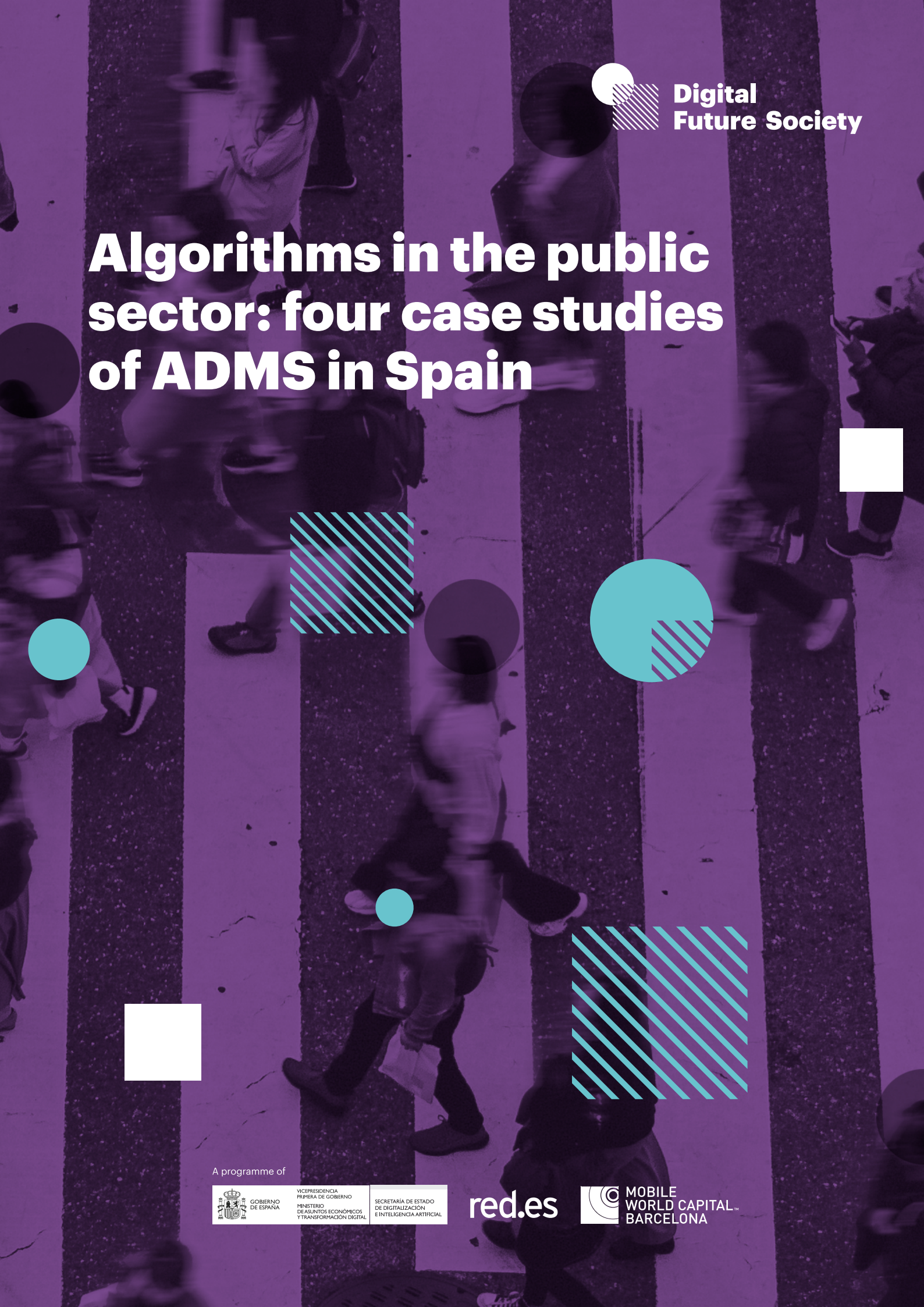


Digital Future Society

Algorithms in the public sector: four case studies of ADMS in Spain

A programme of

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA PRIMERA DE GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

red.es

MOBILE WORLD CAPITAL BARCELONA

About Digital Future Society

Digital Future Society is a non-profit transnational initiative that engages policymakers, civic society organisations, academic experts and entrepreneurs from around the world to explore, experiment and explain how technologies can be designed, used and governed in ways that create the conditions for a more inclusive and equitable society.

Our aim is to help policymakers identify, understand and prioritise key challenges and opportunities now and in the next ten years in the areas of public innovation, digital trust and equitable growth.

Visit digitalfuturesociety.com to learn more

A programme of

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA PRIMERA DE GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

red.es

MOBILE WORLD CAPITAL BARCELONA

Published
January 2023

Disclaimer
The information and views set out in this report do not necessarily reflect the official opinion of Mobile World Capital Foundation. The Foundation does not guarantee the accuracy of the data included in this report. Neither the Foundation nor any person acting on the Foundation's behalf may be held responsible for the use which may be made of the information contained herein.

Digital Future Society

Table of contents

Digital Future Society

Introduction

Artificial Intelligence (AI) has been hailed by some as a magical solution to complex, societal problems, while for others AI is a tool used to justify the increased surveillance and datafication of our societies today. Against this backdrop, governments are rushing towards implementing data driven technologies such as automated decision-making systems (ADMS) to aid the public sector in tasks such as optimising internal processes, anticipating risk and allocating resources.

The use of AI by public administrations is nuanced and complex. As the use of data-driven tools to aid in governance becomes increasingly widespread, it is critical that public administrations and citizens alike understand the **opportunities**, **challenges** and **risks** that their use entails.

Civil society organisations, academia and regulators have all cautioned on the ethical development and use of AI. Following the COVID-19 pandemic, a holistic approach to the digitalisation of the public sector is even more pressing given that countries in mid-recovery may be blindsided by the urgent need to adopt digital solutions.

A promise for a tech-driven public sector that is more efficient and responsive has been a common narrative in the modernisation of administrations around the world.

Following this narrative, in its 2025 Digital Strategy, the Spanish government highlights the public sector as one of the focus areas where it will expedite digitalisation. According to the strategy, emerging technologies, such as artificial intelligence, are key to improving the efficiency and quality of its services and ensuring that its digitally enabled citizens can securely and safely interact with the public administration. Spain has destined 4 billion Euros of the Next Generation EU funding for Spain's National Recovery and Resilience Plan (NRRP) to reform and invest in creating "an administration for the 21st century" (Ministerio de Asuntos Economicos y Transición Digital, n.d.; Mileusnic, 2022).

On the other hand, AI, algorithms and automation have increasingly become part of the daily vocabulary of the general public. As use of AI increases, there is a growing need to understand exactly how the digitalisation of public services and the use of automation can have a direct impact on peoples' lives – a wrong tick on an online application can exclude a group from getting the aid they are eligible for, a miscalculation in a police assessment can be a deciding factor of life or death, an unnoticed alert may result in misuse of taxpayer money. **The more societies delegate decisions to algorithms, the more we need to understand how these decisions are made and what information is used to make them.**

This report builds on previous work of Digital Future Society on automated decision-making systems (ADMS). It continues the effort to understand the social impact that derives from their use. Moreover, it seeks to bridge the knowledge from different experts from academia, the public sector and civil society organisations to illustrate the complexities behind each tool.

The report explores four case studies of ADMS used by the Spanish public sector that have drawn attention in the media. They are:

BOSCO

– Determines eligibility for a nationwide energy bill subsidy

RisCanvi

– Calculates the risk of inmate recidivism in Catalonia

VioGén

– Predicts the risk of gender violence for the national police force

SALER

– Anticipates potential cases of corruption in Valencia

About this report

In 2020, the Digital Future Society Think Tank explored the public sector's use of automated decision-making systems (ADMS) through reports on digital welfare and its implications for gender equality and a whitepaper on the risks and opportunities of AI in the public sector. (Digital Future Society 2020a, 2020b). Along the same lines, the Think Tank produced a podcast in Spanish titled Algoritmos y Gobiernos (Algorithms and Governments) which covered four different tools used in Spain's public administrations.[1] The podcast's mission was to raise awareness of the public sector adoption of ADMS and illustrate the complexities behind their design, governance and use. It features interviews with researchers, public servants, lawyers, civil society organisations and citizens.

This report is based on the content of the podcast, the case studies use material from the recorded interviews, along with additional desk research. Moreover, this report produced in both English and Spanish, seeks to bridge the knowledge and language gap, as research on ADMS and their impact is still incipient in Spain and abroad. This report sheds light on the specific context in Spain, drawing insights from the different actors involved in the implementation of ADMS with a focus on applications that have a direct impact on the welfare of citizens.

It also seeks to bridge the knowledge between Spanish and English-speaking communities. Given the lack of published content about these systems, there is a need to share and learn about what is happening in Spain and to contribute to the body of knowledge being generated on this topic.

Why now?

The pandemic caused by COVID-19 has not only accelerated the digitalisation of all aspects of society, it has revealed the extent of the digital divide. Governments are being encouraged to strengthen their digital strategies and enhance the use of digital technologies. This is reflected in the Next Generation EU funds, which are intended to assist EU member states in their recovery to be "digital, greener and more resilient" for the challenges to come (Cedefop 2021).

[1] See link for more information: https://digitalfuturesociety.com/algoritmos-gobiernos-podcast/

At the same time, Europe is setting regulatory precedence with its proposed AI Act – to ensure human-centric and trustworthy AI. While the act is being negotiated in parliament, Spain has plans to pilot an AI regulatory sandbox to test requirements for high-risk AI and is in the parallel process of creating a monitoring body, the Agencia Española de Supervisión de Inteligencia Artificial (Spanish AI Supervisory Agency).

