Esos casos, que se denominan algoritmos de «caja negra», requieren especial atención. En tales circunstancias, puede ser necesario adoptar otras medidas relacionadas con la explicabilidad (por ejemplo, la trazabilidad, la auditabilidad y la comunicación transparente sobre las prestaciones del sistema), siempre y cuando el sistema en su conjunto respete los derechos fundamentales. El grado de necesidad de explicabilidad depende en gran medida del contexto y la gravedad de las consecuencias derivadas de un resultado erróneo o inadecuado".

7 El artículo 31 de la Ley Orgánica dispone que "1. El Gobierno establecerá, en las Fuerzas y Cuerpos de Seguridad del Estado, unidades especializadas en la prevención de la violencia de género y en el control de la ejecución de las medidas judiciales adoptadas…"; el artículo 32 que "1. Los poderes públicos elaborarán planes de colaboración que garanticen la ordenación de sus actuaciones en la prevención, asistencia y persecución de los actos de violencia de género, que deberán implicar a las administraciones sanitarias, la Administración de Justicia, las Fuerzas y Cuerpos de Seguridad y los servicios sociales y organismos de igualdad. 2. En desarrollo de dichos planes, se articularán protocolos de actuación que determinen los procedimientos que aseguren una actuación global e integral de las distintas administraciones y servicios implicados, y que garanticen la actividad probatoria en los procesos que se sigan… 4. En las actuaciones previstas en este artículo se considerará de forma especial la situación de las mujeres que, por sus circunstancias personales y sociales puedan tener mayor riesgo de sufrir la violencia de género o mayores dificultades para acceder a los servicios previstos en esta ley, tales como las pertenecientes a minorías, las inmigrantes, las que se encuentran en situación de exclusión social, las mujeres con discapacidad, las mujeres mayores o aquellas que viven en el ámbito rural".

8 GONZÁLEZ ÁLVAREZ, LÓPEZ OSSORIO y MUÑOZ RIVAS sostienen que "en el momento actual, la toma de decisiones que realizan los profesionales para la predicción del riesgo de violencia sigue alguna de las tres estrategias tecnológicas básicas: las Clínicas, las Actuariales y las de Juicio Profesional Estructurado. Pese a existir estudios contradictorios, la estrategia que ha mostrado una mayor efectividad y utilidad es la del Juicio Profesional Estructurado, que consiste, básicamente, en una estrategia mixta que se basa en la utilización de guías de evaluación del riesgo que contienen un protocolo de valoración del riesgo construido atendiendo al fenómeno violento específico se va a anticipar, los factores de riesgo, factores de protección propios de ese tipo de violencia y otros aspectos

técnicos propios de esta tecnología. Para facilitar el uso de las técnicas de Juicio Profesional Estructurado se utilizan unas "guías" de valoración del riesgo que están adecuadas a los diversos tipos de violencia (sexual, de género, física, etc.) y que han sido diseñadas para predecir un resultado concreto (un tipo de violencia determinado) y tienen validez en un período temporal delimitado. También, para mejorar la adecuación de estas guías, se contemplan los factores de riesgo particulares de una población determinada y para un contexto sociocultural específico…", *La valoración policial del riesgo de violencia contra la mujer pareja en España — Sistema VioGén*, Ministerio del Interior. Gobierno de España, Madrid, 2018, p. 38; disponible (a 2 de mayo de 2023) en <https://www.interior.gob.es/opencms/pdf/archivos-y-documentacion/documentacion-y-publicaciones/publicaciones-descargables/seguridad-ciudadana/La_valoracion_policial_riesgo_violencia_contra_mujer_pareja_126180887.pdf> (a 2 de mayo de 2023).

9 Puede leerse un resumen sobre las características de otros instrumentos de valoración del riesgo de violencia contra la mujer en el ámbito internacional en el trabajo de GONZÁLEZ ÁLVAREZ, LÓPEZ OSSORIO y MUÑOZ RIVAS. *La valoración policial del riesgo de violencia contra la mujer pareja en España — Sistema VioGén...* pp. 32-35.

