Adversarial Audit

March
2022

Can AI solve gender violence?

Auditing the use of AI to assess risk. The case of Viogén

eticas
Foundation

FUNDACIÓN Ana Bella | RED DE MUJERES SUPERVIVIENTES

To Cathy O'Neil. Her book "Weapons of Math Destruction" was published in 2016, and in 2017 we started looking into VioGén. Her groundbreaking take on algorithmic fairness and accountability has shaped our work and continues to inspire us.

CONTENT

Adversarial Audit of the VioGén System

1- Introduction

Gender-based violence against women is a world-wide phenomenon. The UN estimates 736 million women -almost one in three- have been subjected to intimate (ex)partner violence and non-partner sexual violence at least once in their life around the world.[1] The most developed parts of the world are not exceptional to this trend. Violence against women has also been a key problem in Spain where 1.126 women were killed by their (ex)intimate partners between 2003 and 2021.[2] 32.4% of women in Spain aged 16 years and older women (approximately 6.6 million) have suffered physical, sexual, and/or psychological (emotional, control, economic, and fear) from their (ex)partners throughout their life (Delegación del Gobierno contra la Violencia de Género, 2019b). Faced with the need to provide adequate protection to the women who report instances of gender violence, many countries have developed specialized programs. In some of them, including Spain, such programs include a risk assessment tool that provides an algorithmic risk score that is used to make decisions or recommendations on what police and legal actions should be enacted to protect women.

The use of automated systems to predict risk has been increasing in recent years, often raising concerns about fairness and transparency. In our own work at Eticas, we have highlighted how such systems are often initially deployed in contexts that affect vulnerable populations, which also raises concerns about impact and redress. In order to look at these issues more closely, in 2021 we launched our Adversarial Audit Project, where we are currently reverse-engineering algorithmic systems in different areas (criminal justice, employment, social media, and banking).[3] in collaboration with the affected stakeholders and civil society organizations. VioGén is our first report in what we hope will be a long series.

At Eticas we are well-known for our Internal Audits, which we have been conducting for public and private clients for the last five years. In those cases, we are hired by those developing and /or implementing an algorithmic system to identify and correct instances of lack of fairness and inefficiencies. Our Internal Audit methodology is end-to-end, and so both technical and social, and we look at inputs, models and processes, but also outputs and impacts. One of the things we have learned through our hands-on auditing experience is that reverse-engineering systems is a good exercise even when you can access the code and the relevant data, as some bias dynamics may not be evident before they translate into impacts. This realization led us to consider the possibility of developing an Adversarial Audit methodology for those cases when access to the code or data is not possible.

An adversarial audit is a process by which an independent third party examines the impact and, to the extent possible, the functioning of an algorithmic system in order to detect

[1] https://www.unwomen.org/en/what-we-do/ending-violence-against-women/facts-and-figures

[2] https://violenciagenero.igualdad.gob.es/violenciaEnCifras/victimasMortales/fichaMujeres/pdf/VMortales_2022_01_25_2.pdf

[3] https://eticasfoundation.org/auditorias-externas-algoritmos/

potential anomalies or practices that could be unfair or harmful towards protected groups or society as a whole. The main particularity of adversarial algorithmic audits is that the access to the algorithm and the databases used to design, develop, test, and validate it is usually restricted. In light of this, the impact is often assessed by means of systemic analyses of the populations affected, secondary sources, and databases containing data scraped via different collection mechanisms. When we looked for bibliography on how to externally audit algorithms, the sources we found mostly referred to the adversarial auditing of social media or web-based services (Ada Lovelace Institute, 2021; Sandvig et al., 2014) so we set ourselves to developing and testing different methodologies in different fields and areas, and on different technical systems.

In the case of VioGén, the gender-violence risk assessment tool used by the Spanish Ministry of the Interior, we have a history. Concerned by its functioning and impact, in 2018 we reached out to the Ministry to request information and a meeting, which was only held after we requested the assistance of a Member of Congress. Since then, we have had several meetings with those in charge of the system, where Eticas offered a confidential pro-bono internal audit of the VioGén system and to consider the incorporation of supervised machine learning tools to gather insight from the large amounts of historical data produced by the system. While these suggestions were well received, no action was taken and our offer to conduct an internal audit never materialized. Therefore, in 2021 we set out to start our **Adversarial Audit project of VioGén**, with the collaboration of Ana Bella Foundation, a leading CSO working with women who have survived domestic violence and thus who, at some point or other, have had a risk score produced by the VioGén system.

The questions that we have addressed through this Adversarial Audit exercise on VioGén, are manifold. On the one hand, we are concerned about the **transparency** of the system and the obligations of the Ministry of the Interior in this regard. In 2015, the legal regime of the public sector in Spain was revised to include a provision according to which **automated actuations deployed by the public administration** (e.g. via algorithmic systems) **are subject to a set of governing bodies and processes** to ensure their adequate functioning, including auditing.[4] Even though the scope of this precept covers fully automated systems exclusively (where no human is involved), our data shows that the algorithmic risk assigned by VioGén remains unaltered 95% of the time (Zurita Bayona, 2014). In light of this, we argue that **for publicly-funded, highly automated decision-making systems of enormous social impact such as VioGén** (which, in some cases, makes life-and-death decisions), **independent audits should be required by law.**

We have also noted that **most VioGén studies have been conducted by the same researchers that contributed to its development** (López-Ossorio et al., 2019; López-Ossorio, González-Álvarez, et al., 2020), **and individuals who either work for or have vested interests in the ministry and police forces**. This reinforces our argument for the need for independent oversight of the system, and we hope this Adversarial Audit prompts those responsible for VioGén to commission an independent audit and publish its results.

[4] Ley 40/2015 art. 41

The role of police officers in validating or altering the VioGén risk-score also raises issues of **accountability**. Having a higher or lower VioGén risk score means that a woman will receive different levels of police protection. But it is **unclear who is responsible for that decision**. As 95% of police officers chose to not alter the suggested risk score, it seems clear that they see it as more than just a recommendation, and they are delegating their agency onto the system. Also, if their willingness to question the system decreases when their workload is increased, as some suggest (Estévez Mendoza, 2020), it seems clear that further attention needs to be paid to the impact of the "human in the loop" on the overall performance, fairness and accountability of the system.

Another area of concern is the **lack of participation of the affected populations** in the design and follow-up of the system. Much of the existing literature on VioGén focuses on its technical aspects - especially regarding the predictive validity of the algorithm - and not on its social impact, the role or experiences of the women affected by it. Even though there have been attempts to survey user satisfaction (González-Álvarez & Garrido, 2015), it is concerning that for a system aimed at being used with very vulnerable populations, end users and end-user groups have not been taken into account nor consulted. This is something we extensively address in this Adversarial Audit, focusing not only on technical issues, but also on the impact VioGén has on victims[5] of gender violence, and **our findings are very concerning**.

This is specially relevant at a time when the **Ministry of the Interior seems to be considering the incorporation of Machine Learning** (ML) into the system. Although to the date of this report it has not been officially confirmed, there have been recent initiatives to incorporate ML and advanced data analysis techniques to the VioGén system (Pinedo, 2021). In December 2020 the software company SAS announced that the Ministry of the Interior and the Gender Violence Unit reached an agreement with this software company to incorporate data analytics and what has been branded as the "digital agent" to automate and streamline certain processes to increase protection.[6] Eticas contacted SAS to better understand the nature of this agreement, but they refused to disclose any information given that the proprietor of the system is the Ministry of the Interior. As we describe below, the current VioGén is an actuarial system that uses statistical models to infer future risk. As such, it is a rather simple algorithmic system where information is inputted in a specific format (a questionnaire), and is assessed using different weights. While we would argue that **the data possibilities of VioGén are underutilized**, it is unclear whether *machine* learning - the creation of automated feedback loops into the system, turning it into an Artificial Intelligence model - is desirable from a point of view of accountability and transparency. In any case, **if the debate emerges, it should not be taking place behind**

[5] Following the official terminology of legal and judicial institutions in Spain, this report refers to women who are subjected to male-agression by their ex- or current partners as "victims" (*víctimas*) and men who are perpetrators of this aggression as "aggressors'' (*agresores*). This choice of terminology is adapted for practical reasons and serves for being compatible with official accounts. On the other hand, we acknowledge that the term "victim" is a highly contested concept and is criticized for further victimizing women. That is why many women rights activists and academics opt for the term "survivor". See for example: Ana Bella Foundation, Network of Women Survivors: https://www.fundacionanabella.org/

[6] https://www.sas.com/es_es/news/press-releases/locales/2020/viogen-secretaria-estado-seguridad-y-sas-unidos-lucha-contra-violencia-genero-analitica-avanzada-ia.html

closed doors and without taking into account, at the very least, the point of view of the women affected.

Therefore, when in mid-2021 we set up the team to Audit VioGén, **we had concerns around transparency, independent oversight, accountability, end-user engagement and the transition to ML**. The auditing process has made it possible for us to turn general concerns into specific questions, building our case for more transparency and oversight, better accountability and an assessment of social impact around hard data. While external tools do not allow us to be conclusive, they have provided us with the data to ask and justify our questions. The results we present below are concerning, but the process has convinced us of the usefulness of our approach. **If we have managed to get to this stage without any access to the relevant data, imagine what we, and society as a whole, could have done with access to it.** We hope that this report prompts change in the way VioGén works and evolves.

In the sections below, we present the **results of 7 months of work of Eticas and Ana Bella Foundation** with the available data and affected women and other stakeholders. As mentioned above, it is a part of a broader adversarial auditing project where Eticas, in collaboration with other civil society organizations, reverse engineers and assesses the impact of algorithms in different fields. With this Adversarial Audit project we aim to develop methodological tools to externally audit automated risk assessment systems in the absence of access to the code, input, output, and administrative data to provide methodological tools to community organizations for externally auditing algorithms with social impact and advocating for policy change. In this way, we seek to support bottom-up algorithmic auditing movements conducted by third-party organizations and end-user groups.

The report is **structured** around the process that a victim of gender violence undergoes when filing a police report, from the beginning to the end, with the aim to close the gap between existing literature on VioGén's technical validity and the lived experiences of those women whose life has been affected or even determined by it. We first provide a technical overview of the VioGén system. Then, we critically discuss the system and explore its strengths and pitfalls by inductively studying the perceptions and experiences of some of the major stakeholders – such as survivors of domestic violence who have gone through the VioGén system, their lawyers, and civil society organizations working in this field. By doing so, we establish the grounds to compare the system's design and evaluation with its actual operation and the ways in which it is experienced by its key stakeholders.

2- Assessing risk in the context of gender violence

Both scientists and policy-makers have been working towards developing quick, effective, precise and practical violence prevention solutions. In a hypothetical world with infinite resources, the best strategy would be to provide high level protection and surveillance to all women who feel at risk of suffering gender violence. Unfortunately, police resources are limited and need therefore to be carefully allocated to those who are expected to be at higher risk. In this regard, risk assessment plays a crucial role in gender violence prevention.

Risk assessment tools are designed to categorize gender violence cases according to the level of risk that can be foreseen. Therefore, they aim to provide an accurate prediction of which victims of gender violence are more likely to be assaulted again and therefore are in need of protection. A risk score, in this context, does not evaluate the gravity of the past or current incidents, but rather predicts the likelihood of having a future episode of gender violence -what is assessed is the risk of recidivism of the perpetrator.