Timely measures, considering that the AI Act is part of a larger regulatory trend. Globally, the Act has received mixed reception. While some applaud the European Union's efforts, others deem that the act does not go far enough to protect fundamental rights – 119 organisations, for example, have criticised the AI Act's risk-based approach, stating it does not adequately address the human rights concerns that arise with AI systems. (European Digital Rights, 2021)

As large players make strides to shape AI, it is critical that citizens understand and learn from individual AI cases. This is a crucial moment in which different stakeholders have the opportunity to ensure that AI technologies account for human rights (European Digital Rights, 2021).

Audience

Lastly, this report is for a general audience, especially those interested in learning about how algorithmic tools are being used in the public sector. The report should also prove to be a useful source for policymakers, civil society organisations, activists and any other stakeholder studying the topic and interested in gaining a general understanding of these tools from a social perspective.

Digital Future Society

Case studies

This report showcases four automated decision-making systems in use in different public administrations throughout Spain. **They are not representative of the diverse types of applications in use. AI systems have a wide variety of applications – in fields such as healthcare, finance, education, etc. The cases chosen for the report belong to the areas of social services, policing and administration.** They were selected because all these cases have garnered media attention for different reasons, some negative, such as questionable data practices, lack of transparency and their potential to exacerbate existing inequalities. But also, for their arguable potential to promote efficiency and objectivity. They are:

Case Nº1

BOSCO

– Determines eligibility to an energy bill subsidy nation-wide

Case Nº2

RisCanvi

– Calculates the risk of recidivism of inmates in Catalonia

Case Nº3

VioGén

– Predicts the risk of gender-based violence for the national police force

Case Nº4

SALER

– Anticipates potential cases of corruption in Valencia

RISK ASSESSMENT INSTRUMENTS

This report is intended to offer a global view of each case study, drawing on the main insights of researchers, public servants and users. It is not intended to provide exhaustive information on the workings of each system, but rather to shed light on the big issues surrounding the cases.

Each case study can be read as a stand-alone piece. Each is structured to offer a general understanding of how the system works followed by discussion and analysis. Some applications are more complex than others, such as the risk assessment tools, which, due to the extensive literature on the topic and the complexity of these systems, are slightly longer than the two systems, BOSCO and SALER. At the end of the report, the section “looking ahead” provides an overview of the four transversal themes that arise across the four case studies.

Case Nº1: **BOSCO**

– Determines eligibility for a nationwide energy bill subsidy

Context

In response to the economic crisis, in 2009, the Spanish government introduced an energy bill subsidy for low-income households. Some years later, on review of the benefit, the government imposed tighter restrictions on eligibility. In parallel, it implemented the use of software called BOSCO to review the applicants and decide whether they fit the newly set criteria (Kayser-Bril 2019).

What is BOSCO?

BOSCO is a software program that was created by the public administration in 2017 to assess whether a user is entitled to the energy bill subsidy. The energy companies administer the program and communicate directly with the applicants, not the government.

Who is eligible?

The deciding factor determining whether someone is eligible for the subsidy is their income. This is the base factor; there are other "amplifying" factors that may condition the amount of the subsidy – such as whether the beneficiary has been or is a victim of domestic violence, or whether they have a disability. Registered large families (*familias numerosas*) with three or more children are automatically eligible, regardless of their income. So are some pensioners such as those receiving retirement or disability pensions, but not those on a widow's pension.

Contention

In 2018, over half a million families' applications were rejected. This called the attention of Civio, a non-profit organisation, whose primary mission is to hold the public administration accountable. Once they learned of the mass rejection of applicants, they took on the task of helping claimants navigate the application process.

Civio created an assistant to help applicants determine whether they were eligible for the subsidy. To do so, they used the decree-law the BOSCO software was based on to translate the legal requirements into code. In creating this tool they realised how complicated it was to translate law into code because the law itself has certain ambiguities.

Once Civio had deployed its own tool, many applicants found that they were eligible in Civio's interpretation but were denied aid under the official tool. Alarm bells rang as Civio received many calls about this contradiction. When Civio finally got access to some of the technical information, they discovered that there were errors that needed to be corrected so beneficiaries could get the aid they were eligible for.

BOSCO

To create this assistant, Civio requested the documentation used from the government. The request went through the ministries involved and the Consejo de Transparencia y Buen Gobierno (Council of Transparency and Good Governance) (CTBG). The CTBG provided the technical information and use cases but it did not provide the source code. Without the source code, Civio was unable to determine whether the tool was faulty from its foundations.

Civio's director, Eva Belmonte, requested the source code to determine the basis of the errors; however, the request was denied on copyright grounds. According to Civio's lawyer, Javier de la Cueva, the current interpretation of the law allows the public administration to develop opaque algorithms, which has led Civio to appeal the denial in court (Kayser-Bril 2019).

Belmonte explains: "They gave us what was supposed to be the functionality of the application, and we discovered that some of the mistakes we had detected when talking to people who applied for the subsidy were real, such as the fact that widows were told they were not entitled to it even though they were eligible." Technically those on a widow's pension do not qualify, though if they apply under the income criteria, they meet the low-income threshold. "We don't know if there are more bugs, because we don't have access to the code," she says (Digital Future Society 2022c, 00:11:30).

In June 2022, the case took a turn as the Transparency Council changed its stance and agreed to disclose the source code. Civio is currently awaiting the final ruling in the case.

For Civio, this case is important in the long term as it will set a precedent as to how the Spanish administration will address future cases regarding the transparency of ADMS. Given that BOSCO is relatively simple compared with the more complicated systems expected to grow in use in the future, such as those based on machine learning, this causes great concern for Civio. As the use of ADMS grows in the public sector, Civio's concerns relate to how civil society can anticipate their use and defend citizen's rights should there be any future infringement.

The application process

Once the applicant finds out they are eligible for the subsidy, they can apply through the various power companies listed as distributors on the Spanish government website. Applicants can request the subsidy by phone, email or fax. The process consists of filling out a form and providing the following documents: IDs of household members, the energy contract and proof of residence (other documentation such as proof of disability or vulnerability are also included if applicable). This documentation is sent, the applicant receives an automatic reply and within 15 days they should get a definite reply on eligibility by email or post.

BOSCO

User experience: Mercedes and her caseworker Nerea

Mercedes is a single working mother of two with one child with a disability. She earns 1000 euros per month and pays 550 in rent as well as repayments on a loan she took out to pay the deposit on her flat. "If I have to pay the power and the water bills too, I won't have enough to buy food," she says. (Digital Future Society 2022c, 00:00:35).

Mercedes meets the requirements to receive the energy bill subsidy, but only found out through Insercoop. A non-profit based in L'Hospitalet, Spain, Insercoop helps vulnerable people find gainful employment and assist their clients with bureaucratic tasks, especially ones that require digital skills to receive financial aid. Insercoop plays a fundamental role in informing their public about different forms of aid, as many are unaware that they could be candidates for assistance.

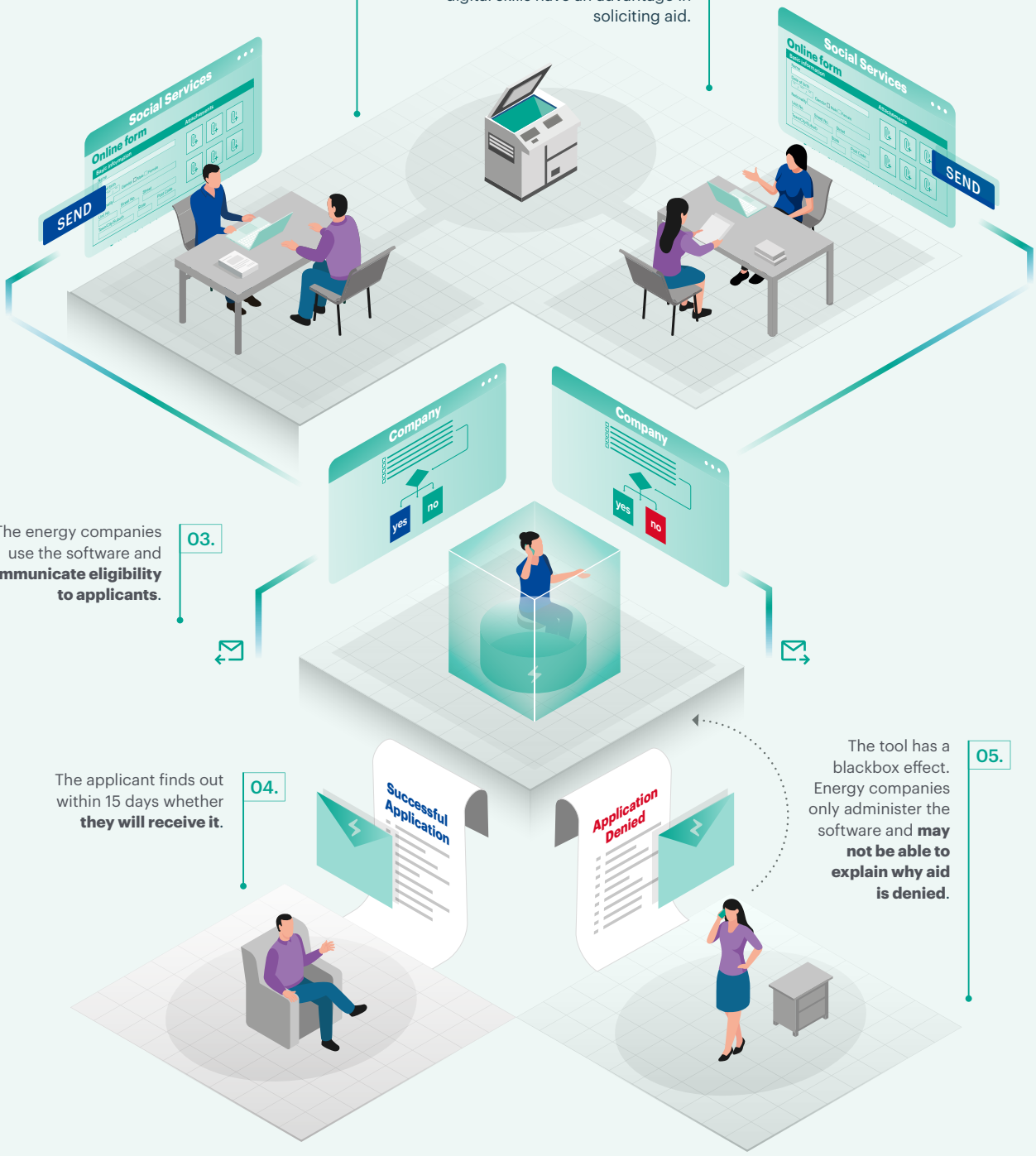
Mercedes' caseworker Nerea helped her fill out her application and send it to the energy company. But they didn't receive the usual response after the set 15 days. Nerea called and was informed that there was documentation missing from the application and they were unable to process it. Mercedes received a letter by post days after the call, informing her there was missing documentation. Mercedes had to wait over three weeks to receive a final answer.