10 <https://violenciagenero.igualdad.gob.es/profesionalesInvestigacion/seguridad/protocolos/pdf/PROTOCOLO_CERO.pdf> (a 2 de mayo de 2023).

11 <https://www.interior.gob.es/opencms/pdf/servicios-al-ciudadano/violencia-contra-la-mujer/estadisticas/2023/Estadistica-31-de-marzo-2023.pdf> (a 2 de mayo de 2023).

12 *Auditoría externa del sistema Viogén*, Fundación Éticas, 2022, p. 12. https://eticas-foundation.org/es/la-fundacion-eticas-realiza-una-auditoria-externa-e-independiente-del-sistema-viogen/ (a 2 de mayo de 2023).

13 *Auditoría externa del sistema Viogén,…*, p. 10.

14 Puede consultarse la *Guía de aplicación del formulario VFR5.0-H en la valoración forense del riesgo* en <https://docplayer.es/204322210-Guia-de-aplicacion-del-formulario-vfr-5-0-h-en-la-valoracion-forense-del-riesgo.html> (a 2 de mayo de 2023).

15 GONZÁLEZ ÁLVAREZ, J. L./SANTOS HERMOSO, J./CAMACHO COLLADOS, M. (2020): "Policía predictiva en España. Aplicación y retos de futuro", *Behavior & Law Journal*, 6(1); en este trabajo se señala que el hallazgo de que no todos los indicadores de la VPR que son útiles para la predicción de reincidencia, lo son para predecir futuros episodios mortales, y el hecho de que, además, fuera necesario recalcular los pesos de los 13 indicadores significativos, es indicativo… de que violencia mortal y no mortal pueden ser fenómenos diferentes, aunque ambas se den en el marco de la violencia de género. Desde el punto de vista de la aplicación de la Inteligencia Artificial (IA) a la predicción del crimen, se plantea una importante reflexión, y es ¿qué nivel de análisis es necesario entonces para desarrollar algoritmos que realmente sean útiles?..." (p. 34). Esos 13 indicadores cuyo peso fue recalculado fueron los siguientes:

- Violencia física grave o muy grave
- Violencia sexual grave o muy grave
- Uso de armas (excepto armas de fuego)
- Amenazas de muerte por parte del agresor
- Acumulación de amenazas o agresiones durante los últimos seis meses
- Signos de celos extremos por parte del agresor en los últimos seis meses
- Comportamientos de acoso por parte del agresor en los últimos seis meses
- Agresiones a otras personas o animales por parte del agresor durante el último año
- Trastorno mental o psiquiátrico en el agresor
- Presencia de ideas o intentos de suicidio por parte del agresor

• Adicción o abuso de sustancias (alcohol, drogas o medicamentos) por parte del agresor
• La víctima manifestó su intención de terminar la relación en los últimos seis meses
• La víctima piensa que el agresor puede hacerle mucho daño o incluso matarla

16 La Ley de Enjuiciamiento Criminal española dispone (artículo 544 ter):

"1. El Juez de Instrucción dictará orden de protección para las víctimas de violencia doméstica en los casos en que, existiendo indicios fundados de la comisión de un delito o falta contra la vida, integridad física o moral, libertad sexual, libertad o seguridad de alguna de las personas mencionadas en el artículo 173.2 del Código Penal, resulte una situación objetiva de riesgo para la víctima que requiera la adopción de alguna de las medidas de protección reguladas en este artículo.

2. La orden de protección será acordada por el juez de oficio o a instancia de la víctima o persona que tenga con ella alguna de las relaciones indicadas en el apartado anterior, o del Ministerio Fiscal.

Sin perjuicio del deber general de denuncia previsto en el artículo 262 de esta ley, las entidades u organismos asistenciales, públicos o privados, que tuvieran conocimiento de alguno de los hechos mencionados en el apartado anterior deberán ponerlos inmediatamente en conocimiento del juez de guardia o del Ministerio Fiscal con el fin de que se pueda incoar o instar el procedimiento para la adopción de la orden de protección.