Even though risk assessment tools predicting gender violence are not new, in the last three decades there have been major breakthroughs in terms of their accuracy and scientific status. First of all, clinical and socio-psychological studies have taken important steps in identifying major gender violence risk factors. While there is still little consensus in the literature about what is meant by *risk* in the context of gender violence (Kropp, 2004),[7] there is considerable agreement on what constitutes a risk factor (Campbell et al., 2001; Riggs et al., 2000). These lists of risk factors have provided the scientific grounds on which risk assessment tools are built. Second, developments in information and communication technologies have enabled public institutions with competences in gender-violence prevention to share information and synchronize their actions. Therefore, risk assessment tools can rely on multiple databases that bring together different types of information. Third, advances in data science and analytics have provided better methods of knowledge extraction/discovery from data, data/pattern analysis, and made predictive models possible (Sarker, 2021).

Despite the progress in the field, there are still controversies over how risk assessment must be conducted, by whom, what role professionals and victims have in this process, and how a risk assessment must inform the process of risk management. The literature mentions **three main approaches** to risk assessment (Heilbrun et al., 2011; Kropp, 2004).

- **Unstructured clinical assessment** entails professional evaluation of each specific case and individualized tailoring of risk management. The advantage of this

[7] This is because there is no such thing as “no risk” in the context of intimate (ex)partner violence as well as there are different types of risks that vary in terms of imminence, nature (such as emotional, physical, sexual etc.), frequency, and seriousness (Kropp, 2004). This also shows the difficulty of assessing risk based on a uni-dimensional scale (e.g. low-high), since it has multiple dimensions that need to be considered.

approach is its ability to account for unique, unusual, and context-specific conditions that need to be evaluated case-by-case by professionals. On the other hand, this professional discretion may come at the cost of having a reliable and valid system, since the evaluation heavily depends on the training, preferences, and biases of the professional.

- **Actuarial assessment** is based on statistical evaluation of pre-determined and scientifically defined risk factors. This approach is designed to produce objective and standardized risk assessments with scientific rigor without relying on how well qualified the evaluator is. Actuarial models are criticized for being "mechanical and algorithmic" (Grove & Meehl, 1996), based on linear assumptions, and not being good enough to capture context-specific information.

- **Structured professional judgment** bridges the gap between the first two methods. Even though some steps are standardized as in the actuarial model, the final step is not done algorithmically and accounts for the responsibility and professional discretion of the evaluator (Kropp, 2004).

In addition to these three classical models of risk assessment, developments in computational methodologies as well as the availability of big digital data make it possible to apply **Machine Learning** (ML) techniques to predicting gender-based crimes. ML systems use inferential and data-driven algorithms to extract patterns from historical data (Tolan et al., 2019). They are capable of making predictions in the context of high uncertainty, being able to deal with a large number of features as input, and allowing for scalability. On the downside, these systems heavily depend on historical data, which can contribute to perpetuating structural biases and inequalities, while its internal complexity can lead to a lack of transparency and the so-called black-box effect.

As of today, VioGén is an actuarial system that uses statistical models to infer the risk that a victim faces (both of aggression and homicide) as well as its evolution based on a set of indicators that have been determined and later evaluated by a group of experts. As discussed later, the possibility of incorporating a machine learning algorithm in the VioGén system, based on the Nearest Centroid technique for classification – or a hybrid model that implements a stochastic mix of the current system and Nearest Centroid – that would seemingly outperform VioGén has been discussed (González-Prieto et al., 2021, p. 6), but there is no evidence that the Ministry may be considering this option..

Each approach to risk assessment has its own advantages and limitations, and offers different perspectives to assessing the risk of recidivism (i.e. predicting the likelihood of new violence) and managing it (i.e. providing information for risk prevention planning). However, continuous efforts are deployed to improve automated risk assessment systems and to provide better protection for victims while effectively managing resources. Against this backdrop, this report presents a new approach to externally studying automated risk assessment systems in general, and a critical evaluation to improve VioGén in particular by analyzing both the nature of the system and the impact it has over gender violence victims.

3- The VioGén System

What is VioGén?

The "Integral Monitoring System in Cases of Gender Violence" (the VioGén System) is a web application, integrated in the Spanish SARA Network (Application Systems and Networks for Administrators). It is designed to coordinate the actions of Spanish public professionals who are in charge of monitoring, assisting, and protecting women who report gender violence and their children. In this way, VioGén aims to establish a dense network of institutions with competencies in the area of gender violence prevention and to provide fast, comprehensive, effective, and high-standard responses to gender violence across the country.[8] It has its legal origins in the mandates of Article 31 and 32 of the Organic Law 1/2004 regarding the "comprehensive protection measures against gender violence"[9].

The system was created by the Spanish Secretary of State for Security (SES) of the Ministry of Interior and launched nation-wide (except Catalunya and the Basque Country) in 2007,. It has so far performed more than **3 million risk evaluations** (López-Ossorio et al., 2019). As of January 2022, **there are 673,912 cases in the VioGén system, of which 69,391 are active cases that require police supervision**,[10] making the Spanish risk assessment system **the first in the world in terms of volume of cases** (González-Álvarez et al., 2018, p. 37).

The VioGén system is officially designed to fulfill the following objectives:[11]

- bringing together all public institutions that have competence in the area of gender violence;
- integrating all relevant information;
- making risk prediction;
- monitoring and protecting victims of gender violence by carrying out preventive work, issuing warnings and alerts, and taking other necessary actions depending on the risk level.

The system aims to integrate different public services, i.e. law enforcement (Guardia Civil and National Police), justice, health, social services, equality, and penitentiary systems to facilitate information exchange. It has more than 30,000 users with different levels of privileges (González-Álvarez et al., 2018). While all these mentioned users can access the

[8] http://www.interior.gob.es/web/servicios-al-ciudadano/violencia-contra-la-mujer/sistema-viogen

[9] https://www.boe.es/diario_boe/txt.php?id=BOE-A-2004-21760

[10] http://www.interior.gob.es/documents/642012/14732358/ENERO+2022/dd906e3b-f2d4-4a61-ade0-275ed40fddfa

[11] http://www.interior.gob.es/web/servicios-al-ciudadano/violencia-contra-la-mujer/sistema-viogen

system – and some of them can even contribute to it with relevant information–, only law enforcement agents (Police officers and Guardia Civil) can register cases. In compliance with the European GDPR (General Data Protection Regulation) and Organic Law of Data Protection (15/1999), users access the system with a username and a non-transferable password.

Guardia Civil	16.239
National Police	5.429
Local Police	1.806
Mossos d'Esquadra, Policía Foral, Attached Units of Galicia, and Comunidad Valenciana	573
Penitentiary Institutions	755
Coordination and Violence Units	128
Social and Equality Services	542
Justice Ministry and Judiciary	8.433
Total users	**33.905**

Table 1: Habilitated Users of the VioGén System as of 31/05/2020. Source: López-Ossorio (2020)

The Spanish Organic Law 1/2004 defines **gender violence** as "a manifestation of discrimination, the situation of inequality, and relations of power that is exercised by men over women by those who are/have been spouses or who are/have been linked to each other by similar affective relationship even without living together".[12] In other countries, this type of violence is often called "Intimate Partner Violence" (IPV) from male aggressor to female victim.

In the VioGén system, a **case** contains a single female victim and a single male aggressor. This means that when a woman becomes the victim of multiple aggressors, there will be a different case for each of her aggressors. In the same way, when a male aggressor targets different women, he will have multiple cases. Therefore, the number of cases outnumbers people (González-Álvarez et al., 2018, p. 33). A gender violence case is registered in the system during the victim's official complaint to the police. An **active case** (*caso activo*) means that it is actively followed and supervised by the police forces. A case becomes **inactive** (*caso inactivo*) when it no longer needs police attention. A case is **deregistered** from the system (*caso de baja*) when there is no expectation for recidivism to occur. There are three conditions for a case to be deregistred (González-Álvarez et al., 2018, p. 33):

[12] Ley Orgánica 1/2004, Articulo 1: "la violencia que, como manifestación de la discriminación, la situación de desigualdad y las relaciones de poder de los hombres sobre las mujeres, se ejerce sobre éstas por parte de quienes sean o hayan sido sus cónyuges o de quienes estén o hayan estado ligados a ellas por relaciones similares de afectividad, aun sin convivencia."

- Firm acquittal of the accused
- Dismissal of the proceedings of the investigated or processed
- Firm conviction that has been executed in which the legal term for cancellation has elapsed.

How does the VioGén Risk Assessment work?

The VioGén system intends to standardize the police assessment of gender violence risk and the protection and preventive measures around it across Spain. The system works through two questionnaires (*Protocolo Dual*): Police Risk Assessment (VPR—*Valoración Policial del Riesgo*) and Police Risk Evolution Assessment (VPER—*Valoración Policial de la Evolución del Riesgo*). The VPR form performs the first risk assessment at the moment of reporting the aggression to the police, whereas the VPER form monitors the evolution of the gender violence risk. These assessment protocols are reviewed and revised by a team of multidisciplinary experts. The fifth and most updated version was released in March 2019. Since then, the risk assessment has been carried out through $VPR_{5.0}$-H and $VPER_{4.1}$.

When a woman makes an official complaint of her aggressor, police agents fill in the $VPR_{5.0}$-H form. This form includes **5 domains** with **35 risk indicators** (see the Appendix). Each item is valued as "present" and "not present". In this way, the collection of information is standardized across the country. Once the form is filled, the system assigns a gender-violence risk score. The levels of this risk score are **"unappreciated"** (no apreciado), **"low"** (bajo), **"medium"** (medio), **"high"** (alto), and **"extreme"** (extremo). Police officers can only modify the score to a higher level of risk, not the other way around: that is, the risk score calculated by the VioGén algorithm cannot be lowered. However, and even though the officers are able to increase the automatically assigned risk score, it is reported that they rarely do this. In 95% of the cases, officers maintain the automatically assigned risk score (Zurita Bayona, 2014). Moreover, as the Covid-19 pandemic increased the workload of law enforcement agents, it has been observed in recent months how officers present a higher tendency to rely on automated decisions than before (Estévez Mendoza, 2020). Unfortunately we do not have data on how workload may be impacting on women's rights, chances and protection.

As can be seen in the table below, the distribution of cases across assigned risk scores seems to be stable across time. **The overwhelming majority of the active cases are considered to be either unappreciated or low-risk situations**, with only a minority of the cases falling into the category of medium/high/extreme risk levels requiring specific protection measures provided by the police. It must also be noted that the number of active cases has been increasing each year. In other words, while the distribution of risk categories has been stable over time, every year there are more cases with higher risk scores that require special police attention.