Without Nerea, Mercedes explains that she would feel lost and overwhelmed by the application process. Apart from lacking the digital skills required to apply for the benefits, many services are difficult to navigate, Nerea attests. Those with digital skills, such as sending emails, uploading scanned documents, have an advantage in soliciting aid, as phone assistance is a tedious and time-consuming process – something Mercedes cannot afford.

BOSCO is a computer program used by electricity companies.
It has been developed by the government to evaluate the applications for the rate subsidy.

01. Once the applicant finds out that they are eligible for the subsidy, **they can apply through various power companies**.

02. The process consists of filling out a form and submitting documents. **Beneficiaries at times apply with the help of a social worker**, given that the process is difficult for some to navigate. Those with digital skills have an advantage in soliciting aid.



03. The energy companies use the software and **communicate eligibility to applicants**.

04. The applicant finds out within 15 days whether **they will receive it**.

05. The tool has a blackbox effect. Energy companies only administer the software and **may not be able to explain why aid is denied**.

Image source: Digital Future Society.

Discussion

Digital welfare and its impact

One of the biggest critiques regarding the 2017 reform (the change to eligibility requirements and implementation of the BOSCO software), is the high number of people denied aid. According to official statistics, in 2021, 1.1 million consumers received the subsidy, which is one third of those who received it in 2010 (3 million), before the reform was enacted.

As mentioned, this gap is in part due to the government's change in eligibility criteria. But Civio claims there are many more reasons why people on a low-income are being rejected. On the one hand, a substantial part of the eligible population is unaware, or finds the online application system overly difficult to navigate or incomprehensible (as in Mercedes' case). And on the other, the software is erroneously rejecting applications.

Cases like BOSCO and their impact on vulnerable populations, is part of a larger trend of digitalising the welfare state. One of the most vocal critiques of digitalised welfare is Philip Alston, former special UN rapporteur on extreme poverty and human rights. In his 2019 report, he exposes the seemingly benign initiative of the digitalisation of welfare systems, as an opportunity to reduce the welfare budget, which involves "a narrowing of the beneficiary pool, the elimination of some services... and **a complete reversal of the traditional notion that the State should be accountable to the individual**"(Alston, 2019).

A blackbox within a blackbox

As mentioned, energy (private) companies administer the software and communicate eligibility to applicants. This adds a layer of complexity for the end user – the beneficiary – who is not in direct contact with the public administration. Such diffuse liability presents another obstacle to contest the automated decision. Energy companies do not have a vested interest in assuring the subsidy is correctly attributed – they are just following orders, and applying the rules of eligibility laid out by the software. If a user who meets the eligibility criteria wants to know why they have been excluded there is no one to ask and no way to find out why.

This effect produces a blackbox within a blackbox. Energy companies are limited to the reasoning behind why a subsidy has been denied, they may know whether there is a lack of documentation but cannot go further to understand if the ineligibility is due to an administrative error, or whether the claimant does not meet the criteria. Because they do not own the software, only administer it, they have no agency to understand why aid was denied. Nor do citizens have direct access to the administration to contest their request for aid.

A fight for transparency

Although the BOSCO software was developed internally by the Spanish administration, the algorithms behind such systems are often bought from private sector companies and are therefore proprietary (AI Now Institute, 2018). Most often, source code is not shared on the

grounds of trade secrets. In this case, the administration's reluctance to share the source code under intellectual property rights is even more surprising, given that the public administration developed the tool. Whether these systems are proprietary or not, reluctance to share the source code is, unfortunately, a common challenge for external actors like Civio. It is also a significant hurdle to be able to evaluate the system and identify errors in software design or implementation.

Without the Civio platform, the many families that were denied aid wouldn't have understood that their rejection could have been due to a software error. Civio's case uses Act 19/2013 of 9 December on transparency, access to public information and good governance, to argue that the beneficiaries have the right to understand how automated decisions are used to determine access to services. Some even argue that the General Data Protection Regulation (GDPR) requires that the information be presented in a way that recipients can understand (Selbst and Powles, 2017).

When a public sector algorithm is brought to court (other models abroad)

The larger trend in digitalising welfare systems, can take on more sophisticated forms, such as tools to detect benefit fraud. One such tool is the Netherlands' infamous Fraud Signaling Facility (FSV), a risk profiling system, which led to the resignation of the Rutte administration because of its infringement of the GDPR in early 2021.

For years, the Dutch government had implemented an algorithm that was designed to create risk profiles of residents who were most likely to commit childcare benefit fraud. The scandal began in 2012, and up to 26,000 parents were falsely identified of fraud; many families were required to pay back large sums, amounting to tens of thousands of euros. Adding to the scandal, 11,000 dual-national families were under special scrutiny, as the algorithm illegally used data such as the claimants' nationality. (Holligan 2021).

In the Netherlands, automation has drastically changed the way that civil servants work. Marlies van Eck, an assistant professor at Radboud University, alleges that before the use of automation, benefit payments would go through extensive review. The increased use of ADMS, has in a way, substituted frontline worker discretion (Geiger 2021; *NL Times* 2021).

The lack of transparency persists on an institutional level. Early in 2021, the Dutch tax office sent letters to 60,000 people stating that it shut down the FSV system due to its breach of data laws. Some of those who received the letter were unaware they were on the list or why.

Risk Assessment Instruments: an explanatory note

Before delving into case studies two and three, this section provides a brief overview of risk assessment instruments, the role they play within the field of criminal justice, and how they are used. This context will provide the necessary background information to the discussion points of both cases.

What are Risk Assessment Instruments?

A risk assessment instrument (RAI) is a type of algorithmic tool used to predict future risk. In criminal justice, RAIs are used to predict a defendant's future risk of misconduct and are commonly used for pre-trial judicial decisions. This application of RAIs became commonplace in the 1970s in countries like the United States and Canada, where clinical assessments based on professional judgement had been perceived as highly subjective. These assessments that rely on clinical expertise, formally called unstructured clinical assessments, are carried out by experts, such as psychiatrists or psychologists, to aid judges in sentencing and other pre-trial decisions. RAIs, in contrast to clinical assessments, provide structured, evidence-based prediction and were introduced to reduce discretion and increase objectivity (Marco Francia, 2020).

One of the most common RAIs are Actuarial Risk Assessment Instruments (ARAI). These tools are based on statistical evaluation of pre-defined risk factors. Supporters of these tools advocate for their potential in making decisions more consistent, accurate and transparent. Nonetheless, the promised objectivity of these assessments is still open to debate (Heilbrun et al., 1999). Those who caution their use are concerned about their accuracy in prediction, lack of individualisation (something that clinical assessments do provide), bias and absence of transparency (Silver and Miller, 2002). Not only are there concerns about whether these tools are as accurate and fair as human operators, they are also controversial as they have exhibited biases against race and gender, as in the case of the UK's Offender Assessment System (OASys) (Angwin et al., 2016). The OASys, a tool comparable to the tool used in Spanish criminal justice, RisCanvi, was found to generate different predictions for race, gender and age (Big Brother Watch 2020). Moreover, these tools, like Northpointe's COMPAS, have been the centre of controversy in the United States, given their proprietary nature and absence of transparency under trade secret claims.

Actuarial tools, like those mentioned above, contain **static factors** which include age, nationality, prior history; and **dynamic factors**, which are socio-economic and personal in nature. Static factor indicators cannot be modified, while dynamic factors (such as substance abuse) are potentially changeable.

An assessment that bridges both assessments (unstructured clinical assessments and actuarial assessments) is called structured professional judgment, in which clinical experts use statistical risk indicators as a guide, and use them to make a decision in their analysis.

How are actuarial risk assessment tools evaluated?

Risk assessment tools are evaluated on their accuracy, calibration and discrimination. If a tool is accurate, it will correctly identify true positives – referred to as **sensitivity** – and it will correctly identify true negatives – referred to as **specificity**. The more accurate a system is the lower the error rate will be; in this case, an accurate tool would correctly predict if an inmate reoffends or does not.

Calibration measures how well an actuarial tool captures absolute risk. If a tool is well calibrated, it will predict the probability of an outcome with the observed probability. If, for example, based on historic data it has been found that 9.5% of Catalonia's inmate population will reoffend, a well-calibrated tool will assess the current population in line with the observed probability.

Discrimination measures capture relative risk. That is, they measure how well a tool separates those who are high risk from those who are low risk. A common way to measure discrimination is with the area under the curve (AUC). A tool with an acceptable **AUC score** (between 0.7–0.8), is better at identifying a high-risk inmate than flipping a coin. If the AUC score is at 0.5 this means the tool is no better than flipping a coin in deciding who is high risk.

According to the Urban Institute, AUC scoring is a good metric in determining whether a tool effectively discriminates between people who are high risk and people who are low risk, but does not measure the probability of reoffending. So "AUC might then be a good metric for stakeholders to consider if they are using the tool to determine how to prioritize and target a fixed amount of resources" (Tiry and Kim 2021).

