AI that suggests solutions. The Spanish system "VioGén 2" and its impact on decisional freedom

by *Leonor Sanz Gallardo*

This paper critically analyses the use of artificial intelligence systems with recommendation functions in legal and security contexts, focusing on their significant yet indirect effects on decision-making autonomy and the protection of fundamental rights. Starting with the conceptual distinction between decision support systems and automated decision-making systems, the paper examines the Spanish gender-based violence risk assessment system, "VioGén", paying particular attention to its evolution in the "VioGén 2" version. The analysis reveals that, in practice, algorithmic recommendations can have *de facto* binding effects through cognitive anchoring, procedural standardisation, and methodological opacity. This reduces the effectiveness of human oversight. These dynamics impact both victim protection and the information environment that informs judicial decision-making. In light of European and international law on artificial intelligence and human rights, the study argues for the need to guarantee transparency, verifiability and human oversight.

1. Introduction: what "AI that suggests solutions" means in the judicial sphere

A system that formulates algorithmic recommendations is a type of artificial intelligence that processes data through a mathematical or statistical model to generate an orientative output, such as a score, category, prediction, or suggestion. This output is not binding and does not replace human decision-making; its declared purpose is solely to inform the person who must adopt the final decision. These systems are characterized by not replacing human judgment, transforming data into a structured interpretation of reality, and producing standardized and repeatable results that allow information to be organized without themselves executing any action or adopting autonomous decisions. Unlike automated decision-making systems, they do not trigger measures on their own; instead, they provide elements that must be evaluated by a human.

In the judicial sphere, where resolutions must be personal, reasoned, and the result of critical assessment, this distinction is essential.

Although algorithmic recommendation systems do not make decisions or have binding authority, their functioning produces psychological, operational, and institutional effects that can influence, orient, or even *de facto* determine human decision-making, with particular impact on fundamental rights. By processing data and generating recommendations, such

systems may introduce errors, biases, or distortions that affect rights such as equality, non-discrimination, privacy, or access to justice, as noted by both the UN High Commissioner for Human Rights and the EU Fundamental Rights Agency (FRA)[1].

The convergence among the main international principles, such as the Council of Europe Framework Convention on Artificial Intelligence (2024)[2], the UNESCO Recommendation on the Ethics of AI (2021), the Inter-Parliamentary Union Ethical Guidelines on Human Autonomy and Oversight, and the emerging doctrine on the right to human control, shows that regulation of AI systems must start from the respect for and guarantee of human rights, imposing clear limits and safeguards. All these instruments coincide in that when these systems are used in sensitive public contexts, they must operate with transparency, auditability, methodological comprehensibility, and the real possibility of human correction, avoiding that their recommendations become uncontrolled automatisms conditioning institutional decisions.

Once this conceptual and legal framework, requiring preservation of effective human control, is established, it becomes especially useful to analyze a concrete case in which these tensions clearly arise: The Spanish system for assessing the risk of gender-based violence known as "VioGén", operating since 2007 and reformed in 2025 under the name "VioGén 2".

2. Background: "VioGén" and its evolution to "VioGén 2"

Before analyzing the "VioGén" system in depth and explaining the extent to which it constitutes a paradigmatic example of the impact AI systems can have on decisional freedom in the legal sphere, it is important to recall that it is an instrument designed to assess risk in cases of gender-based violence, whose operability may in practice condition the actions of those responsible for victim protection and thereby affect the fundamental rights at stake. The VioGén System was created by the Ministry of the Interior and launched on 26 July 2007, in compliance with Organic Act 1/2004 of 28 December on Comprehensive Protection Measures against Gender Violence. Its immediate regulatory basis was Instruction 10/2007 of the Secretariat of State for Security, which approved the first protocol for police risk assessment, later modified through various administrative instructions.

Thus, VioGén was conceived with several institutional objectives: to gather all competent public institutions, integrate all relevant information on active cases into a single system, generate risk predictions based on empirically validated factors, assign protection and monitoring measures according to the level of risk, issue automatic warnings or alerts in case of significant incidents, and create a national network of coordinated and homogeneous protection throughout the territory[3].

The system feeds its database through police reports, interviews with victims, forensic reports, protection orders, police and judicial records, as well as personal and social information about both the victim and the aggressor. Using all this information, VioGén applies automated actuarial tools, particularly the Police Risk Assessment (VPR) and the Police Assessment of the Evolution of Risk (VPER), based on statistical models following methodologies similar to epidemiological approaches used in public health[4].

It establishes five levels of risk (not observed, low, medium, high, and extreme), each associated with concrete protection measures ranging from basic monitoring to intensive surveillance or the request for a judicial restrictive order. Although agents formally maintain the ability to adapt the assessment, in practice, the algorithmic recommendation sets the initial framework of measures to be considered.