	Total Cases	Active Cases	Unappr.	Low	Medium	High	Extreme
2021	670,061	69,469	45.30%	41.82%	12.02%	0.83%	0.02%
2020	621,907	63,656	48.78%	40.62%	9.93%	0.66%	0.01%
2019	577,907	61,355	49.65%	39.49%	10.18%	0.65%	0.02%
2018	529,762	58,498	43.48%	45.95%	10.09%	0.45%	0.04%
2017	485,439	54,793	49.62%	41.58%	8.38%	0.39%	0.03%
2016	439,307	52,635	56.37%	36.11%	7.19%	0.32%	0.02%
2015	396,552	52,005	68.05%	26.11%	5.63%	0.19%	0.01%

Table 2: Number of total and active VioGén cases and distribution of risk scores. Source: Monthly Statistical Bulletin by Government Delegation against Gender Violence (Ministry of Equality).[13]

The monitoring of how the risk evolves is conducted via VPER$_{4.1}$. If the continuous evaluation is intended as a matter of periodic control without incidents, it is called VPER$_{4.1}$-S ("*Sin incidente*"). The period of this evaluation is determined by the risk level: extreme level: before 72 hours, high level: before 7 days, medium level: every 30 days, and low level: every 60 days. After the application of VPR$_{5.0}$-H, if a new incident occurs, then VPER$_{4.1}$-C ("*Con incidente*") is conducted. Between 2007-2019, more than 3 million evaluations (VPR and VPER) were made. According to the developers of VioGén, this is one of the highest number of risk evaluations in the world (López-Ossorio, Muñoz Vicente, et al. 2020).

In 2019, the VioGén system was adjusted to detect the cases with lethal assault risk as well as the cases where children are exposed to violence.[14] The updated VioGén system runs two evaluations in parallel (VPR$_{5.0}$ plus the H-Scale): one is recidivism (the likelihood of a new assault from the same aggressor) and the other is the risk of homicide. The cases that carry the risk of homicide are reported as cases of "special relevance" (caso de especial relevancia). In López-Ossorio et. al. (2020), the authors identified 13 indicators that bear a positive significance in terms of the risk of Inter Partner Homicides (IPH). These indicators capture certain variables of the aggressor's criminal record, mental and psychiatric disorders, or certain behavioral patterns and life- or work-related problems, as well as the victim's mental health and substance addiction (López-Ossorio, González-Álvarez, et al., 2020, p. 50). In this sense, the OR coefficients reflect the significance of each variable as a predictor of the homicide risk, being suicidal threats (OR=8.087, p<.001), economic and

[13] https://violenciagenero.igualdad.gob.es/violenciaEnCifras/boletines/boletinMensual/home.htm

[14] https://www.rtve.es/noticias/20190313/entra-vigor-nuevo-protocolo-policial-para-valorar-riesgo-victimas-violencia-genero/1900900.shtml

work-related problems in the last six months (OR=6.324, p<.001), and any kind of addiction or substance abuse in the victim (OR=5.101, p<.001) the most salient ones (Ibid.). The authors also provide sensitivity (TPR=.084) and specificity (TNR=0.60) values, in order to attest for the classification capacity of the H-Scale

(López-Ossorio, González-Álvarez et al., 2020, p. 51). With respect to children, the new VioGén system identifies two categories: "children in a situation of vulnerability" and "children in a situation of risk" (instruction 4/2019).

In these three cases - special relevance, with children in a situation of vulnerability and with children in situation of risk - an Automated Diligence (*Diligencia Automatizada*) is attached to the Police Risk Evaluation in order to call the attention of judges and prosecutors and recommend additional expert evaluation of the case.

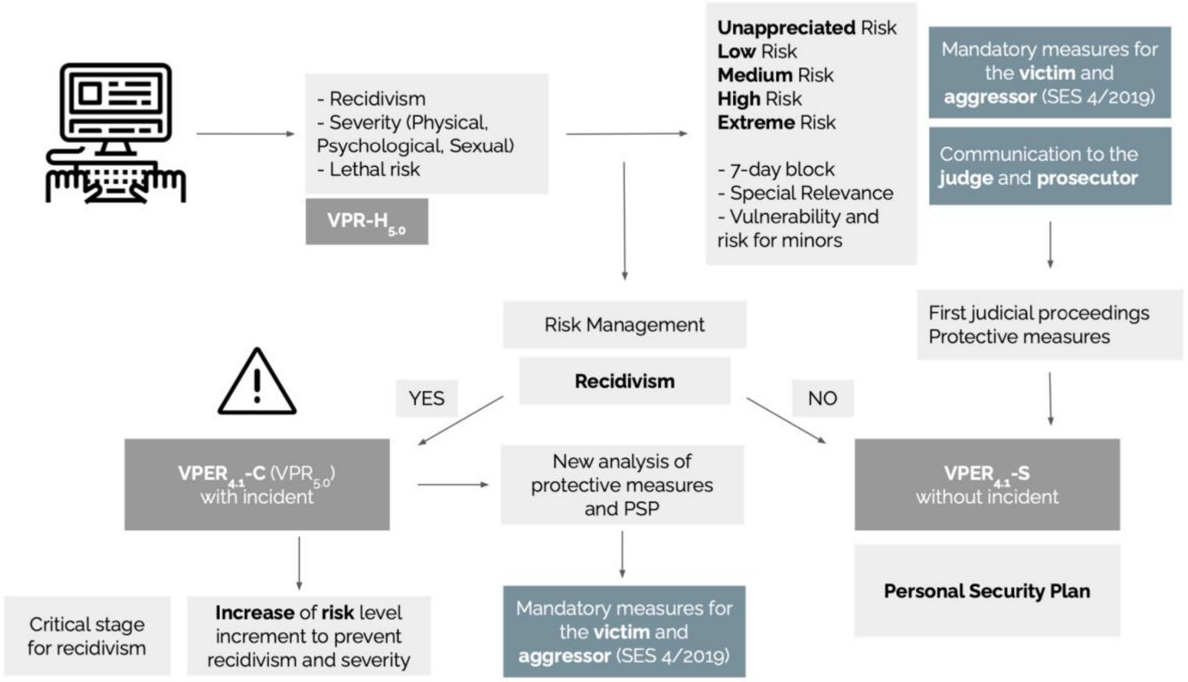

- Recidivism
- Severity (Physical, Psychological, Sexual)
- Lethal risk
VPR-H$_{5.0}$
Unappreciated Risk
Low Risk
Medium Risk
High Risk
Extreme Risk
- 7-day block
- Special Relevance
- Vulnerability and risk for minors
Mandatory measures for the **victim** and **aggressor** (SES 4/2019)
Communication to the **judge** and **prosecutor**
First judicial proceedings
Protective measures
Risk Management
Recidivism
YES
NO
VPER$_{4.1}$-C (VPR$_{5.0}$) with incident
New analysis of protective measures and PSP
VPER$_{4.1}$-S without incident
Personal Security Plan
Critical stage for recidivism
Increase of **risk** level increment to prevent recidivism and severity
Mandatory measures for the **victim** and **aggressor** (SES 4/2019)

Figure 1: The process of the VioGén system (adapted from López-Ossorio, Muñoz Vicente, et al., 2020, p. 9).

The **VioGén system automatically activates police protection measures for each victim according to their risk score**. The 2019 update also offers a Personalized Security Plan (*Plan de Seguridad Personalizado* - PSP). The PSP takes into account the specific conditions of each victim, including whether the victim has children, works outside of her home, lives with her aggressor etc. There are also increasing efforts to adapt PSPs to available technologies, and include measures such as changing the phone number, blocking calls from the aggressor, installing the AlertCop app[15], registering emergency numbers, activating geolocation, etc. Technically speaking, **VioGén uses classical statistical models to perform a risk evaluation based on the weighted sum of all the responses according to pre-set weights for each variable** (González-Prieto et al., 2021, p. 11).

In order to validate the $VPR_{4.0}$, a stratified prospective longitudinal study was conducted in order to then develop the predictive model. The sample used to that end was constituted by victims of gender violence between September the 24th and December the 1st 2015, with a follow-up window spanning until April 29, 2016 to allow recidivism detection. The sample (n=3907) was then split into two groups: 60% to build the model and 40% to validate it. Risk indicators were extracted from $VPR_{4.0}$, and Pearson's chi square was used to calculate statistical significance, with the Odds Ratio (OR) being the main dependent variable (López-Ossorio, 2017, p. 191).

Moreover, 13 indicators were included as critical identifiers - with expert weighing - in order to screen the risk of homicide. The weight of these indicators was multiplied by two, except in cases of mental disorder, which were directly given a score of 3 (López-Ossorio, 2017, p. 192). These indicators are:[16]

- Grave or very grave physical violence
- Grave or very grave sexual violence
- Use of weapons (except firearms)
- Death threats from the aggressor
- Threats or aggression build-up during the last six months
- Signs of extreme jealousy from the aggressor over the last six months
- Harassment behaviors from the aggressor over the last six months
- Aggressions to others or animals from the aggressor over the last year
- Mental or psychiatric disorder in the aggressor
- Presence of suicidal ideas or attempts by the aggressor
- Addiction or abuse of substance (alcohol, drugs or medicines) by the aggressor
- The victim manifested her intent to end the relationship in the last six months
- The victim thinks that the aggressor may hurt badly or even kill her

Since the system is **actuarial**, the overall value is obtained by adding the weighted presence of an indicator. This means that all significant indicators add up to the final result, while the weight of each indicator varies according to the empirical data obtained. Thus, the theoretical scale for $VPR_{4.0}$, calculated by adding all the indicators, was (0 to 77.019) and the empirical scale (0 to 68.062). The authors estimated that a weighted sum increased

[15] AlertCop is a mobile application service provided by the Spanish Ministry of Interior that is designed to send security alerts with images or videos to the nearest emergency center, chat directly with a support agent or receive security news and notifications sent by public security services. Source: https://alertcops.ses.mir.es/mialertcops/

[16] The indicators were defined in Spanish and have been translated by the authors. Available at: López-Ossorio (2017, p. 193).

the predictive capacity of the model in 2 percentual points of AUC (López-Ossorio, 2017, pp. 193–194).

In order to classify each case, the thresholds for each category were set at:

Risk Class	$VPR_{4.0}$ Risk Intervals
Unappreciated Risk	0 ≤ x ≤ 9.353
Low Risk	9.353 < x ≤ 21.886
Medium Risk	21.886 < x ≤ 34.715
High Risk	34.715 < x ≤ 45.284
Extreme Risk	x > 45.284

Table 3. Risk thresholds for $VPR_{4.0}$. Source: López-Ossorio (2017, p. 194).

In order to establish the thresholds, the empirical scale was used according to three main criteria: the first threshold should be set so that the False Negative Rate was low – regardless of the increase in the false positive rate –. The majority of grave cases should be classified above "medium risk". Last, the overall clustering should be proportional to the resources available to implement the protection measures (López-Ossorio, 2017, p. 194). However, all of the victims who go through the VioGén system consider they are under a sufficient threat to file a police report, leaving those with lower risk scores in a vulnerable position, as it is further discussed later in this report.

The development of VPER-4.0, on the other hand, was conducted with significant methodological differences, using a retrospective design method, with the first recidivist assessments as cases, and the first periodic assessments as controls (López-Ossorio, 2017, p. 197). In this case, the theoretical scale for $VPER_{4.0}$ was (-11.257 to 73.018) and the empirical scale was (-11.257 to 25.080). The thresholds for each category were set at:

Risk Class	$VPER_{4.0}$ Risk Intervals
Unappreciated Risk	-11.257 ≤ x ≤ -3.087
Low Risk	-3.087 < x ≤ -2.185
Medium Risk	-2.185 < x ≤ 12.069

High Risk	12.069 < x ≤ 19.751
Extreme Risk	x > 19.751

Table 4: Risk thresholds for VPER$_{4.0}$. Source: López-Ossorio (2017, p. 198).