Case Nº2: **RisCanvi**

– Predicting criminal recidivism in Catalonia

Context

In the summer of 2007, Francis Evrard, a 62-year-old man in France who was incarcerated for serial rape and paedophilia, was released. Within one month, he reoffended and kidnapped and raped a 5-year-old child (Savary, 2009). In Catalonia later that summer, Alejandro Martínez Singul, a serial rapist and sex offender who had been recently released from prison, caused great outrage and concern as Barcelona residents were fearful of Singul replicating what Evrard had done months earlier.

That same year, spurred by the outrage in the media, legislators were called to action and the effectiveness of the prison system was called into question. In an effort to prevent cases like these from happening, Montserrat Tura, former Minister of Justice of Catalonia set up an investigative committee to propose initiatives to inhibit the cycle of recidivism. The commission was formed of various experts, from psychologists, jurists, endocrinologists, etc. and there was special concern around sexual violence and reoffence in sexual crimes. As a result, the commission came out with a series of recommendations, which had the following results: they implemented supervised release (*libertad vigilada*), and second, they commissioned the design of a risk assessment protocol to evaluate the risk of reoffence, now known as RisCanvi.

RisCanvi – origins

Risk assessment instruments, as mentioned, are increasingly used for sentencing, correction and parole decisions worldwide. In 2008, Catalonia's Department of Justice commissioned a Group (Group of Advanced Studies in Violence) led by Professor of Violence and Criminal Psychology, Antonio Andres Pueyo, to design an actuarial assessment tool for the Catalan context.

The Group took two years to develop a protocol designed closely with the Department of Justice. They had access to five years of internal data (from 2003-2008) on the inmate population to understand and identify risk factors of ex-inmates who later reoffended. Based on the data of approximately 600 inmates, the research group created RisCanvi (named from the Catalan words for risk and change) that initially was used to predict the probability of four outcomes. Since its creation in 2010, the tool has gone through three iterations and now calculates five outcomes:

1. Self-directed violence – suicide attempts or self-inflicted injuries in the penitentiary centre.
2. Violence towards other inmates or staff
3. Violent recidivism
4. Breaching parole
5. General recidivism

In order to predict the probability of each, the protocol uses 43 risk factors. A risk factor for such tools is a variable that is found to be highly correlated with re-offence. The research group conducted an extensive literature review of other methods and protocols to understand how other tools were developed and which risk factors were taken into consideration.

How is it used?

When an inmate is registered in the prison system, they first go through a screening tool, called RisCanvi – Screening, an assessment tool with no/yes questions to determine the initial risk level of the offender. The screening tool is composed of 10 factors that are completed with the initial report and interviews. These factors cover information such as the age at which the offender first engaged in violent activity and whether the offender has family support.[2] This first tool is composed of static factors that cannot be modified. The screening tool categorises the offender as low or high risk. If the offender is categorised as low risk, they undergo the screening again in six months.

If the offender is categorised as high risk, they undergo RisCanvi-Complete or RisCanvi-C. A team of multidisciplinary professionals collects data regarding each factor alongside the clinical history, observations and interviews. They then input the information into the tool which assigns the inmate a risk score. The algorithm's outcome is only a three-level classification meaning risk can be low, medium or high.

The final evaluation is then assessed by the team to determine the type of treatment the inmate receives. Inmates are evaluated at least every six months, unless there is misconduct in the prison such as a suicide attempt or violence against other inmates. If such an event occurs, the evaluation is modified. The predicted levels are also used in reports sent to prosecutors and judges to consider a conditional release.

The tool is used to assess the potential risk of the inmate, and the probability of the inmate either being violent within the prison or if released the type of risk the inmate poses of reoffending. The outcomes are consulted throughout the inmate's "lifecycle" within the penitentiary system. This helps management allocate resources and determine subsequent treatment of the inmate.

RisCanvi-C uses 43 risk factors in five main areas. Below are some factors used in the assessment tool: (Moreno Yuste, 2015)

1. Criminal history
- Type of violent offence
- Age they committed their first offence
- Offence committed under the influence of drugs/alcohol
- Offence caused victim injuries

[2] See Annex I for complete factors of RisCanvi and RisCanvi-C

RisCanvi

2. Prison history
    - Whether they have been previously incarcerated
    - Conflicts during incarceration
3. Personal history
    - Family members also incarcerated
    - Level of education
    - Employment status
4. Clinical history
    - Drug use
    - Severe mental illness
    - Promiscuous sexual behaviour
5. Personality
    - Pro-criminal attitudes or antisocial values
    - Impulsive, emotional instability

The tool is composed of these 43 risk factors, both static and dynamic, along with four other variables that include sex, nationality, age and sentence status (whether the offender is awaiting final sentencing or serving their sentence). According to Pueyo, these variables have significant weight on the final outcome, for example, whether the inmate is a national has a different weighting on the mental health risk factor because in general, immigrants tend to have better mental health in the Catalan prison system. On the other hand, risk factors such as social integration might be higher for immigrants given that they lack resources (Digital Future Society 2022a 00:09:43).

RisCanvi is a tool used to assess the risk of recividism and violent recividism of the inmate population.

01.

When an inmate is registered in the prison system, they first go through a screening tool called **RisCanvi – Screening**, which is an assessment tool with **no/yes questions to determine the initial risk**. The screening tool categorises the offender as low or high risk.

Evaluated again in 6 mths

02.

If the offender is categorised as high risk, they undergo **RisCanvi Complete (RisCanvi-C)**. A multidisciplinary team enters the additional data into the tool which assigns the inmate a risk score. The tool calculates five outcomes of violence and general recidivism. The scores are used to determine the type of treatment the inmate receives.

Evaluated again in 6 mths

03.

The predicted levels are also used in reports sent to **prosecutors and judges** to consider conditional release.

Image source: Digital Future Society.

Discussion

Measuring the effectiveness of risk assessment tools

According to a 2015 study by Manel Capdevila Capdevila's team at the Centro de Estudios Jurídicos y Formación Especializada (Centre for Law and Specialised Studies) (CEJFE), the accuracy of this system to predict the recidivism of a prisoner was good. According to the standards of this type of tool 77% of the prisoners classified as medium-high risk reoffended and 57% of the non-reoffenders were classified as low risk by the algorithm (Capdevila 2015).

A subsequent study by criminologist Lucía Martínez Garay, however, criticized the way in which these results were framed, stating that the conclusions were too optimistic and confused the sensitivity of the tool with its overall predictive accuracy. She argues that Capdevila's results actually show the tool has a lower sensitivity – in only 17.94% of cases did it correctly predict recidivism in high and medium risk inmates; however, it more successfully predicted low risk profiles, showing a high level of specificity – 95% (Martinez Garay 2016). Regarding discrimination, calculated with the AUC score, Martínez Garay concludes that RisCanvi falls within the performance parameters of other actuarial risk assessment tools. These measurements are essential in understanding how to best use the tool. As mentioned, each measurement can have its purpose at one point in the decision-making process, given that the tool is used at different instances throughout the inmates' "lifecycle" and by different professionals.

Experts stress that a clear understanding of the limitations of these risk assessment tools is necessary to effectively use them. Overall, actuarial risk assessments are recommended as advisory tools, and their predictive efficacy according to some experts may have reached its limits. Ultimately, Martinez Garay cautions on the use of Riscanvi to make decisions about the freedom or fundamental rights of inmates (Martinez Garay 2016).

Replicating Bias

Risk assessment tools like RisCanvi have been promoted in their role in reducing bias and providing an objective analysis of risk. Advocates of the tool stress the advantages of implementing a set protocol within the penitentiary system. They claim that relying on clinical assessments only was found to provide disjointed and biased information. Nonetheless, many authors have criticised the argument that risk scores are both more objective and accurate tools. Because these tools are based on historic inmate population data, many caution that these tools are actually replicating biases.

Bernard Harcourt, lawyer and critical theorist, argues that risk assessment tools, as they rely on an inmate's past, specifically their criminal history, actually predict policing – that is, how likely it is they are going to get arrested, rather than a person and their dangerousness (Cardoso, 2020).

Protocols, objectivity and biases

In a recent study exploring the professional use of RAIs, the authors found that professional users of the RAI were aware of over- and under-representation of data and the potential discriminatory outcomes that may arise from their use (Portela et al. 2022). Some experts argue there is no way to eliminate bias, and highlight that when using these tools professionals need to be well-equipped to identify and counteract these biases.

According to researcher Manuel Portela, the professionals that know how to work with RisCanvi are aware of the tool's limitations. His research shows that the users see the tool as a positive asset in their work. They believe it provides a wealth of data, which gives a more "objective" perspective than the caseworkers' discretion alone.

Again, context is important in understanding the potential risk of amplifying bias. There is a margin of error for both humans and algorithms in the decision-making process. For this reason, RisCanvi is evaluated by a group of caseworkers – the decision does not fall upon one. Nonetheless, there is contention regarding the use of RisCanvi, the value of the information it provides and the objectivity of its users. RisCanvi was designed for use in correctional and rehabilitation facilities. But it is also used by judges to determine whether the inmate should be released with parole, etc. (Garrett and Monahan 2019).

Defending inmates' rights

According to lawyer Nuria Monfort, expert in criminal law and member of IACTA – a non-profit cooperative that perceives the law as a tool for social change – for those who practice law, RisCanvi is a bit of a mystery (blackbox): there is very little understanding of how it works. She says that while we now understand what the 43 risk indicators are, there is still much to know about how risk is calculated and the weighting given to each variable (Digital Future Society 2022b 00:17:20).

What they do know is that it is used to predict the future risk of reoffending – which according to Monfort is a complete shift in the practice of law, because before the use of RisCanvi there was a dialogue. But with RisCanvi in the picture, there is little room for debate. The risk factor is presented as a scientific truth – hard to contest.

She says, "The need for RisCanvi is part of a whole paradigm shift toward societies that try to avoid risk... of security versus freedom." She notes that risk or dangerousness is not something new; what is new is that now there is an algorithm to measure this danger (Digital Future Society 2022b 00:22:00).