3. La orden de protección podrá solicitarse directamente ante la autoridad judicial o el Ministerio Fiscal, o bien ante las Fuerzas y Cuerpos de Seguridad, las oficinas de atención a la víctima o los servicios sociales o instituciones asistenciales dependientes de las Administraciones públicas. Dicha solicitud habrá de ser remitida de forma inmediata al juez competente. En caso de suscitarse dudas acerca de la competencia territorial del juez, deberá iniciar y resolver el procedimiento para la adopción de la orden de protección el juez ante el que se haya solicitado ésta, sin perjuicio de remitir con posterioridad las actuaciones a aquel que resulte competente.

Los servicios sociales y las instituciones referidas anteriormente facilitarán a las víctimas de la violencia doméstica a las que hubieran de prestar asistencia la solicitud de la orden de protección, poniendo a su disposición con esta finalidad información, formularios y, en su caso, canales de comunicación telemáticos con la Administración de Justicia y el Ministerio Fiscal.

4. Recibida la solicitud de orden de protección, el Juez de guardia, en los supuestos mencionados en el apartado 1 de este artículo, convocará a una audiencia urgente a la víctima o su representante legal, al solicitante y al presunto agresor, asistido, en su caso, de Abogado. Asimismo será convocado el Ministerio Fiscal... Cuando excepcionalmente no fuese posible celebrar la audiencia durante el servicio de guardia, el Juez ante el que hubiera sido formulada la solicitud la convocará en el plazo más breve posible. En cualquier caso la audiencia habrá de celebrarse en un plazo máximo de setenta y dos horas desde la presentación de la solicitud.

Durante la audiencia, el Juez de guardia adoptará las medidas oportunas para evitar la confrontación entre el presunto agresor y la víctima, sus hijos y los restantes miembros de la familia. A estos efectos dispondrá que su declaración en esta audiencia se realice por separado.

Celebrada la audiencia, el Juez de guardia resolverá mediante auto lo que proceda sobre la solicitud de la orden de protección, así como sobre el contenido y vigencia de las medidas que incorpore. Sin perjuicio de ello, el Juez de instrucción podrá adoptar en cualquier momento de la tramitación de la causa las medidas previstas en el artículo 544 bis.

5. La orden de protección confiere a la víctima de los hechos mencionados en el apartado 1 un estatuto integral de protección que comprenderá las medidas cautelares de orden civil y penal contempladas en este artículo y aquellas otras medidas de asistencia y protección social establecidas en el ordenamiento jurídico.

La orden de protección podrá hacerse valer ante cualquier autoridad y Administración pública.
6. Las medidas cautelares de carácter penal podrán consistir en cualesquiera de las previstas en la legislación procesal criminal. Sus requisitos, contenido y vigencia serán los establecidos con carácter general en esta ley. Se adoptarán por el Juez de instrucción atendiendo a la necesidad de protección integral e inmediata de la víctima y, en su caso, de las personas sometidas a su patria potestad, tutela, curatela, guarda o acogimiento.
7… Cuando se dicte una orden de protección con medidas de contenido penal y existieran indicios fundados de que los hijos e hijas menores de edad hubieran presenciado, sufrido o convivido con la violencia a la que se refiere el apartado 1 de este artículo, la autoridad judicial, de oficio o a instancia de parte, suspenderá el régimen de visitas, estancia, relación o comunicación del inculpado respecto de los menores que dependan de él. No obstante, a instancia de parte, la autoridad judicial podrá no acordar la suspensión mediante resolución motivada en el interés superior del menor y previa evaluación de la situación de la relación paternofilial.
8. La orden de protección será notificada a las partes, y comunicada por el Secretario judicial inmediatamente, mediante testimonio íntegro, a la víctima y a las Administraciones públicas competentes para la adopción de medidas de protección, sean éstas de seguridad o de asistencia social, jurídica, sanitaria, psicológica o de cualquier otra índole. A estos efectos se establecerá reglamentariamente un sistema integrado de coordinación administrativa que garantice la agilidad de estas comunicaciones.
9. La orden de protección implicará el deber de informar permanentemente a la víctima sobre la situación procesal del investigado o encausado así como sobre el alcance y vigencia de las medidas cautelares adoptadas. En particular, la víctima será informada en todo momento de la situación penitenciaria del presunto agresor. A estos efectos se dará cuenta de la orden de protección a la Administración penitenciaria.
10. La orden de protección será inscrita en el Registro Central para la Protección de las Víctimas de la Violencia Doméstica y de Género.
11. En aquellos casos en que durante la tramitación de un procedimiento penal en curso surja una situación de riesgo para alguna de las personas vinculadas con el investigado o encausado por alguna de las relaciones indicadas en el apartado 1 de este artículo, el Juez o Tribunal que conozca de la causa podrá acordar la orden de protección de la víctima con arreglo a lo establecido en los apartados anteriores".