However, an alternative **machine learning** algorithm based on the Nearest Centroid technique for classification – or a hybrid model that implements a stochastic mix of the current system and Nearest Centroid – that would seemingly outperform VioGén has also been discussed (González-Prieto et al., 2021, p. 6). To support the ML approach, González Prieto et al. have suggested using a new evaluation metric called Police Protection, which can be obtained as the sum of the precision for the "inexistent / no" risk class, the F1 score for the "low" risk class, and the recall for the "high" risk class (González-Prieto et al., 2021, p. 6). This measure intends to ensure the right identification of risk for models that, for example, have good precision in the recidivism risk of class "high" but a bad recall. In those cases, it is possible to have the F1 score (i.e. the harmonic mean of precision and recall) within the range of admissible values, even if this means leaving most of the worst cases without protection, which would be inadmissible. However, the authors also identify two main **difficulties with the incorporation of machine learning into the VioGén system** that are intrinsic to the problem and stand in the way of a machine learning-based solution. The first one is the impossibility – for legal, ethical, and social reasons – to establish a control group in order to measure the impact of the model. The overall evolution of this type of crime can be observed throughout the years, but other components such as demographic changes, policy and law enforcement initiatives, and all the other variables that affect the incidence of IPV must also be accounted for. The second one is derived from the attempt to predict the risk of suffering gender violence itself. At first, it seems that the system measures relapses, and not a direct assessment of the risk (González-Prieto et al., 2021, p. 9). However, the historical data does not reflect recidivism itself, but rather rearrest and reconviction (Christian, 2020, p. 75). This means that the ML model - in a similar way that experts do – should learn how to classify the risk of victimization according to historical data on rearrest and reconviction, and not on recidivism or the actual risk. However, this problem is intrinsic to risk assessment, and the hybrid approach could facilitate a seamless transition from the actuarial to the ML-based approach.

4- Methodological Framework

This adversarial audit has been carried out through **multi-methods** research. In our experience with internal audits, we have learned that the best approach to understanding how algorithms work and impact on different groups is through the combination of quantitative and qualitative approaches. High-impact systems are always socio-technical systems, as the data comes from social and sociological processes and impacts on personal and social dynamics. Therefore, and as seen in Eticas' Algorithmic Auditing Guide (Eticas, 2021), any method or process designed to open up the black box of algorithms and AI will require to go beyond and above a purely technical analysis, which would not only be incomplete but also misleading.

The chosen methodology brings together a **statistical analysis** of IPH (Intimate Partner Homicide) cases to evaluate the predictive validity of VioGén's homicide risk assessment across different social groups of women, as well as **qualitative methods** that explore different perceptions of and experiences with the VioGén system. This approach has allowed us to maximize the use and contribution of the data available, both that coming from databases and the information we have derived from conducting interviews with different stakeholders.

Since Eticas' petition to access the original dataset used to build and validate VioGén was never granted, **the technical analysis has been based on the public record of 1,000 IPH victims** provided by the General Council of Judicial Power (CGPJ - Consejo General del Poder Judicial),[17] as this is the only source of data that is publicly available and allows us to approach the issues we are trying to tackle through the Adversarial Audit. The quantitative analysis has been designed to identify false negative rates and disparities in recall across the different strata (i.e., protected groups), in order to contextualize the predictive accuracy of VioGén when dealing with extreme risk cases. Recall is a metric that reflects the number of positive cases that are correctly classified. Thus, disparities across protected groups would indicate an underlying bias in the algorithm.

The dataset was first constrained to reflect exclusively those cases that happened between 2009 and 2019 (585 IPH cases) and, later on, restricted to cases from all of Spain except Catalunya and the Basque Country, where VioGén is not active, resulting in a total of 475 IPH cases. Since VioGén only acts upon reported cases of IPV, and the original data reflects both reported and unreported cases of IPH, the final sample is constituted by 126 reported cases of IPV that resulted in homicide. Each IPH case used includes information

[17] Available at: https://www.poderjudicial.es/cgpj/es/Temas/Violencia-domestica-y-de-genero/Actividad-del-Observatorio/Datos-estadisticos

about the place where the report was filed, the victim's and aggressor's nationality and age, their relationship, the cause of death, the number of children and underage children that the victim had, the aggressor's response, and whether there was a previous police report and a subsequent protection order. Finally, False Negatives were defined as any case of IPH with a previous police report yet lacking police protection, whereas Insufficient Protection was defined as any case of IPH with a previous police report that derived into some form of protection. The results of this analysis are described in the next section.

The **qualitative fieldwork** of this audit included semi-structured phone interviews and survey research with closed and open ended questions. **We interviewed 31 women who suffered from gender violence and went through the VioGén system**. All these interviews were conducted by an expert in the field of gender violence experienced in qualitative interviewing methods. Informed consent of the participants was obtained at the beginning of each interview. The interview questions were designed to inquire about women's experiences and perceptions during their journey in the VioGén system. We avoided questions that may create emotional distress and burden on the side of the participants and reminded them that they could skip any questions and/or withdraw from the study anytime. The interviews were recorded (audio only) and transcribed, but any personal information including the names of people and locations were anonymized to ensure confidentiality. We used the code of W (e.g. W1, W2, etc) to identify each victim in our analysis.

We reached out to the survivors of gender violence through the network of Ana Bella Foundation, which includes more than 27,000 survivors in their Network of Women Survivors (Red de Mujeres Supervivientes).[18] As a sampling criteria, we included women who went through the VioGén system between 2019 and 2021 and who reported the aggression in Andalucía, Valencia, Madrid, or Galicia, which are among the top-five regions in Spain with the most active VioGén cases.

Lawyers specialized in gender violence were also contacted through an online survey that included both open and closed-ended questions. **7 lawyers responded to our questionnaire**. **We also interviewed representatives of Ana Bella Foundation**, to get the perspective of the civil society on this issue. All the data collected during the fieldwork was analyzed through a thematic analysis to identify the emerging patterns and themes. For these expert interviews we followed the consent procedure mentioned above, and coded respondents with "L" for lawyers and "C" for civil society representatives.

	Number of participants	Interview code
Women survivors	31	W
Lawyers	7	L
Civil Society	2	C

[18] https://www.fundacionanabella.org/

Table 5: Distribution of fieldwork participants.

5- Auditing VioGén

Accessing the VioGén system

The VioGén risk assessment system can only be activated if the victim officially reports her aggressor to the police. Although by law all women in Spain have the right to report a gender-based aggression, in practice just a small percentage do. According to the Macrosurvey of Violence against Women (2019a), only 21.7% of women over 16 years old who suffered from gender violence did report their agressor. The remaining 78.3% of women did not report and therefore were not evaluated by the VioGén system. Our analysis of the CGPJ report[19] shows that, among 347 mortal victims of gender violence in Spain between 2009-2019, only 126 of them previously filed a police report. This means that **73% of women who were killed by their (ex)intimate partners did not previously report their aggressor and did not receive a VioGén risk score**.

Knowing this data, we set out to find out what were the factors that could be stopping women from reporting their violent partners, as this has a large influence on the efficiency of the VioGén system overall. There is a wide range of hurdles that can dissuade women from officially reporting their aggressor. Based on the results of the interviews conducted, we can group these barriers into three categories to more systematically analyze them:

- Individual emotional barriers
- Group-based structural barriers
- Institutional barriers

Individual emotional barriers against reporting gender-based aggression relate to the factors that are personal to the victim. They might exist regardless of victims' socio-economic status, education, age, race or ethnicity. Some of these individual emotional hurdles mentioned during our fieldwork are: not being conscious of the ongoing gender violence (denial), reluctance or fear to change the status quo, avoiding the emotional burden of filing a police report, fear of the aggressor, fear of societal judgment, feelings of shame, feelings of guilt, or hopes that the aggressor will change his behavior (C1 and C2). These barriers are very common and almost all victims face some of them.

Group-based structural barriers take place when a victim finds herself in a structurally disadvantaged position to report the aggression because of the group that she belongs to. While reporting intimate (ex)partner aggression is not easy for any women, there are some groups of women that encounter more structural challenges than others. These vulnerable

[19] Available at: https://www.poderjudicial.es/cgpj/es/Temas/Violencia-domestica-y-de-genero/Actividad-del-Observatorio/Datos-estadisticos

groups include but are not limited to women with small children, with few resources, with disabilities, living in rural areas, migrant women, and LGBT+ members and various intersections of these categories. In this report, we focused on three categories of vulnerability in order to have a closer look at the hurdles they face in their effort to officially report their aggressors. These groups are women with small children; socio-economically disadvantaged women; and undocumented migrant women.

- **Women with small children:** Not only women but also their children suffer from gender violence. The 2019 updates in the VioGén protocol intend to better capture children at risk of suffering the effects of gender violence. Between 2019 and 2021, there have been 7,008 active VioGén cases with children in a vulnerable situation and 2,376 cases with children in a situation of risk.[20] Victims of gender violence with underage children find themselves in a rough situation when the perpetrator is also the father of their children. Our fieldwork has shown that women in such a situation are highly concerned about the safety and wellbeing of their children when the perpetrator keeps his custody and visiting rights. As W17 said:

  > "the perpetrator is not a good father. He takes drugs, he does not even care about his own life. … He does vicarious violence to me and asks for shared custody. … He should not have the same rights as a good father. … He treated me badly even when I was pregnant with my daughter in my arms. The system fails."

  In some cases, **the fear of leaving the children alone with the perpetrator delayed their decision of filing a police report** against the aggressor and asking for a protection order. They pointed out how a possible restraining order (*orden de alejamiento*) decided by the court to protect women against their aggressor would also mean that women cannot be next to their children during the visiting hours of their father who is also the aggressor. As C1 said:

  > "many women do not report their aggressor because of the fear that after that the father will visit the children without her being there. When they file the complaint, they get separated from the aggressor, but what about the children?"

  Since January 2022, there has been an important change in the law, which has made it harder for perpetrators to keep their shared custody rights.[21] The new law states that *"[s]hared custody will not be considered if any of the parents is currently undergoing a criminal process for having attempted against the life, physical integrity, freedom, moral integrity or sexual freedom and liberty of the other partner or children*

[20] https://violenciagenero.igualdad.gob.es/violenciaEnCifras/boletines/boletinMensual/2021/docs/BE_diciembre__2021.pdf

[21] Law 17/2021, from December 15th, which amends the Civil Code, the Mortgage Law, and the Code of Civil Procedural, regarding the legal regime of animals («B.O.E.» 16 December); of art. 92.7 among others.

that live together."[22] While it is early to see the results of this legal change, it is still considered by the civil society as an important step in protection of women and children from gender-violence. However, women that we interviewed were not aware of this legal change.

- **Socio-economically disadvantaged women**: While victims of gender violence can come from all socio-economic classes, women with few resources find it particularly challenging to file a police report against their aggressor and seek a legal remedy. Economic dependence on the aggressor is an important handicap that hinders women from reporting their aggressor, even though there are designated social and economic subsidies for gender violence victims in Spain. For instance, and depending on their eligibility, victims can be granted various subsidies such as Active Inclusion Income (Renta Activa de Inserción-RAI), Minimum Inclusion Income (Renta Mínima de Inserción - RMI), and in some cases Minimum Living Income (Ingreso Mínimo Vital - IMV). They can have priority access to social housing and can be given special help to find a job. But all these measures require certain types of paperwork and can become available if and once the victim is provided the protection order by the court. Therefore, they are not often considered adequate to "take a leap out of an abusive relationship" (W 16).