RisCanvi becomes significant when inmates can get permission to leave, or conditional leave. This occurs in Spain with its three degrees of confinement based on the type of offence committed and the degree of danger an offender poses. Prisoners of the third degree, low risk, are allowed to enjoy a regime of semi-freedom and the tool is employed to assist in calculating this risk.

Monfort notes that societies must understand the political and ideological influence of these tools, and that they are far from being objective and scientific, that they prioritise efficiency over the human rights of inmates. She says, “We have a huge prison population... a limited number of professionals and they have to issue a number of resolutions. So we are going to have a machine. It’s got nothing to do with the social function, the collective responsibility, with the individual’s conduct, it’s about getting through the pile” (Digital Future Society 2022b 00:22:25).

According to Monfort each legislative change in the criminal code followed a violent case that got a lot of attention in the media, usually sexual violence. This legislative change was made based on one perspective of social risk, focusing on a type of crime that is not among the most common. But RisCanvi is applied to all types of crime and mostly applied to a segment of the population that has nothing to do with the type of crime the tool was developed for and is at risk of exclusion. “RisCanvi arose for the prevention of violent crimes,” Monfort says, “and now it turns out that its focus is non-violent crimes... which pose the highest risk of recidivism” (Digital Future Society 2022b 00:19:45).

Similar cases abroad

Risk prediction tools are widely used in many countries, though there is little research on the scope (where and what tools are used) and their impact on the local inmate population. Risk prediction tools, specifically those used to predict criminal recidivism, have been in the spotlight for criticism received in countries like the US and UK. Among the many risks these systems pose, is discrimination, as they are entrenched with society’s biases. A UK system called the Offender Assessment System (OASys), comparable to RisCanvi, is used in pretrial hearings, sentencing and parole.

A 2014 National Offender Management Service (UK) analysis, found that the reoffending predictor and violence predictor of the system generated different predictions for race, gender and age: the validity “was greater for female than male offenders, for white offenders than offenders of Asian, black and mixed ethnicity, and for older than younger offenders.” For COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), a similar tool used in the US, it has also been reported that “the system predicts that black defendants pose a higher risk of recidivism than they do, and the reverse for white defendants” (Big Brother Watch 2020).

Case Nº3: **VioGén**

– Risk assessment for gender violence

Background

In 2004, the Spanish parliament unanimously passed a landmark piece of legislation against gender violence, the first of its kind in Europe, "to prevent, punish and eradicate gender violence and assist its victims" (Ministerio de Justicia 2009). This act, in its integrated approach, laid out directives for national law enforcement for a protocol. As the act states, "These protocols will facilitate the prevention and early detection of gender violence and ongoing assistance to women suffering or at risk of suffering it." The act laid the groundwork for the risk protocol assessment called VioGén (from the Spanish for Violence and Gender) designed in 2007 by a team of experts led by scientist and professor Antonio Andrés Pueyo, who would a year later create RisCanvi.

Much like RisCanvi, the legislation for VioGén's creation was driven by public outrage after the death of Ana Orantes, a woman who was murdered by her ex-husband in 1997 after a television appearance testifying to the violence she had endured during and after their marriage. In front of a live audience, she explained that she had gone to the police several times but to no avail. Her brutal killing and the ensuing public outrage prompted policymakers and society at large to take gender violence seriously. Prior to her death, gender violence in Spain had been seen as a private matter, only occurring on rare occasions – and there was practically no legal support for victims (Minder 2020).

What is VioGén?

VioGén[3] is a system used by Spanish police to help estimate the risk of recividism in gender violence. The tool is used by national Spanish police forces, which include the Policía Nacional and Guardia Civil and local police forces not including those in Catalonia or the Basque Country. While the system is under the jurisdiction of the Ministry of the Interior, it is used by other ministries, such as the Ministry of Equality, Social Services, and Justice.

How does it work?

When an incidence of gender violence is reported to the police (which could be presented by a relative, witness or law enforcement agent – it need not be the victim) this initiates an administrative process in which the police officer opens an investigation and fills in an online form with the victim. The form is called Police Risk Assessment (*Valoración Policial de Riesgo, VPR 5.0*). It has 35 indicators under five categories, which the police can check off if they apply to the victim's case.

[3] *Sistema de Seguimiento Integral en los casos de Violencia de Género* (Comprehensive Monitoring System in Cases of Gender Violence)

These items include:

1. A history of violence in the relationship such as
   - severity of previous assaults, whether sexual or physical,
   - use of weapons
   - death threats
   - signs of extreme jealousy
2. Characteristics of the aggressor such as
   - employment status,
   - addiction or substance use,
   - mental health
3. Information pertaining to the victim, whether she is
   - pregnant,
   - economically dependent
4. Information related to minors
   - whether the victim has minors in her care
   - whether the victim fears for their well-being/life
5. Aggravating factors
   - the victim has expressed her wish to end the relationship
   - the victim believes the aggressor is capable of inflicting strong violence or murder

(See Annex II for complete list of items)

Some items can be weighted. For example, when addressing the history of violence, physical assaults are weighted according to the severity of the assault. For example, if there is physical violence, the officer can distinguish the intensity of the violence, ranging from scratches and bruises to stabbing or intent to asphyxiate the victim. This checklist allows officers to assess the severity of the situation.

There are five resulting outcomes: no risk, low, medium, high and extreme risk. The officers can manipulate the score depending on their perceived risk of the situation. Depending on the risk level of the first report (if they are classified as medium, high, very high risk), periodic forms known as the Valoración Policial de la Evolución del Riesgo (Police Assessment of the Evolution of Risk) (VPER) are filled out in order to assess whether the situation has improved or worsened.

The time at which VPER is filled out depends on the level of risk assigned by the first assessment. If there are no incidents (such as new complaints) that require the police to follow up beforehand, those classified as very high risk need to be followed up within 72 hours, while those that are high risk within seven days, (medium risk 30 days and low risk 60 days). Cases can be followed up across different regions. Depending on the risk assigned to each case, officers come up with a "custom security plan". If a case is filed as low risk, the aggressor might be lightly surveilled, while if the aggressor attains a higher score, the victim's house might be guarded, and the aggressor put under tight surveillance (Catanzaro 2020; González-Álvarez et al. 2018).

What happens to a risk score once the report has been filed?

A case in the VioGén system corresponds to the victim that files a police report against her aggressor. One victim can have multiple cases, and the same applies to the aggressor. That is, if an aggressor has multiple victims, there is a case for each of his victims. Thus, when interpreting how many VioGén cases there are, the cases outnumber the people who are in the system.

Filing a report is the first step in the process. Once the victim files a report, the aggressor is detained and there is a judicial hearing in which the court decides whether the case pertains to gender violence, and if so, what measures should be taken. The risk assessment results are considered along with other information provided by the police, witnesses, etc.

A VioGén case remains active if it requires police attention following the periodic assessment mentioned above. In the process of risk assessment, should the risk decrease to a point that the officer believes there is no further risk of violence, the case becomes inactive. However, if there is a new incidence of violence, or another report is filed, the case is reactivated. A case is deregistered from the system only when there is legal justification.

How has the tool been perceived?

This year, 2022 marks the 15th anniversary of the tool. According to the Ministry of the Interior it has issued more than 700,000 cases (Ministerio del Interior 2022). However, since its implementation, there has been controversy on the degree of risk assigned to each case. In 2014, daily newspaper *El Mundo* published a leaked document that showed how 14 out of 15 women who were killed that year had been classified as low risk (Catanzaro 2020).

In the same vein, in 2018 Itziar Prats was also determined low risk, after she had reported her ex-partner's violence to officials. While she had reported his threats to murder her children to both police officers and judge, the officials did not deem her case high risk, as violence to children was not an item on the checklist at the time. In September of that year, her ex-husband murdered their two children (Álvarez 2019).

In March of 2019, VioGén was modified to include items such as the risk of murder of women and minors, which since its implementation in 2007, had not been factors for consideration in the process.

In addition, other improvements have been made to the tool since its implementation. For example, victims of gender violence, generally, are not good evaluators of their own risk for future violence as they underestimate the abuse they suffer. They usually do not admit their victimhood. If asked about their children, however, they do admit the danger they are in. These types of questions have been reformulated taking into account the psychology of the victim and their perceptions of violence (Eticas Foundation 2022).

How does the tool perform?

According to a recent study published in 2020, researchers found that VioGén performs in line with other actuarial tools, with a sensitivity (true positives) of 84%, specificity (true negatives) of 60% and an AUC at 0.80 (López-Ossorio et al. 2020).[4] Based on these measurements, the research team states that the tool can identify any extreme risk cases that lead to homicide. Nevertheless, a team from the Eticas Foundation highlights that in Spain only one in four cases of homicide occurs after the victim files a report. So unfortunately, many of the homicide victims do not undergo the police risk assessment process and therefore would not be identified as high risk (Eticas Foundation 2022).

A lawyer's perspective

Sonia Márquez is a lawyer for the Fundación Ana Bella, a civil society organisation based in Seville, Spain, dedicated to helping survivors combat gender violence. In her experience, she has found that the tool and the system that surrounds it, could do a lot more to engage its end users (the victims), even if the protocol itself guarantees certain rights to victims who file a report, such as the right to be accompanied by a legal professional (Digital Future Society 2022d 00:29:11).

Firstly, there is a general unawareness of the system itself – victims are unaware that they are assigned a risk score, and what that means. This highlights an area of concern as the design and evaluation of the tool do not consider the social impact these have on the affected population.

According to an external audit conducted by Eticas, 48% of the 31 women interviewed, evaluated their experience negatively (Eticas Foundation 2022). Márquez attests that the interview process may be disorienting due to a series of factors, such as the way that questions are asked. Some questions are about the violence itself and others are to determine what type of protective measure should be implemented. Little explanation is provided to give the victim clarity as to why these questions are being asked.