Since 2007, VioGén has recorded nearly three million risk assessments, becoming a national platform for coordination among police, judicial, penitentiary, and social institutions. The Ministry of the Interior itself acknowledges that the system is an "evolving instrument," subject to periodic revisions through instructions that update its actuarial logic and functionalities. The experience accumulated over nearly two decades shows that VioGén presents structural failures with real impact on victim protection. Among the most significant, an analysis cited by the Human Rights Research Center[5], based on data published by *The New York Times*, revealed that 55 of 247 women murdered

1. FRA *Getting the Future Right – Artificial Intelligence and Fundamental Rights* (2021).
2. 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.
5. www.humanrightsresearch.org/post/spain-s-gender-violence-detection-algorithm-reveals-failures-in-saving-domestic-violence-victims.

after a police evaluation had been classified as "low" or "negligible" risk. This outcome indicates that the tool fails not only in marginal cases, but in critical situ ations where the assessment was insufficient to trigger adequate protective measures.

A well-known example is that of Lobna Hemid, murdered in 2022 after being evaluated by VioGén as low risk. The initial classification, perceived as technically neutral, conditioned the institutional response and contributed to the failure to adopt necessary protection measures

Another recurrent criticism concerns the automatic acceptance of results. According to an audit by Eticas Foundation[6], agents maintain the system-generated score in 95% of assessments, even though the protocol allows adjustments based on professional judgment. As a result, human oversight becomes a formality and the algorithmic recommendation becomes a *de facto* operational decision.

This problem is aggravated by the fact that victims often do not disclose critical information during police interviews due to fear, economic dependence, or migration status, generating incomplete or biased data. These initial biases are transferred to the algorithm, which underestimates risk precisely in the most vulnerable victims.

Moreover, the actuarial approach has inherent limitations. According to the Ministry of the Interior, the models used in VPR and VPER are based on statistical markers associated with violent recidivism but lacking direct causal links. Consequently, small variations in responses can shift the assigned risk level, while the models fail to capture qualitative or dynamic factors specific to each case.

Finally, the lack of transparency is among the most persistent criticisms. The Eticas Foundation audit noted that VioGén "does not allow knowing the exact formula, the assigned weights, or the complete empirical basis of the model," and that it has not undergone independent external technical audits. This opacity violates key international standards on ethical and responsible use of AI, which require traceability, auditability, and effective human oversight in systems affecting fundamental rights. In this sense, the Council of Europe Framework Convention on Artificial Intelligence—the first binding international treaty on the topic—imposes on States the obligation to guarantee systems' traceability and adequate human supervision throughout their life cycle. The UNESCO Recommendation on the Ethics of AI (2021) similarly requires transparency, explainability, and human oversight, safeguarding dignity and fundamental rights. The Inter-Parliamentary Union Ethical Guidelines also stress that any AI used in public functions must preserve human autonomy and avoid concealed delegation of decisions, insisting on guaranteed continuous and effective human control.

All these factors erode the guarantees that must govern AI use in fundamental rights contexts. Furthermore, the algorithmic classification conditions the intensity of protection provided: according to data cited by *Politico*, only one out of seven women requesting police protection ultimately receives it, largely because priority assistance is conditioned by the risk level assigned by the system[7].

These accumulated criticisms led the Ministry of the Interior to implement, in 2025, a comprehensive reform of the system resulting in "VioGén 2". This update is structured around three lines of action directly related to the failures detected in the previous model: first, elimination of the "no risk" category, which had led to particularly serious underestimations in practice; second, the mandatory incorporation of qualitative and contextual information in police interviews, aimed at reducing incomplete data resulting from victims' fear, economic dependency, or administrative situation; and third, the introduction of personalized protection plans and improved inter-institutional coordination, designed to prevent responses excessively dependent on the algorithmic classification. These lines of reform appear in the editorial material on the 2025 system update.

With these measures, VioGén 2 aims not only to improve the system's technical accuracy, but above all to correct the risks of concealed automation detected in practice, expanding the margins of human oversight and restoring operators' ability to exercise a more complete and contextualized judgment. VioGén 2 therefore constitutes an institutional response to the challenges posed by AI in the protection of fundamental rights.

3. Actual functioning: suggestion or concealed decision?

Although the VioGén system was formally conceived as a support tool intended to guide professional action, its practical functioning reveals that algorithmic recommendations often become, in a substantial number of cases, concealed decisions. This transformation is not intentional but responds

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

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

to specific mechanisms common to AI systems integrated into institutional circuits. As shown earlier, although the official design of VioGén states that the algorithm provides only an orientative risk level and that the police authority must independently assess the case, documented practice indicates that operators follow the algorithmic recommendation with minimal human intervention, maintaining the automatic result in 95% of evaluations, according to Eticas Foundation.