  Education is another important factor that needs to be considered when assessing victims' access to the VioGén system. The role of education needs to be considered along with other interrelated factors such as employment opportunities, economic sources, timing of family formation, partner selection, and family attitudes to gender equality (Weitzman, 2018). The level of education has an important role in supporting women's bureaucratic literacy and their understanding of how/where to ask for formal help against the ongoing gender violence. In this respect, women with a lower level of education find it more challenging to file a complaint against their aggressor. As L7 explained: "the level of education makes it easier to understand what is being asked and to explain how she is feeling and what she is suffering of". Having said that, the experts highlight that professional women with higher levels of education also avoid reporting their aggressor, because they intend to protect their social reputation and professional careers. In the words of one civil society representative that we interviewed, "women without education wait 8 years and women with a doctorate wait 13 years to report their aggressor. In other words, "professional women with an education find it very difficult to report their aggressor, because they believe that this would question their professional career" (C 1).

- **Migrant Women:** Migrant women encounter specific challenges in the process of reporting gender-violence. These challenges include: being raised in cultures that lack the notion of gender equality; being enclosed within the aggressor's circle without having their own family and friend support network and/or economic

[22] Original article: "No procederá la Guarda conjunta cuando cualquiera de los padres esté incurso en un Proceso Penal iniciado por intentar atentar contra la vida, la integridad física, la libertad, la integridad moral o la libertad e indemnidad sexual del otro cónyuge o de los hijos que convivan con ambos." https://www.boe.es/eli/es/l/2021/12/15/17

> autonomy; not being fluent in host country languages; lacking the knowledge of laws, regulations, and official processes in Spain; and being in an irregular status for their stay in Spain and fearing deportation. As W3 explained to us that her irregular status blocked her from seeking help. Her aggressor constantly told her that she cannot do anything, because otherwise she gets deported. She said: "I didn't get in touch with any organizations, I didn't have any support and I was completely alone".
>
> The Spanish laws (Organic Law 4/2000, of 11 January, on the Rights and Freedoms of Foreigners in Spain and their social integration) offer special protection to irregular migrant women who are gender violence victims.[23] According to this law, if the presence of gender violence situation is confirmed by the court and protection measures are allocated, this opens the way for regularization of undocumented women's stay in Spain. But if the court does not confirm the gender violence situation and does not allocate the protection order, then the process of deportation is initiated. While the law provides an important protection for the undocumented gender violence victims, it also creates a major dilemma considering the fact that proving a gender violence case is not always straightforward and often requires properly collected evidence and good legal assistance to prepare and present the case before the court. The fact that VioGén classifies approximately 45% of cases as no-risk cases ("unappreciated"), validates the dilemma undocumented women may have to seek protection.

Institutional barriers: When a woman overcomes the previously discussed hurdles and decides to report her aggressor, the attitude of police officers plays an important role in her overall experience. Unwelcoming attitudes, lack of empathy and understanding, and judgemental behavior on the side of the police can even result in victims' leaving the police station without being able to officially report the aggression. Our fieldwork shows that women have very different experiences and perceptions of police officers while filing the aggression report. In one case, the victim described her experience with the police officers as "I am very satisfied, they took care of me and calmed me down" (W 16). In other cases, victims expressed their experience with the police as "unpleasant" (W 4), "passive and no empathy" (W 10), and "unprepared" (W21) to deal with gender violence cases. Three of the cases in our fieldwork showed how the same gender-violence case received different treatment by different police officers. W 13, who is a gender-violence victim of migrant-origin, had to travel to another city to be able to report the aggression by her partner. Only after traveling to another city, she could be evaluated by the VioGén system, and she received a high risk score.

> "In my first try, they did not accept my complaint and they told me to go back to my country and that it was all my fault. Later, in another city, they treated me very well". (W 13)

[23] https://violenciagenero.igualdad.gob.es/informacionUtil/extranjeras/proteccion/home.htm

In the other case, W14 had to wait 12 hours at the police station for the duty-shift, so that she could avoid the unwelcoming police officer she initially encountered and file the report with another officer.

> "I left without filing the complaint, and had to wait a long time. The attitude of the police officer was judgemental. I felt uncomfortable and guilty. … Horrible treatment and he called me a 'liar'. … I stayed there until the duty-shift. Another police officer attended me very well and I could file the *denuncia.* It took me 12 hours from 4pm to 4am to file the complaint*."* (W 14)

The fieldwork conducted for this Adversarial Audit points to the importance to assess technical systems also from the perspective of their users and even before the technical system intervenes, as what happens before and after the technical solution can have a big impact on participation rates, representativity and overall assessment of a particular problem.

The VioGén interview and questionnaire

When a woman officially reports a gender-violence case to the police (or guardia civil), **her case is activated and evaluated by the VPR questionnaire**. This questionnaire includes 35 risk indicators that are evaluated as "present" or "not present". Once the victim answers the questions covering these risk indicators, **the system assigns a risk score that assesses her likelihood of encountering future aggression by the same perpetuator**. In this way, the VioGén system promises to standardize the risk assessment process throughout the entire country (i.e. the same VPR is conducted in every single gender violence case supposedly in the same manner) and to provide an objective risk score that is, presumably, free from individual officers' biases and level of expertise. **The VioGén system is built on the assumption that women suffering from gender violence understand and respond clearly to all 35 items in the VPR form and the police officers objectively transform women's statements into binary answers (present/not present)** in the VPR form. But in reality, the process rarely works in this idealized way. In our fieldwork, **over 80% of the women interviewed reported different problems with the VioGén questionnaire.** This means that the quality of the data fed into the algorithmic system could be compromised during the input generation stage, resulting in possible sources of bias and misrepresentation within the system.

Some of the most salient problems identified during the fieldwork are listed below:

- **Lack of information**: In order for women to understand and answer the VioGén questionnaire properly, they must have adequate information on what the VioGén system is, how it works, what it does, and what is expected of them. The police officers/guardia civil have a key role in informing victims about the system before they start conducting the questionnaire. The gender violence survivors we interviewed, especially when this was their first time filing a complaint, stressed that they were not fully informed about the VioGén system. One interviewee stated that: "They (the police officers) were using the word VioGén, but I did not know what this word meant" (W 29). Another interviewee (W 28) thought that the VioGén risk score is something that calculates the meters of a restraining order. **35% of the women we interviewed were not informed about their VioGén risk score and therefore did not know what risk level the system assigned to them**.

- **Timing of the VioGén questionnaire**: The VioGén questionnaire is conducted at the moment of filing the police report for gender-based violence. Since many women suffering from gender violence arrive at the police station and file the complaint right after a violent incident, they find themselves in **a state of shock** and not physically or emotionally ready to provide accurate answers. Our interviewees mentioned how they found it challenging to recall all the past information, organize their thoughts, and provide thorough answers to the VioGén questions. In one of the cases, **W 25 had an anxiety attack during the VioGén questionnaire and was taken to the hospital. Another interviewee (W 23) described the moment of answering VioGén questions as "surreal"** and explained that at that moment, she didn't know what she was doing. For her, it was "a cloudy moment with absurd questions where errors are made while filling the questionnaire." (W 23)

- **VPR questions:** Our interviewees highlighted how even though some VioGén questions were clear and straightforward, others were "**ambiguous**" (W15) and did not make much sense in that context. Some lawyers find VPR questions to be "**rigid**", not allowing for explanations (L2), and "**generic**" (L4), lacking the capacity to adequately refer to individual situations. According to one of the lawyers we interviewed, cited above (L7), understanding and answering VPR questions highly depends on the victim's level of education.

- **Lack of legal support**: Women suffering from gender violence have the right to request a lawyer to file their complaint, but just a few of them are aware of this right (C1). Often, they are provided with a lawyer just before their case is heard by a court. Therefore, they answer the VioGén questions without being able to talk to a lawyer who would inform them about the process and what is expected of them.

- **Lack of psychological support**: One of the important issues that arises during the VioGén interview is the emotional distress it creates on the side of the victims. Not having psychological help and support before and during this process worsens this emotional burden, especially when the VPR questions are conducted by police officers that are not specialized in gender violence and lack adequate training on the VioGén system.

As with any data system, the quality of the inputted data is key to the quality, representativity and fairness of its results. The factors identified translate into **poor or biased inputs in 80% of the cases.** While we may not generalize from our small sample, the results obtained point to the urgent need to revise the conditions in which women access the system and the questionnaire, as acting at this stage is key to ensuring data quality. As mentioned above, **the quality of the data fed into the algorithmic system may be compromised during the input generation stage** (i.e., when women respond to the VPR risk indicators), **resulting in possible sources of bias and misrepresentation within the system**.

We have been able to complement this qualitative assessment with the **results of the data analysis conducted on the dataset of 1,000 homicide (IPH) cases** described above, which provides some insight into the issues at stake from a technical perspective. The VioGén questionnaire $VPR_{5.0}$ has been equipped with an additional protocol (the H-scale) designed to identify potential victims of IPH which, according to its developers, has a high predictive capacity.[24] However, and as the quantitative findings below suggest, IPH is a singular gender violence mechanism that, in most cases, is not the end result of an escalation of historical abuse, but rather a one-off event of extreme violence, making it more difficult to predict. In light of this, and despite the denial of Eticas' petition to access the original database of IPV (Intimate Partner Violence) in Spain from the ministry, a critical assessment of the H-scale has been performed based on the data gathered in the public record of homicide (IPH) victims released by the Spanish CGPJ (Consejo General del Poder Judicial).[25]

- **H-scale and homicide predictability**: The initial aim of the analysis was to identify possible sources of bias and variance in the prevalence of homicide (IPH) throughout different protected groups. As Eticas did not have access to the original database, we have used the available data to identify issues that point to potential problems that should be addressed in the VioGén system. The data from CGPJ only grants visibility on what we have defined as the "death zone" (see Figure 1). **The main objective was to identify recall variations across protected classes, as this would hint at problems of bias and representativity in the predictions made by VioGén**, in order to identify and discuss correction mechanisms. The data used to conduct the study captures the effects of the overall system (including the judge's decision) and not those of the algorithm exclusively. Thus, and for the results to be generalizable, this study should be extended onto the original database to identify and address potential discrimination patterns in VioGén.

Figure 2 is an abstract representation of the population of reported gender violence cases (i.e., the cases that are processed by the VioGén system). The left half of the picture represents the relevant elements (i.e., those cases that are at risk of suffering further violence), being false negatives (**FN**) the subset of cases predicted as negative but actually positive, and true positives (**TP**) the subset of positive cases classified as such. The right half of the picture represents the actual negative elements (i.e., those cases are not at risk of suffering further violence), being false positives (**FP**) the subset of cases predicted as

[24] The main performance metrics presented in (López-Ossorio, González-Álvarez, et al., 2020) claim a sensitivity of 84%, a specificity of 60%, an OR = 8.130, an AUC = .80, a PPV = .19 and a NPV = .97

[25] Available at: https://www.poderjudicial.es/cgpj/es/Temas/Violencia-domestica-y-de-genero/Actividad-del-Observatorio/Datos-estadisticos

positive but actually negative, and true negatives (**TN**) the subset of negative cases classified as such. This analysis focuses on the "death zone", a subset that exclusively includes the TP and FN cases that resulted in homicide.