Secondly, victims are at a great disadvantage if they do not have the administrative or bureaucratic understanding of how the filing process works. As a lawyer, Márquez provides valuable information to her clients to navigate the system. This is important because the risk score can be very different depending on:

1. Who takes down the report – police officers have their own biases on gender violence.
2. The time and day the report is filed – some stations have specialised staff to receive victims of gender violence. But these specialised officers have specific hours within each department.
3. Where the report is filed – some cases occur in a small town or capital; this makes a difference to how a case is handled – not all courts have a team specialised in dealing with gender violence victims.

[4].These pertain to the tool's newly designed homicide scale, the H-scale, used to improve the prediction of Intimate Partner Homicide (IPH) – which is an assessment designed to run alongside the global VPR tool.

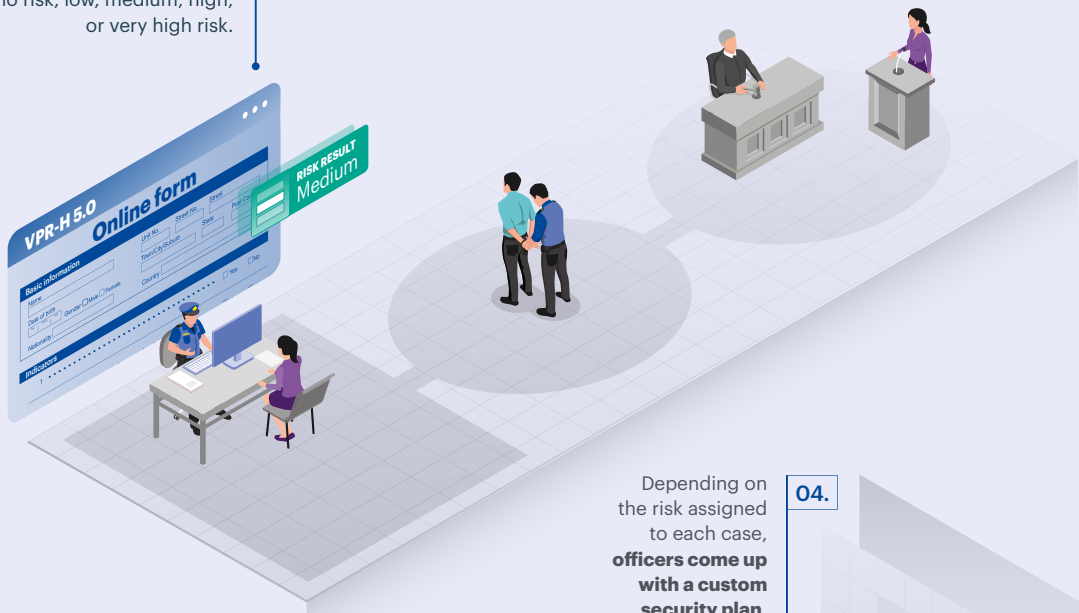
VioGén is a system used by Spanish police to help estimate the risk of recividism in gender violence.

01.

When an incidence of gender violence is reported to the police it initiates an administrative process in which the police officer opens an investigation and fills in an online form with the victim. The report can classify the risk of future violence as no risk, low, medium, high, or very high risk.

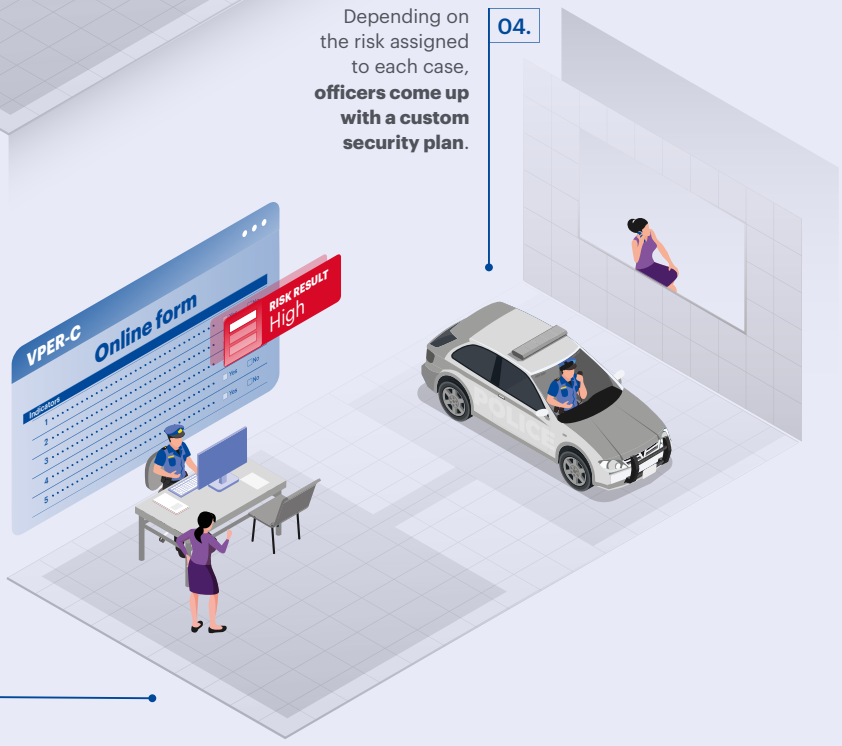
02.

Filing a report is the first step in the process. Once the victim files a report, **the aggressor is detained and a court decides whether the case pertains to gender violence**, and if so, what measures to take. The risk assessment results are considered along with other information provided by the police and witnesses.



04.

Depending on the risk assigned to each case, **officers come up with a custom security plan**.



03.

If the risk decreases to a point that the officer believes there is no further risk of violence, the case becomes inactive. **However, if there is a new incidence of violence, or another report is filed, the case is reactivated**. A case is deregistered from the system only when there is legal justification.

Image source: Digital Future Society.

Discussion

A step in the right direction

Gemma Galdón, founder of the Eticas Foundation – attests that the survey does a good job at putting academic findings into practice, as the design of the questions reflects the complexity of victims of gender violence. VioGén is also applauded by victims' rights organisations as an effort in collecting data to better understand the reality of victims of gender violence.

Galdón sees room for opportunity in that **automation has opened the door to collecting and analysing information that would not have been possible without the help of technology**. According to Galdón, as a static algorithm, VioGén cannot learn from previous outcomes such as cases incorrectly identified as low risk. Therefore, the **questionnaire cannot be adjusted according to whether the algorithm is successful in determining a victim's past risk or not**.

Galdón sees huge potential in analysing the data collected by the tool, which could allow stakeholders to better understand the problem at hand. For example, there are certain correlations that the data has shed light upon that have not been previously covered by existing literature such as the victim's place of employment greatly conditioning her likelihood of future abuse.

Human oversight

As mentioned, actuarial tools provide a certain structure as statistical analysis tools in decision making. With regards to VioGén, the system was designed to improve the day-to-day work of police officers and account for the potential biases and variations of each individual. The risk assessment is designed to allow for modification and is intended as a tool to compliment the police officer's professional experience.

However, a 2014 study showed the majority of the officers (95%) relied on the automated outcome (Catanzaro 2020). This study calls into question whether the police officers, as the first line of contact, are the experts needed to adequately assess the risk of gender violence. Experts – **psychologists and forensic physicians – highlight the limitation of using police officers for risk assessment**, as they do not have the professional knowledge required to assess the indicators. Specially trained professionals would ideally fill the gap.

Not only does this lie beyond the expertise of police officers, as Sonia Márquez points out, there are a number of institutional, organisational and cultural factors that influence the way these assessments are made. According to an Eticas external audit, **out of the 27,000 officers involved in monitoring, only 2,000 are specialised in gender violence.**

There is a lack of specialised professionals on the frontline that understand the victims' needs, which adds to the inherent mistrust that victims have in the system. The problem of gender violence goes far beyond the scope of this system. There was a notable decline in gender violence reports in 2020, during the pandemic, which brings into question whether police reports are an adequate prevention measure (Álvarez 2021).

When it comes to ADMS, human oversight is often implemented as an afterthought to comply with regulations. The Eticas external audit of the system raises the concern of how human oversight is designed as part of the system. The lack of clarity in how officers should use the tool is telling. Galdón explains that there are varying interpretations of how a public servant can interact with an automated tool – they may perceive their job as simply operating the tool, or use the tool as part of the decision-making process in which they feel like the "owners" of the assessment.

The victim's perspective

Another critique of VioGén is that the system design does not account for the victim's experience. As mentioned, many victims are not aware they are being assessed by an algorithm, or what their rights are as victims of violence. The survey has not been tested with victims of violence, nor have there been workgroups to evaluate the system itself – whether there should be new questions, etc.

According to Galdon, "It is essential that algorithms with social impact involve the end-user to ensure the quality and transparency of these systems. The women who undergo VioGén have the right to know how these decisions are being made and understand that this risk assessment conditions what type of police and legal protections are going to be implemented to protect them" (Digital Future Society 2022d 00:23:32).

The case of Itziar Prats highlights various faults in the system design that could have been addressed if it had been more in line with the victim's perspective, though some issues have been addressed like the inclusion of the variable of threats of murder. Critics of VioGén claim that such degrees of risk should not exist. They argue that the category of "no risk" should not be included, that all cases should be considered cases of risk, so as to avoid false negatives. Like Itziar Prats, there are more cases in which the women who seek support are revictimized or their case is disregarded because the quality of information, or the victim's testimony alone, does not meet the officers' or judges' expectations (Coronado 2022).

Other models

Standardised risk assessments, which include actuarial tools like VioGén or structured professional judgement (SPJ) scales, have been in use since the 90s in North America. They include well-known tools such as the Danger Assessment, Spousal Assault Risk Assessment (SARA), Domestic Violence Screening Inventory and Ontario Domestic Assault Risk Assessment.[5] These tools originated in clinical settings and later were adopted by victim advocates. Now, they are in use by police forces (Helmus and Bourgon, 2011).