17 "Eficacia predictiva de la valoración policial del riesgo de la violencia de género", *Psychosocial Intervention*, 25, pp. 1-7.

18 Sentencia disponible en <https://vlex.es/vid/727894245> (a 2 de mayo de 2023).

19 GONZÁLEZ-ÁLVAREZ ET ALII "Integral Monitoring System in Cases of Gender Violence. VioGén System", *Behavior & Law Journal*, 4(1), 2018, pp. 29-40; en particular, p. 37.

20 *Auditoría externa del sistema Viogén*… p. 34, <https://eticasfoundation.org/es/la-fundacion-eticas-realiza-una-auditoria-externa-e-independiente-del-sistema-viogen/> (a 2 de mayo de 2023).

21 <https://www.europarl.europa.eu/doceo/document/TA-9-2021-0405_ES.html> (a 2 de mayo de 2023).

22 Según la Recomendación sobre la Ética de la Inteligencia Artificial de la UNESCO, aprobada en la reunión del 9 al 24 de noviembre de 2021, "la explicabilidad supone hacer inteligibles los resultados de los sistemas de IA y facilitar información sobre ellos. La explicabilidad de los sistemas de IA también se refiere a la inteligibilidad de la entrada, salida y funcionamiento de cada componente algorítmico y la forma en que contribuye a

los resultados de los sistemas. Así pues, la explicabilidad está estrechamente relacionada con la transparencia, ya que los resultados y los subprocesos que conducen a ellos deberían aspirar a ser comprensibles y trazables, apropiados al contexto. Los actores de la IA deberían comprometerse a velar por que los algoritmos desarrollados sean explicables. En el caso de las aplicaciones de IA cuyo impacto en el usuario final no es temporal, fácilmente reversible o de bajo riesgo, debería garantizarse que se proporcione una explicación satisfactoria con toda decisión que haya dado lugar a la acción tomada, a fin de que el resultado se considere transparente".

Conforme a las *Directrices éticas para una IA fiable* del Grupo de expertos de alto nivel sobre inteligencia artificial de la Unión Europea, "la explicabilidad es crucial para conseguir que los usuarios confíen en los sistemas de IA y para mantener dicha confianza. Esto significa que los procesos han de ser transparentes, que es preciso comunicar abiertamente las capacidades y la finalidad de los sistemas de IA y que las decisiones deben poder explicarse –en la medida de lo posible– a las partes que se vean afectadas por ellas de manera directa o indirecta. Sin esta información, no es posible impugnar adecuadamente una decisión. No siempre resulta posible explicar por qué un modelo ha generado un resultado o una decisión particular (ni qué combinación de factores contribuyeron a ello). Esos casos, que se denominan algoritmos de «caja negra», requieren especial atención. En tales circunstancias, puede ser necesario adoptar otras medidas relacionadas con la explicabilidad (por ejemplo, la trazabilidad, la auditabilidad y la comunicación transparente sobre las prestaciones del sistema), siempre y cuando el sistema en su conjunto respete los derechos fundamentales. El grado de necesidad de explicabilidad depende en gran medida del contexto y la gravedad de las consecuencias derivadas de un resultado erróneo o inadecuado".