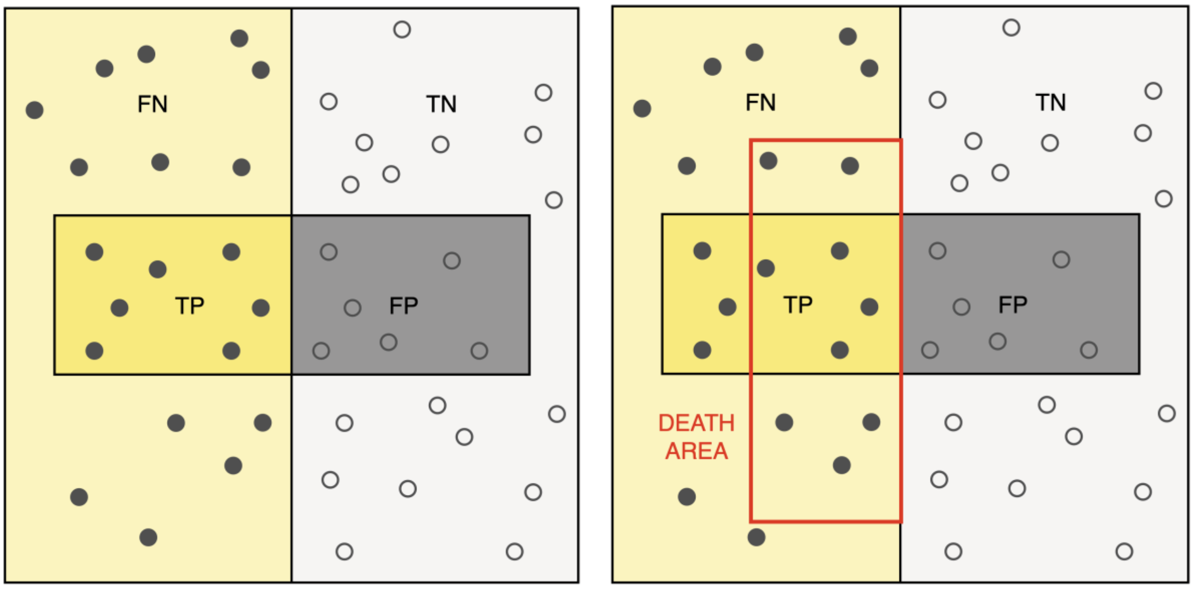


Figure 2: Abstract representation of the population (left) vs the visibility area due to data availability (right)[26]. The areas are not proportional to the actual population. Source: The authors

In order to conduct the analysis, after "cleaning" the data available, we segmented it into the following groups:

- False Negatives: this refers to cases where murdered women had previously filed a police report but did not receive a protection order as VioGén did not appreciate risk ("riesgo inapreciado") or predicted a low risk. 56% of the victims (N=126) fall into this category. Their distribution according to the groups provided by the CGPJ in the dataset is the following:

	Total FN (homicide)	
Victim's Nationality	ESP = 46	nESP = 25
Aggresor's Nationality	ESP = 45	nESP = 26
Children (C/nC) *	C = 52	nC = 15
Underage Children (mC/nmC)*	mC = 38	nmC = 29
Age (-54,+55) **	x≤54 = 61	x>54 = 9

* In four cases it was not specified whether the victim had children or not

[26] Note that there are true positive cases that fall outside the death area. This set is constituted by cases of IPV victims that were identified as being under risk, but recidivism was not in the form of IPH.

** One victim was underage and has not been included in the group x≤54

Table 6: Count of False Negative cases (i.e. previous police report without protective measures that resulted in homicide) calculated for each protected attribute, from the data released by CGPJ. Source: The Authors

- Insufficient protection: this refers to cases where women murdered had previously filed a police report and did receive a protection order. 44% of the victims (N=126) fall into this category. This class is a subset of true positive cases for which protective measures did not suffice to prevent homicide. Their distribution according to the groups provided by the CGPJ in the dataset is the following:

	Total IP (homicide)	
Victim's Nationality	ESP = 34	nESP = 21
Aggresor's Nationality	ESP = 36	nESP = 19
Children (C/nC) *	H = 48	nH = 6
Underage Children (mC/nmC)*	Hm = 30	nHm = 24
Age (-54,+55)	x≤54 = 50	x>54 = 5

* In one case it was not specified whether the victim had children or not

Table 7: Count of Insufficient Protection cases (i.e. previous police report with protective measures that resulted in homicide) calculated for each protected attribute, from the data released by CGPJ. Source: The Authors

The stratified counts of negative and insufficient protection cases allow the calculation of recall as a function of different protected attributes. Thus, by comparing these with the recall value over the whole population allows identifying bias and discrimination. Recall can be formulated as:

$$Recall = \frac{TP}{TP + FN}$$

From the available data, we observe a significant difference in recall regarding one particular dimension: having children vs not having children.[27] This means that those victims who did not have children were significantly perceived as being at lower risk. Prima facie, this is inconsistent with the configuration of the H-Scale in VPR$_{5.0}$, where none of the 13 predictors included reflect the presence or absence of children.

	Recall (homicide)	**|x1-x2|/rc**
Recall (TOTAL)	rc = 55 / (55+71) = 0.437	-

[27] Note that the ratio of IPH victims that did have children (regardless of whether there was a protection order in place or not) over those that did not have children was 5:1, and the size of the sample is relatively small.

Victim's Nationality	ESP = 0.425 (34/80)	nESP = 0.457 (21/46)	7%
Aggressor's Nationality	ESP = 0.444 (36/81)	nESP = 0.422 (19/45)	5%
Children (C/nC)	C = 0.48 (48/100)	nC = 0.285 (6/21)	**44.62%**
Underage Children (mC/nmC)	mC = 0.442 (30/68)	nmC = 0.453 (24/53)	2.5%
Age (-54,+55)	x≤54 = 0.450 (50/111)	x>54 = 0.357 (5/14)	21.28%

Table 8: Recall calculated for each protected attribute, from the data released by CGPJ. *Source:* The Authors

Intimate Partner Homicide (IPH) is a heinous crime, yet it constitutes a small percentage of the overall population of Intimate Partner Violence (IPV) victims. Taking, for instance, the first semester of 2021, the total number of police reports for IPV was 70.723, while the number of IPH cases in the same period amounted to 21 (roughly 0.03%).[28] **The main problem related to IPH, however, is not the size of this group, but the fact that most of the cases of IPH are not previously reported cases of IPV.** The data provided by CGPJ shows that the number of cases of IPH without previous police reports between 2009 and 2019 (constrained for the CC.AA. in Spain where VioGén is active) ascended to 347 (73%), whereas the number of IPH victims that had previously filed a police report ascended to 126 (27%). This means that **73% of homicide (IPH) victims during that period had not previously filed a police report, and thus a VPR or a VPER score was not assigned to them**. Moreover, if we consider all IPH victims for that period, **only 55 (11.6% over the total) received some form of police protection**. These findings are consistent with the Macrosurvey of Violence against Women, as it has been pointed out at the beginning of this section, hinting at a structural problem regarding the visibility on IPV that VioGén needs to account for. This reinforces the findings of our qualitative approach, described above in "Accessing VioGén", and point to the urgent need to work to remove or lessen existing barriers that stop women from seeking protection.

The predictive capacity of $VPR_{5.0}$-H is not challenged by these findings: previous studies show how the predictive validity of the VioGén tool has been improving, reaching an AUC (Area Under Curve) value of 0.80 in its newly designed homicide scale (H-scale).[29] Yet, **the existing research has evaluated the performance of the VioGén system at an aggregate level and has not studied whether the system performs differently for different groups of women according to their age, origin, or whether they have children or not. In this study, we have conducted a stratified analysis of IPH cases across different categories of women with the purpose of inquiring about the kinds of vulnerabilities that might be produced and enforced by the system**. In doing so, we have found that women without children are systematically assigned a lesser risk score than women with children. And the

[28] Source: La violencia de género en 10 indicadores 2021 (Primer semestre)

[29] The Area Under the Curve (AUC) is one of the most common means to assess predictive validity of risk assessment instruments. Its value can range between 0 (perfect negative prediction) and 1 (perfect positive prediction), with 0.50 indicating chance prediction (Douglas et al., 2005).

same should be done for VioGén, in order to better understand how it is calibrated and whether disparities across protected groups are accounted for.

Moreover, the fact that only 1 in 4 victims of homicide (IPH) enter the VioGén system via previous IPV reports raises difficult questions about VioGén's recall of extreme risk cases. One reason for the seemingly high predictive capacity of the H-scale could be that the indicators of extreme risk (especially within the H-scale) have been tailored to identify a particular subset of IPH cases, i.e. those that derive from a build-up in violence from a longer history of IPV, where the weight of other factors is sufficient to discriminate between extreme risk and lesser-risk labels. But these constitute 1 in 4. For the other 75% of cases, a more comprehensive analysis of the context and profiles should be conducted to trace back the causes of this lack of visibility, and a disclaimer about the validity of the H-scale over the total population of IPH victims should be made explicit: the high predictive accuracy of the H-scale only covers 25% of the IPH cases. **This means that the majority of IPH cases (75%) remain unaddressed by the new protocol VPR$_{5.0}$-H.**

Perception of the VioGén algorithm

While the predictive accuracy of the VioGén system is technically studied by its developers (López-Ossorio et al., 2017, 2019), there is less attention on perception of and trust in the system by the key stakeholders. The 2015 survey study by Gonzales et. al. (2015) mainly focuses on victim satisfaction with police performance, but not their perception of the VioGén algorithm and their assigned risk score. Our fieldwork and interviews with **gender violence survivors and their lawyers indicate some general distrust in how the VioGén algorithm works**.

- **48% of the women we interviewed negatively evaluated their experience with the system; 32% of them highlighted both negative and positive aspects, and only 19% of them positively evaluated their overall experience with the VioGén system.**
- **All lawyers that answered our questionnaire also had low trust in the VioGén system.**

While our sample is not representative of the broader population of survivors and lawyers and therefore our findings are not generalizable, our fieldwork raises important questions that need to be studied more systematically, and ideally addressed at the institutional level.

One of the main concerns about the VioGén algorithm is that approximately 45% of the cases receive the score of "unappreciated". In the context of gender violence, the category of "no risk" is already a very contested issue (Kropp, 2004). As it is mentioned by our respondents, the act of filing a police report against the aggressor itself is a risky behavior that would result in backlashes. When a victim receives an "unappreciated" or "low" risk score, this creates a wide gap between how she self-evaluates and perceives her risk of re-victimization and what the system predicts. As one lawyer interviewee explained: "Victims feel deceived once they file the complaint and see how they are not believed". (L5)

The other issue that was pointed out during our interviews was the perception of how the VioGén algorithm under-values psychological violence and perhaps newer forms of non-physical violence (such as stalking through social media), putting the emphasis on physical violence. In this regard, there is a general belief by both victims and their lawyers that the VioGén parameters do not adequately account for psychological violence. As L4 states: "It does not take a beating, nor an aggression, for the risk to exist. It seems that the parameters forgot the psychological abuse".

Perception issues are important as, ultimately, the trust in a system like VioGén will have an important effect on the quality of the data it receives from the women. As all data-intensive systems, the VioGén system relies on human data, and so if the data sources are scared, reluctant or hesitant, the quality of the data inputs will suffer. Our means and the data available do not allow us to go further in our analysis, but provide, in our opinion, a strong case for a systematic improvement of the weak points that our Adversarial Audit has identified.