In Europe, these tools began getting attention in the 2000s. First mentioned in the European parliament in 2004, the institution encouraged the development of risk assessment tools as an adequate measure to reduce gender violence. Spain and Sweden are the frontrunners in this

[5] For more information on these tools visit: https://www.rma.scot/wp-content/uploads/2019/09/RATED-Collated-Intimate-Partner-Violence-Awaiting-Validation.pdf

aspect with risk assessment and management embedded in their legal frameworks (European Institute for Gender Equality, n.d.).

B-Safer, an adaptation of the US tool, SARA, was implemented in Sweden in the 90s. The UK has been using DASH, an SPJ scale, since 2009. In 2019, Portugal adapted the Spanish model (*La Vanguardia* 2019). On the international scene, there is much debate as to whether risk assessments are an adequate tool in predicting gender violence.

Some tools are more successful than others at predicting violence. There is still a lot to learn about gender violence in general, given that globally, the majority of cases go unreported, and about patterns of abuse, such as those that do not show visible harm (UN Women 2022).

Across the literature, academic researchers agree on a cautionary approach when implementing risk assessment tools, and emphasize that users must be aware of their strengths and limitations to guarantee responsible use of these tools. This includes the importance of having (skilled) evaluators that understand how to interpret the results. Missing or unreliable data also affects the accuracy of the tools, therefore awareness of the way the questionnaire is administered, setting, etc., are all important factors the evaluators should take into account.

There is still much to learn about how these tools perform in diverse populations, something that experts have especially noted for tools such as SARA and VioGén (Helmus and Bourgon 2011). This is one of the major concerns, especially for the AI-based tool piloted by the Queensland Police Service in Australia to predict domestic violence (Smee 2021). Although the Queensland police acknowledge the potential for the model to disproportionately impact Indigenous and other minority communities, and has taken steps to mitigate bias in the system, local civil society organisations are wary of the risk of such models reinforcing the biases found in historical data.

Case Nº4: **SALER**

– Anticipates potential cases of corruption in Valencia

Context

In the early 2000s, Valencia, the fourth largest autonomous community in Spain, had a bad reputation for rampant political corruption. In 2015, an article in *El País* counted the number of officials facing trial, from regional deputies to premiers, at a total of 150 elected politicians (Barbería 2015). The voters' 20-year support of the People's party was about to change. In 2015 they voted in two parties that with the support of a third, formed a coalition government known as the Pacto del Botànic, to the Valencian parliament.

One of the main concerns of the new coalition was to regain trust and recover from the bad press of the regional government in past decades. It created specific entities to combat corruption such as the Anti-Fraud agency and the Regional Ministry of Transparency, Social Responsibility, Participation and Cooperation (Ministry of Transparency).

The institutional narrative was not to prosecute old cases of corruption but to start anew and adopt new ways of working to re-establish the citizens' trust in the regional administration. Under this new scenario, the new Ministry of Transparency sought to create an alert system to detect misconduct and prevent a series of events that could result in fraud, corruption or malpractice.

The alert system's main objective was to prevent previous cases of corruption from "contaminating" the newly formed government. It was very important that the system be designed to identify possible bad practices in real time and accompany the public servant in identifying it and resolving it.

What is SALER?

The Early Warning System of the Generalitat Valenciana, (SALER is from name in Spanish – *Sistema de Alertas*) was officially created in 2018 when the Valencian parliament passed a law to regulate its use.[6] It is regulated by the General Inspection of Services, which is the internal control body of the regional administration.

SALER is a **computer system based on data analysis designed to anticipate potential cases of corruption in the public administration**. This tool is the first of its kind developed in Spain; however, it has not been fully implemented in the public administration.

[6] See law: https://dogv.gva.es/datos/2018/11/08/pdf/2018_10294.pdf

SALER

Alfons Puncel Chornet, former undersecretary of the Ministry of Transparency in the first Pacto del Botànic[7] legislature from 2015 to 2019, was the father of the tool. He recounts the moment he was inspired to create it...

> A construction company was asking for reimbursement of a deposit they had paid to guarantee completion of a project. The comptroller overseeing the public works contract realised the company was asking for the deposit back before the project was finished. They had been using a paperwork management process and I saw the possibility to digitise the system so it could alert public servants of similar cases immediately.

The conceptualisation and design of the system technically started in 2016 under a different name – SATAN8 – with the help of the Computer Science department of Valencia Polytechnic University (Cid 2018). At the time, the General Inspection of Services was understaffed and did not have enough resources to develop the tool. So, in parallel, the Ministry of Transparency worked to lay the groundwork so that the system could operate internally. TWhis meant collaboration with other departments, providing the legal framework and hiring staff.

How would this tool work in practice?

The SALER system is designed to use data from different sources within the public administration such as:

- contract records (procurement and bidding processes)
- information on direct payments (suppliers, services contracted)
- data on subsidies (grantor, beneficiaries, invoices, etc.).

Additionally, SALER collects data from notary databases, the property registry and the business registry.

It cross references data from these different sources to flag conflicts preventively or while they are taking place, such as during a bidding process when it can raise an alert of a conflict of interest. It flags the following types of conflicts: conflicts of interest, duplicate funding, collusion and the avoidance or mishandling of public procurement procedures.

Once the system flags a case, an inspector from the General Intervention Service is alerted. The inspector opens an investigation, which can lead to further inspection or dismissal. Upon further investigation the inspector can determine whether it is a mistake or negligence, bad practice or a potential case of fraud. As an early-warning system it is designed to accompany the official while they carry out their work. These alerts are not open to the public.

[7] Coalition party formed by three parties on the left, marking a shift in the 20 years that the People's Party governed in Valencia.

[8] *Sistema de Alertas Tempranas Anticorrupción* (Anti-Corruption Early Warning System)

SALER

How was the system designed?

The system is trained on patterns from real cases of malpractice and searches a set of databases to identify risk. A team composed of the General Inspection Services defined the indicators and queries to generate the algorithms.

What stage is SALER at now?

SALER is at an initial stage incorporating a series of databases. One of the main challenges SALER has faced is the slow digital transformation of the public sector. For the computer system to operate, it has to have access to a consolidated system that unifies the datasets and has a sufficient amount of data. This is difficult to accomplish as not every department or organisation has data that is good quality, accurate and accessible. As the system uses databases from different sources, the data is not easy to analyse – not all databases are encoded the same way. And the system requires data to be prepared so it can be cross-referenced, which takes time and effort.

Not only are there technical challenges stalling the full implementation of SALER, the original advocates for the system, the secretary and undersecretary of the Ministry of Transparency, were moved to other roles after the first term of the Pacto del Botánic.

SALER is a tool used to anticipate potential cases of corruption in the public administratio.

The **SALER** system is designed to use data from different sources within the public administration such as: contract records, information on direct payments, data on subsidies. 01.

It cross references data from these different sources to flag the following types of fraud: conflicts of interest, duplicate funding, collusion and the avoidance or mishandling of public procurement procedures. 02.

Once the system flags a case, an inspector from the **General Intervention Service** is alerted. The inspector opens an investigation, which could lead to further inspection or having the case dismissed.

With further investigation the inspector can determine whether **the case is a mistake** or negligence, bad practice or a potential case of fraud 03.

Image source: Digital Future Society.

Discussion

Managing trust in SALER

One of the main concerns about alert systems is the suspicion that these tools are in use to surveil and punish their users (public servants). Wary that inspection can garner distrust and result in users undermining the system, the advocates of SALER were very cautious to explain the difference between SALER and tools that do punish and surveil. Nonetheless, they met with internal resistance.

Transparency plays a key role in managing this trust. With more transparency, there is more institutional buy-in. However, systems like SALER that pursue an aim of public interest such as the fight against fraud and corruption, are not held to the same transparency standards as others.

There is debate as to how much transparency the SALER system should have. The argument is that too much transparency might be counterproductive. A bad actor could use the information to find a way to carry out corruption, thus subverting the system.

Compliance with data regulation

Another major challenge with using large amounts of data as the SALER system requires is compliance with data regulations. The system had a rough start. In 2017, as legislation for the tool was being debated, the Spanish Agency for Data Protection issued a report that was highly critical of the alert system – declaring it non-compliant with the provisions of article 23 of the GDPR.

To comply with the data protection regulation, systems like SALER need to have a clear objective of what data they are going to collect and analyse, for what purpose and the duration. Regulation can be perceived as stifling innovation. One of the main advantages of using these systems is the potential that AI offers to cross-reference data from different databases and determine the risk indicators. Data protection laws limit the potential of tools like SALER, where machine learning can enable the system to detect bad practices previously undetected by humans.

However, the system would need to justify the data it cross-references and specify what databases it has access to and what rights people have with regards to that data. The data has to be proportionate and use the minimum number of data points to accomplish the objective of the system. Before using the system, specific targets must define which datapoints must be analysed to assess the risk of corruption.

Robust concept, uncertain impact

Oscar Capdeferro, expert in administrative law, commends the tool as an effective anti-corruption measure (Capdeferro 2018). On a conceptual level, SALER follows internationally backed recommendations, especially in the way the alert system is based on previous risk analysis. Another factor worth noting about SALER is that it is designed to coordinate with

other anticorruption measures. Cases are later followed up by the Inspection Service and if the case turns out to be one of corruption, it is backed by legislation that authorises the sanctioning power of the General Intervention Service.

The law delimits the use of the tool and establishes guarantees and rights to users and end-users of the system and the organisation that is accountable for the tool. It also requires regular updates to guarantee the effectiveness of the alert system.

However, there is still a lot to know with regards to how the tool functions in practice. How cases are dealt with in practice can result in a series of outcomes, some unexpected. The way the tool is used by inspectors and the impact the tool has on civil servants, has to do with organisational and institutional factors as much as the design of the tool. One potential application that Capdeferro sees for SALER is scaling it to other local administrations, given the concentration of corruption at a local level.