This occurs because, in practice, the system's classification functions as a cognitive framework acting as a preliminary diagnosis that conditions how interviews, warning signs, or the victim's narrative are interpreted. This is a typical anchoring effect, where the first value received becomes a dominant reference for all subsequent analysis. Added to this are workload pressures, operational constraints, and the need for standardization, which lead agents to adopt the recommendation as an operational starting point, triggering protocols, priorities, and determining the intensity of monitoring. In this context, the recommendation functions as a pre-decision because it is already connected to a set of preconfigured police measures.

At a psychological level, the fact that the algorithm does not produce an interpretative text but a closed category ("low risk," "medium risk," "high risk") simplifies a complex reality into an apparently precise result, which acquires technical authority over professional intuition. The combination of cognitive anchoring, organizational standardization, and lack of transparency turns the algorithmic recommendation into a decisional reference that conditions police practice and limits the real capacity for human revision.

These tensions become especially relevant in the judicial sphere, as AI systems used in justice or security are classified by the EU AI Act (Regulation (EU) 2024/1689)[8] as high-risk systems, requiring transparency, traceability, significant human oversight, and fundamental rights impact assessments, precisely to prevent undue conditioning of judicial decisional freedom in contexts where judges receive information previously processed by automated systems.

4. Impact on judicial decisional freedom

The use of algorithmic systems within the justice ecosystem raises a crucial question: how to protect decisional freedom when the information reaching the judge has been previously configured by an automated system. In judicial proceedings, such freedom requires the ability to analyze the case without external conditioning, evaluating facts and risks with full autonomy. However, when the information reaching the court comes from a system whose internal logic is not auditable, the cognitive environment in which the judge must deliberate is altered. This prior configuration—resulting not from an intention to interfere, but from the way AI systems organize, synthesize, or filter preliminary police information—may influence the assessment of urgency, proportionality, and risk in adopting precautionary or protective measures.

In the European and international framework, judicial decisional autonomy is not conceived merely as the absence of direct external pressures, but as the right and duty to decide based on understandable and verifiable information, free from conditioning by automated mechanisms. This is required by Article 6 of the European Convention on Human Rights, which demands an "independent and impartial" tribunal, a guarantee that includes independence from technological influences limiting the judge's real capacity to form judgment. Likewise, Article 47 of the Charter of Fundamental Rights of the European Union enshrines "effective judicial protection," presupposing the availability of sufficient and clearly accessible information for the judge to reason decisions without structural constraints. This understanding is reinforced by Council of Europe standards, especially those of the CCJE, whose Opinions (notably Opinion No. 26 (2023) on support technologies in justice) emphasize that judicial independence includes the capacity to control the rationality of the decision-making process, requiring that the judge understand, audit, and question the information incorporated into judicial reasoning.

The Council of Europe Framework Convention on AI (2024) strengthens this requirement by imposing obligations of transparency and adequate human supervision throughout the life cycle of algorithmic systems, preventing AI from operating as an uncontrolled filter of the reality reaching the courts. Similarly, the UNESCO Recommendation on the Ethics of AI (2021) requires the primacy of human agency in all public decision-making, warning that non-transparent systems may displace professional judgment and weaken individualized reasoning. This instrument states that human supervision must be real, meaningful, and constant, especially when AI affects essential rights. For its part, the Inter-Parliamentary Union Ethical Guidelines underline that AI used in public functions must preserve

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

human decisional autonomy and avoid any form of concealed delegation, insisting on the need for permanent human control.

In the European Union, the EU AI Act (Regulation (EU) 2024/1689) classifies AI systems applied to justice and security as "high risk," subjecting them to the strictest regulatory requirements: enhanced transparency, documentary traceability, significant human oversight, stringent data governance, and fundamental rights impact assessments before deployment. According to the summary by the Oxford Institute of Technology and Justice, this category reflects the fact that AI used in justice can materially affect due process, equality of arms, and the tribunal's effective impartiality, requiring reinforced safeguards to avoid AI deforming the factual environment in which judicial deliberation takes place.

From this European and international perspective, the emergence of VioGén poses not merely a technical problem but a constitutional one: the judiciary cannot operate with full independence when the procedural reality it examines is already conditioned by a system whose methodology, internal logic, and biases are not accessible to the judge. Judicial independence includes independence from non-transparent technological infrastructures, and effective judicial protection requires that judges be able to reconstruct the case critically and provide reasoning without depending on a closed statistical category or an unexplained algorithmic assessment.

Therefore, the true challenge posed by automated risk assessment systems lies not in formally replacing the judge, expressly prohibited without exception in the European framework, but in reducing the epistemic control the judge must exercise over relevant information. Preserving judicial decisional freedom requires ensuring that AI does not define the contours of procedural reality, but merely provides comprehensible, verifiable, and audited information, always subordinated to human judgment.