Findings and recommendations

When we started the Adversarial Audit of VioGén, we had concerns around transparency, independent oversight, accountability, end-user engagement and the transition to ML. After conducting the audit, we can confirm that:

- **VioGén is not transparent.** We could not access any system data or information beyond what has been produced by experts involved in the definition of the system. Neither adversarial auditors nor women groups have any kind of access to the VioGén data. For a publicly-funded, high-impact system like VioGén, this is unacceptable.
- **VioGén has not been independently assessed or audited**. The publicly available resources and surveys regarding the validity and desirability of VioGén have been conducted by individuals who either work for or have vested interests in the ministry and police forces. External auditors or researchers have no official or public path to access the data, and access seems to be provided by the Ministry at their discretion.
- **VioGén is not accountable**. While the Ministry of the Interior sees VioGén as a recommender system, the high rates of prima facie acceptance of the algorithmic results (95%) points to an automated system, which should be held to further scrutiny as per the Régimen Jurídico de la Función Pública.
- **VioGén does not engage end-users**. In our fieldwork we have found that women and women organizations have never been approached about the system, neither in its design phase nor later on during the different decisions on how to alter the VioGén system. Also, we have found that 80% of the women who have used the system have negative

comments about it. They are not informed of what it does or how it works, which leads to distrust.

- **The VioGén transition to ML raises new questions**. Even though the literature explores the process of transitioning to a ML version for VioGén, the nature and extent of the collaboration between SAS and the ministry has not been publicly disclosed. While the lack of a public and open debate on this process would in itself be concerning, the fact that the technical evolution of the system is being decoupled from state of the art research and oversight is bound to lead to further problems.

The auditing process has also allowed us to go beyond our initial concerns to identify new issues that deserve attention.

Firstly, we want to highlight that through this audit, we have found that the VioGén system adapts the clustering of risk assessments to the resources available. This means that the system only gives the number of "extreme" risk scores it can afford, and so **funding cuts have a direct and quantifiable impact on the chances that women will receive effective protection after seeking police protection**. As the number of VioGén cases is growing each year, there are more women receiving police protection. While in 2015 around 3,000 women received police protection -with medium, high, and extreme risk scores-, in 2021 this number rose up to almost 9,000 women. Yet, there is still a big gap between women who receive police protection over those who do not, despite reporting the case of gender violence to the police. In terms of calibration,[30] **we are concerned by the number of cases that the VioGén system "discards" by giving them an "unappreciated" risk score**.. As it is currently designed, the risk score given by VioGén is not only determined by the objective facts that the questionnaire intends to unearth, but also the overall distribution of gender-violence cases, which is determined by the available resources. Therefore, in 2021, **only 1 out of 7 women who reached out to the police for protection actually received it**.[31]

This is even more serious if we take into account the barriers we have identified to accessing VioGén, which are one of the reasons why only 21.7% of women victims of domestic violence seek protection. These figures mean that **only 3% of the women who are victims of gender violence receive a risk score of "medium" or above and, therefore, effective police protection**.

Secondly, we have identified that not having children has a significant negative impact on how extreme risk cases are perceived. Our data analysis shows that **women who were killed by their partners and did not have children were systematically assigned lower risk scores than those who did, with a recall difference between groups of 44%**.

We would also like to call into question the representativity of the AUC value of the H-scale claimed by the lead researchers of VioGén. While it is true that with the data available

[30] Calibration error can be understood here as the difference between the predicted probabilities of the outcomes and the true probabilities of those outcomes.

[31] All numbers come from the Monthly Statistical Bulletin prepared by the Delegation of Government against Gender Violence.

the H-scale is capable of identifying extreme risk cases that can lead to homicide, **the fact that only 1 in 4 cases of homicide occur after a previous police report indicates how the majority of homicide victims will remain unprotected, even with the deployment of VPR$_{5.0}$-H**. This means that even though VioGén is now better equipped to identify certain cases of extreme risk, most homicide cases will remain unaddressed.

Fourthly, we have observed that **VioGén is, in practice, an automated system with minimal and inconsistent human oversight**. Police officers only increase the risk observed in 5% of cases, a figure that goes down when they feel overworked. This is highly problematic, as a non-accountable implementation of human oversight ("human in the loop") can lead to explainability and transparency problems. If police officers do not have clear instructions on when and how to intervene, their role can re-introduce bias into the system, and women may receive different scores depending on who files their case. Assessing the role of human oversight over time should be part of any audit and transparency efforts.

While our sample is not representative of the broader population of victims and lawyers and therefore our findings are not generalizable, our fieldwork raises important questions that need to be studied more systematically, and ideally addressed at the institutional level.

In light of the above, **we recommend**:

- Enacting policies aimed at **removing access barriers at individual, group-based, and institutional levels.** Women who suffer from gender violence must be able to report this aggression and seek official help. As there is a wide-rage of barriers against women's access to the VioGén system, the measures to alleviate these barriers must address the specific issues at each level and provide effective solutions.

- **Increasing the number of officers specialized in gender violence.** Police officers have an enormous impact on how the VioGén system works in practice, how it is experienced and perceived by the victims. Currently, 27.000 officers are involved in VioGén monitoring, but only 2.000 are specialized in gender violence.[32] There is an urgent need for increasing the number of police officers specialized in gender violence.

- When women actually access the VioGén questionnaire, they should receive **legal and psychological support**. Women suffering from gender violence often lack information about their legal rights and duties. While they have the right to ask for a lawyer, this legal help often comes at the trial stage. However, earlier legal support before and during the VioGén interview would help women to evaluate and present their situation better. Also, it is not easy for women to report against their (ex)intimate partners and answer the VioGén questionnaire. As the process is highly

[32] https://english.elpais.com/society/2021-11-26/what-life-is-like-for-the-women-and-children-in-spain-under-police-protection-due-to-gender-violence.html

emotionally charged and distressing, victims must be provided with psychological support early on, in some cases at the stage of filing their complaint.

- In order to address **accountability** issues, we recommend **accompanying the VioGén score with the justification of the police officers.** As the police officers conduct the VioGén protocol, they get a good grasp of the case and its contextual details, and they are the ones who approve the VioGén score. Their professional point of view matters and cannot be only expressed in a risk score value. Therefore, the VioGén risk score should be accompanied by a police report that justifies the score and provides further professional opinion if needed, both when they decide not to alter the score and when they use their discretionary powers to increase it. This would support the **accountability** of the Police for the risk assessment. It would also help the judiciary in their effort to understand and interpret the VioGén risk scores.

- As this report proves, there is great value in **auditing automated systems,** especially when they are publicly funded and have an impact on vulnerable communities like women victims of gender violence. The impact that an automated risk assessment system has on the victims of gender violence but also on society as a whole justifies enabling third-party audits. Doing so would not only confirm the proper functioning of the system, but would also allow independent researchers to identify possible harmful consequences of its use. Auditing algorithms is emerging as one of the practices that can ensure the accountability of technical systems, and promote trust. **Public institutions can and should lead the way in promoting responsible data practices**.

- **Using historical data to infer patterns of gender violence.** Given that VioGén relies upon a considerably large database to make its predictions, it would be advisable to contrast the actual configuration with advanced data analytics techniques, in order to validate the risk factors and identifiers used by the system to assess the risk. Moreover, these evaluations should be made available to the general public to foster transparency and trust in the system.

- If VioGén aims to **incorporate ML techniques,** this should be accompanied by a **public debate on the benefits and risks** of this. Academic research has shown that some ML techniques outperform the current design of VioGén (González-Prieto et al., 2021). In light of this, a space for public deliberation regarding the benefits and risks of migrating towards a ML approach – in its different forms – should be promoted from the ministry, where all stakeholders and third-party experts could discuss its desirability.

- Finally, it is urgent to **seek regular feedback from the victims and other stakeholders.** The performance evaluation of the system must not be limited to its technical analysis; but also needs to take into account the experiences and perceptions of stakeholders who go through or work with the system. The VioGén system cannot be only improved through scientific armchair contemplation, but also requires active fieldwork research, co-design methodologies and feedback mechanisms with women organizations as a way to explore ongoing and emerging vulnerabilities, and how the system is applied and experienced in practice. The **distrust in the system shown by women is alarming**.

6- Conclusions

The conclusions and recommendations emerging from the Adversarial Audit of the VioGén system have been highlighted above. In this conclusion section, we want to focus on the lessons learned on how to externally audit algorithms, its possibilities, limits and risks.

As mentioned in the Introduction, this is the first report from our Adversarial Audit project series, which will ultimately result in an Adversarial Audit Guide with practical recommendations on how to reverse engineer data and algorithmic systems. The VioGén case is a case of serious lack of transparency, where not only accessing the technical system is difficult (we would need to be women victims to access it), but also accessing those that have been impacted by it is laborious. Therefore, the auditing methods suggested by most authors, which focus on reverse engineering social media or internet services, were not useful. After a thorough review of the available sources for this case, which are mainly academic papers produced by experts involved in the design of the VioGén questionnaire and the CGPJ dataset, we decided to proceed with a mix-methods approach that we knew would not allow us to reach representative conclusions, but we hoped would allow us to ask the right questions. The **results described above have exceeded our expectations in terms of the possibilities of an external approach to algorithmic transparency**.

The lessons learned, which we will include in a forthcoming Adversarial Audit Guide, include:

- **Adjusting expectations to the available data**: no external assessment will be able to reach the representativeness of the conclusion reached by an internal auditing process. However, Adversarial Audits should be able to inform the right questions, and prompt systems developers to tackle the issues identified and address transparency issues. But while external exercises will never be conclusive, they can identify bias dynamics that may not be obvious during the design phase (in the case where one or more of the variables used in the algorithm is a proxy for a protected quality or group), provide an opportunity to involve end-users and impacted communities, and empower traditional CSOs to address

- **Analyzing the system end-to-end**: During the design, development, and validation stages of VioGén, neither the user experience nor the data collection processes were thoroughly evaluated. By means of this adversarial audit, however, many of the factors that alter the quality of the input data, as well as the overall experience of victims with the system have been identified and discussed. Thus, and in order to fully grasp the impact of an algorithmic system, analyses should transcend evaluation metrics, focusing also on design and deployment limitations that play a crucial role regarding the quality of the predictions such systems make – even more so with sensitive matters as in the case of VioGén.

- **Assessing the system via a multi-method approach**: An exhaustive assessment of algorithmic systems requires the use of multiple methods given the wide range of questions that need to be answered. While the technical assessment of the system requires the computation and critical discussion of varying evaluation metrics, which can be impaired by a lack of data availability, the exploration of how the system is experienced and perceived by different stakeholders can be done through qualitative research means including ethnographic fieldwork, interviews, and questionnaires. Combining both quantitative and qualitative methods facilitates a more comprehensive approach to the problem at hand.

- **Seeking alternative data to avoid proprietary barriers**. Most algorithmic systems cannot be directly evaluated due to proprietary limitations. Yet in this case, the statistical model was available but the original dataset from which the indicators and their weights were inferred was not. In this regard, resorting to alternative data sources such as CGPJ enabled an indirect impact assessment, identifying patterns that would need to be further evaluated within the original dataset.

- **Assessing the gap between the design and the experience context.** One of the main limitations of algorithmic systems to assess criminal risk that we have identified is the gap between the design and the application context. Failing to account for language, cultural, or socioeconomic barriers when designing the questionnaire can hinder the quality of the data that is collected, putting into question the validity of the predictions made by the system. In this regard, evaluating whether the design process included impacted stakeholders to increase its context-sensitivity can trigger a set of valuable research questions for the auditor.

7- Acknowledgements

We would like to thank Fundación Ana Bella for putting us in contact with the main stakeholders. Second, we would also like to thank the victims, lawyers, and members of the civil society whose invaluable responses have contributed to the findings and discussions presented in this report. Last, we want to thank Amparo Quiroga Albino for conducting the interviews with survivors of gender violence.

Project team: Adversarial Audits
Project Lead & Research Director: Dr. Gemma Galdon-Clavell, Founder of Eticas
Researchers:

- Dr. Evren Yalaz, Ethics & Technology Researcher at Eticas
- Toni Lorente, Ethics & Technology Researcher at Eticas

Recommended citation: "Eticas. (2022). The Adversarial Audit of the VioGén System. Association Eticas Research and Innovation."

8- Bibliography

Ada Lovelace Institute. (2021). *Technical methods for regulatory inspection of algorithmic systems*. Ada Lovelace Institute. https://www.adalovelaceinstitute.org/wp-content/uploads/2020/11/Inspecting-algorithms-in-social-media-platforms.pdf

Campbell, J. C., Sharps, P., & Glass, N. (2001). Risk assessment for intimate partner homicide. In *Clinical assessment of dangerousness: Empirical contributions* (pp. 136–157). Cambridge University Press. https://doi.org/10.1017/CBO9780511500015.009

Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values*. Norton & Company.

Delegación del Gobierno contra la Violencia de Género. (2019a). *Delegación del Gobierno contra la Violencia de Género*. Ministerio de Igualdad. https://violenciagenero.igualdad.gob.es/violenciaEnCifras/macroencuesta2015/Macroencuesta2019/home.htm

Delegación del Gobierno contra la Violencia de Género. (2019b). *Macroencuesta de Violencia Contra la Mujer*. Ministerio de Igualdad. https://violenciagenero.igualdad.gob.es/violenciaEnCifras/macroencuesta2015/

pdf/Macroencuesta_2019_estudio_investigacion.pdf

Douglas, K. S., Guy, L. S., Reeves, K. A., & Weir, J. (2005). *HCR-20 Violence Risk Assessment Scheme: Overview and Annotated Bibliography*. Implementation Science and Practice Advances Research Center Publications. https://escholarship.umassmed.edu/psych_cmhsr/335

Estévez Mendoza, L. M. (2020). Inteligencia artificial y violencia contra las mujeres: ¿funcionan los sistemas automatizados de evaluación del riesgo? *Perspectivas*, *3*, Article 3. https://revistas.ucalp.edu.ar/index.php/Perspectivas/article/view/148

Eticas. (2021). *Guide to Algorithmic Auditing*. Eticas Research and Consulting. https://www.eticasconsulting.com/resources/

González-Álvarez, J. L., & Garrido, M. J. (2015). Satisfacción de las víctimas de violencia de género con la actuación policial en España. Validación del Sistema VioGen. *Anuario de Psicologia Juridica*, *25*(1), 29–38. https://doi.org/10.1016/j.apj.2015.02.003

González-Álvarez, J. L., López-Ossorio, J., Urruela, C., & Díaz, M. (2018). Integral Monitoring System in Cases of Gender Violence-VioGén System. *Behavior & Law Journal*, *4*, 29–40. https://doi.org/10.47442/blj.v4.i1.56

González-Prieto, Á., Brú, A., Nuño, J. C., & González-Álvarez, J. L. (2021). Machine learning for risk assessment in gender-based crime. *ArXiv:2106.11847 [Cs, Stat]*. http://arxiv.org/abs/2106.11847

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293–323. https://doi.org/10.1037/1076-8971.2.2.293

Heilbrun, K., Yasuhara, K., & Shah, S. (2011). Violence Risk Assessment Tools: Overview and Critical Analysis. In R. K. Otto & K. S. Douglas (Eds.), *Handbook of Violence Risk Assessment* (pp. 1–18). Routledge.

Kropp, P. R. (2004). Some Questions Regarding Spousal Assault Risk Assessment. *Violence Against Women*, *10*(6), 676–697. https://doi.org/10.1177/1077801204265019

López-Ossorio, J. J. (2017). *Construcción y validación de los formularios de valoración policial del riesgo de reincidencia y violencia grave contra la pareja (VPR4.0-VPER4.0) del Ministerio del Interior de España* [Universidad Autónoma de Madrid, Facultad de Psicología, Departamento de Psicología Biológica y de la Salud]. http://hdl.handle.net/10486/678510

López-Ossorio, J. J. (2020, June 19). *The VioGen System* [Espai Societat Oberta]. Predecir el futuro: herramientas predictivas y sociedades segmentadas. https://www.youtube.com/watch?v=b7ytXGfKAUg

López-Ossorio, J. J., González, J., Buquerín, S., Garcia, L., & Buela-Casal, G. (2017). Risk factors related to intimate partner violence police recidivism in Spain. *International Journal of Clinical and Health Psychology*, *17*. https://doi.org/10.1016/j.ijchp.2016.12.001

López-Ossorio, J. J., González, J., Muñoz Vicente, J., Urruela, C., & andres-pueyo, A. (2019). Validation and Calibration of the Spanish Police Intimate Partner Violence Risk Assessment System (VioGén). *Journal of Police and Criminal Psychology*, *34*, 1–11. https://doi.org/10.1007/s11896-019-09322-9

López-Ossorio, J. J., González-Álvarez, J. L., Loinaz, I., Martínez-Martínez, A., & Pineda, D. (2020). Intimate partner homicide risk assessment by police in Spain: The dual protocol VPR5.0-H. *Psychosocial Intervention*, *30*(1), 47–55. https://doi.org/10.5093/pi2020a16

López-Ossorio, J. J., Muñoz Vicente, J., Andrés-Pueyo, A., & Pastor Bravo, M. (2020). *Guía de Aplicación del Formulario VFR5.0-H en la Valoración Forense del Riesgo*. Gobierno de España, Ministerio del Interior.

Pinedo, M. (2021, September 2). Matemáticas e inteligencia artificial contra el maltrato machista. *El País*. https://elpais.com/sociedad/2021-09-02/matematicas-e-inteligencia-artificial-contra-el-maltrato-machista.html

Riggs, D. S., Caulfield, M. B., & Street, A. E. (2000). Risk for domestic violence: Factors associated with perpetration and victimization. *Journal of Clinical Psychology*, *56*(10), 1289–1316. https://doi.org/10.1002/1097-4679(200010)56:10<1289::AID-JCLP4>3.0.CO;2-Z

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*, 23. http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf

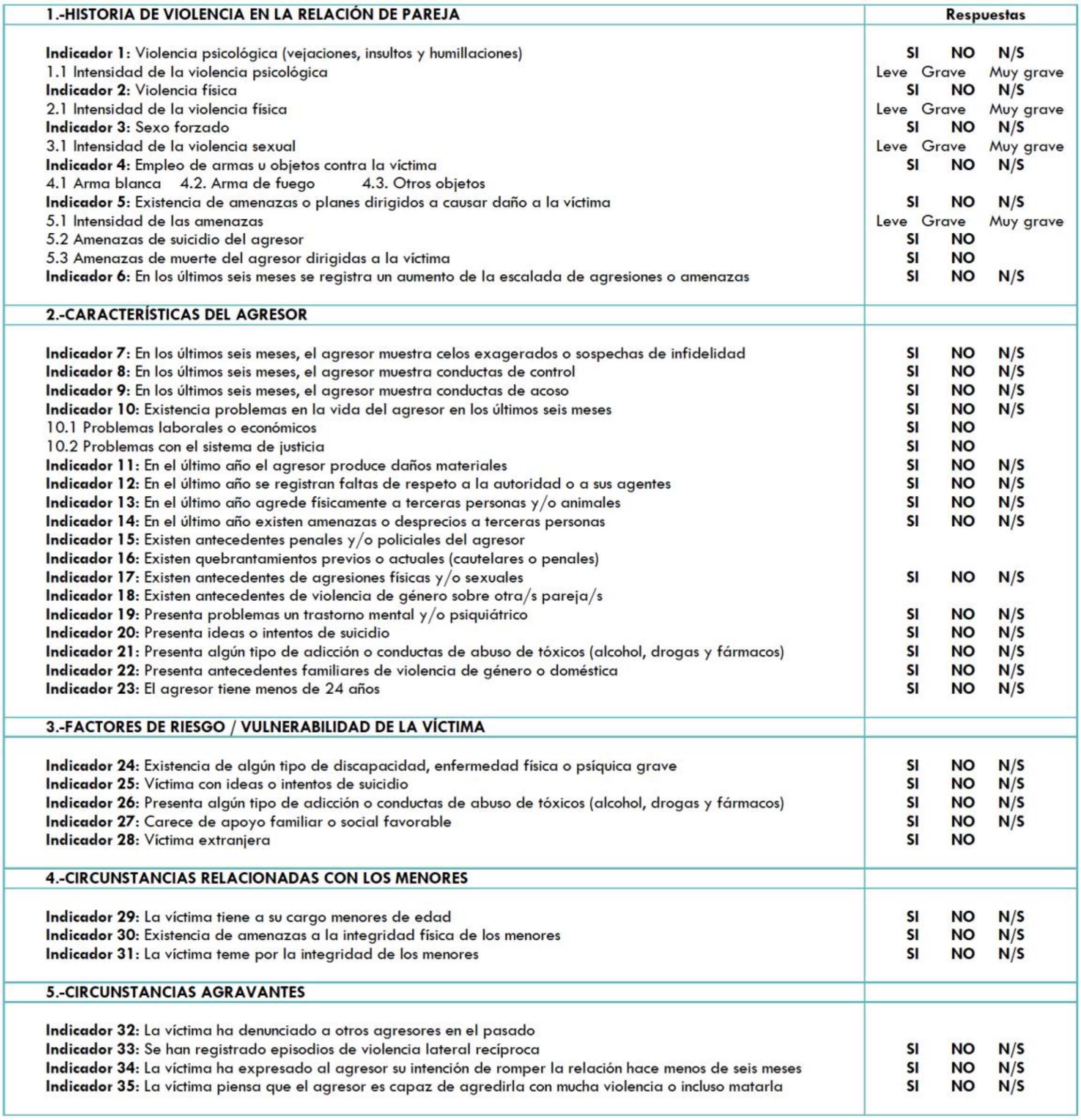
Sarker, I. H. (2021). Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. *SN Computer Science*, *2*(5), 377. https://doi.org/10.1007/s42979-021-00765-8

Tolan, S., Miron, M., Gómez, E., & Castillo, C. (2019). Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, 83–92. https://doi.org/10.1145/3322640.3326705

Weitzman, A. (2018). Does Increasing Women's Education Reduce Their Risk of Intimate Partner Violence? Evidence from an Education Policy Reform. *Criminology : An Interdisciplinary Journal*, *56*(3), 574–607. https://doi.org/10.1111/1745-9125.12181

Zurita Bayona, J. (2014). *Violencia contra la mujer: Marco histórico evolutivo y predicción del nivel de riesgo* [Universidad Autónoma de Madrid, Facultad de Psicología, Departamento de Psicología Biológica y de la Salud]. https://repositorio.uam.es/handle/10486/661810

9- Annex - The VPR$_{5.0}$ Protocol



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

Adversarial Audit