Other tools in use

The use of data-driven tools to detect fraud and corruption is increasingly seen by governments as an effective and efficient solution to managing public funds. Some tools have gained worldwide attention, like the Netherland's System Risk Indication (SyRi), a system designed by the Dutch government to detect welfare fraud. However, tensions arose on the implementation of the tool. In 2018, civil society interest groups called for a halt to SyRi on the grounds that the system constituted an invasion of a person's privacy and did not have sufficient safeguards. Subsequently, in 2020, the District Court of the Hague ruled that the system violated the right to privacy, and the use of data was disproportionate to the aim of public interest in detecting welfare fraud.[9]

On a wider scale, the European Union has also seen a need to curb corruption of funds. According to a 2019 report by the European Court of Auditors, an estimated 390 million euros is stolen from structured funds every year – this estimate is based on detected cases alone (OECD 2019). Since 2014, the EU has been using ARACHNE, a risk scoring tool to help the authorities detect fraud for the European Social Fund and the European Regional Development Fund (European Commission n.d.a). Based on a set of risk indicators, ARACHNE identifies the risk of fraud and irregularities at the project approval and implementation phase. For now, approximately 20 member states use the tool on a voluntary basis, though the plan long term is to have the tool be compulsory. The Spanish government, for example, has committed to using ARACHNE in its management of EU-funded projects (EuroEFE 2021).

[9] Judgement of the District Court of The Hague 2020: ECLI:NL:RBDHA:2020:1878

Looking Ahead

In a post pandemic world, governments across the globe are confronted with problems, evermore complex, that will be increasingly difficult to address without the help of digital technologies. AI systems allow government officials to streamline services, make data-based decisions, anticipate risks and allocate resources.

Nevertheless, the use of AI in the public sector has begun to draw the attention of the media, civil organisations and the general public for its potential negative impact on citizens. The case studies show that the harms these systems can cause are tangible and if they are not addressed have the potential to exacerbate inequalities. As societies become mindful of the risks, it is important that they understand the whole picture.

The aim of this report is to fill in the knowledge gap between the general public, experts and technologists. It also seeks to bridge the knowledge between Spanish and English-speaking communities. Not only is there a lack of published content about these systems from a non-technical perspective in both languages, there is also a need to share and learn about what is happening in Spain and to contribute to the body of knowledge being generated on this topic.

There are four main issues that must be considered from the case studies.

1. **Transparency**. There are a number of difficulties that stem from a lack of transparency. Without transparency, there cannot be an effective evaluation of the system. Consequently, there cannot be redress if there are any algorithmic harms such as discrimination or unfair treatment caused by the use of an algorithm.

   When addressing transparency and algorithmic systems, there should be a distinction between the lack of transparency of a system itself (which may be an obstacle for a civil servant to understand how an algorithm makes a particular decision) and the transparency of actors, who for many reasons, such as data privacy laws, may not wish to disclose information. For example, the transparency that Civio is asking from the Spanish government, is required to understand whether the software is compliant with the law.

2. **Human oversight and accountability.** AI systems are often implemented to aid in human decision-making rather than serve as a substitute. In practice, however, human oversight takes on many forms and may be inconsistent or a token gesture. The risk assessment tools and the SALER alert system are both designed for specific contexts in which specialised individuals use the information provided by the algorithm as one of many inputs to aid them in their decision-making. In the case of SALER, the alert is just the first step in assessing a potential case of corruption. However, the police use of the VioGén tool raises questions on effective human oversight, given that the majority do not modify the risk assessment. Although experts recommend that risk assessment tools like VioGén are more effective when used by specialised professionals, there is also a strong argument that the tools serve to provide a structured protocol for police on the frontline.

Nonetheless, there are systems in place that explicitly take the administrative weight off civil servants, as demonstrated in the case of BOSCO. The implementation of the software changed an internal process, given that a decision that was once made by a civil servant has now been transferred to the energy companies that have limited understanding of why an application was denied. The complexity of human oversight of these systems, shows that there is a need to better understand how these tools work in practice and whether the tool is aiding in decisions or actually being delegated to take the weight off decisionmakers.

3. **Discrimination.** Three of the four cases studied, BOSCO, Viogén and Riscanvi have a **clear social impact on a vulnerable part of the population**. As discussed in the Bosco case, there is little information available about the software itself to determine whether the exclusion of beneficiaries is due to a software error or the algorithm. Nonetheless, there are significant improvements that could be made to the application process, given that the administrative burden does not ensure that this public aid is accessible to all who need it. The tools RisCanvi and VioGén use historical statistical data. For example, RisCanvi deals with sensitive data that is inextricably linked to groups who have been traditionally discriminated against in the judicial system. Other tools used abroad have also shown a risk of discrimination. Further investigation is needed to understand what biases may be reinforced by the algorithms.

4. **Inclusiveness.** Civil society organisations play a fundamental role in raising awareness of the tools and aiding those impacted gain a better understanding of how to navigate these systems. Their work highlights the need to involve "end-users" in the design and implementation of these tools to drive inclusiveness and prevent discrimination, which unfortunately is rarely considered in their deployment.

   One vivid example is how the initial design of VioGén excluded children from the risk assessment – a fact that might have been avoided had the creators of the system considered the different perspectives of those who go through or work with the system on a daily basis.

   Clear communication with all stakeholders and co-design are tools that generate trust in these systems and help them evolve to be more inclusive. In the case of SALER, communication was key from the beginning in laying the groundwork to help the different departments within the Generalitat in Valencia to understand the aim of the tool. Should officials believe that the SALER tool is in use to surveil, this might cause officials to undermine the system and therefore render it ineffective.

On a final note, it is important to mention that both Spain and the European Union are taking steps to mitigate the potential harms that algorithmic systems may pose, addressing the points mentioned above: **transparency, accountability, human oversight** and **discrimination.**

This is most evident in the proposed AI Act the European Commission is currently deliberating – a world-first attempt at horizontal regulation of AI systems. In anticipation of this regulation, Spain is in the process of creating the Agencia Española de Supervisión de Inteligencia Artificial (Spanish AI Supervisory Agency) – the first of its kind in the EU – to monitor the use of AI. Another example advancing these efforts is the Observatory of Algorithms with Social Impact created by the Eticas Foundation. This project consists of a search engine of social impact algorithms providing relevant information, such as their use, scope of application and aims.

While these are steps in the right direction, many questions remain on whether these measures are effective in addressing the risks AI poses to fundamental rights. The uncertainty makes it plain that coordination among the various actors and relevant stakeholders is necessary to ensure that these systems do not become tools to harm those who are most vulnerable – by perpetuating biases and excluding them from their design. Only by understanding the social impact of these tools, could stakeholders understand how these systems can protect fundamental rights, and ensure that they do not hinder the very institutions they are intended to help.

Annexes

Annex I: **Factors included in RisCanvi-Screening and RisCanvi-Complete**

RisCanvi Screening:
RisCanvi includes basic data on the inmate: age, gender, marital status, procedural and penitentiary status, degree of confinement, type of crime and relationship to the victim.

This is followed by 10 items: 1. age of first violent incident or initiation of violent conduct; 2. violence prior to (the main crime); 3. prior prison behaviour (serious or very serious offences); 4. escapes, breakouts; 5. problems with alcohol or other drug use; 6. prior mental health problems (previous diagnoses of disorders, anger, emotional instability, impulsiveness); 7. prior self-harm attempts or behaviours; 8. lack of family and social support, lack of a relational network; 9. work/economic problems; 10. hostile attitudes or anti-social values.

RisCanvi Complete:
Criminal/Penitentiary Factors: 1. Violent base crime; 2. Age at the time of the basic offence; 3. Intoxication while committing the basic offence; 4. Injured victims; 5. Duration of the sentence; 6. Uninterrupted time in prison; 7. History of violence; 8. Initiation of criminal or violent activity; 9. Increased frequency, severity and/or diversity of crimes; 10. Conflicts with other inmates; 11. Non-compliance with judicial measures; 12. Disciplinary records; 13. Escapes; 14. Regression in degree of confinement; 15. Violation of Permits Personal/
Sociofamily factors: 16. Childhood maladjustment; 17. Distance between usual residence and the prison centre; 18. Educational level; 19. Problems related to education; 20. Lack of financial resources; 21. Absence of viable future plans; 22. Criminal history in the family; 23. Problematic socialisation in the family; 24. Lack of family and social support; 25. Criminal/delinquent friendships; 26. Belonging to social groups at risk; 27. Prominent criminal role; 28. Victim of gender-based violence (only applicable to women); 29. Current family responsibilities.
Clinical factors/personality: 30. Drug abuse or dependence; 31. Alcohol abuse or dependence; 32. Severe mental disorder; 33. Promiscuous sexual behaviour; 34. Limited response to psychological or psychiatric treatment; 35. Anger-related personality disorder; 36. Poor stress management; 37. Self-harming attempts or behaviours; 38. Pro-criminal attitudes or anti-social values; 39. Low mental capacity and intelligence; 40. Recklessness; 41. Impulsiveness, emotional instability; 42. Hostility; 43. Irresponsibility

Annex II

1. HISTORY OF INTIMATE PARTNER VIOLENCE	ANSWERS		
Indicator 1: Psychological violence (harassment, insults and humiliation)	**Yes**	**No**	**No reply**
1.1 Degree of the psychological violence	Mild	Severe	Very severe
Indicator 2: Physical violence	**Yes**	**No**	**No reply**
2.1 Degree of physical violence	Mild	Severe	Very severe
Indicator 3: Forced sex or rape	**Yes**	**No**	**No reply**
3.1 Degree of sexual violence	Mild	Severe	Very severe
Indicator 4: Use of weapons or objects against the victim	**Yes**	**No**	**No reply**
4.1 Bladed weapons 4.2 Firearms 4.3 Other objects	Mild	Severe	Very severe
Indicator 5: Existence of threats or intentions to cause harm to the victim	**Yes**	**No**	**No reply**
5.1 Degree of threats	Mild	Severe	Very severe
5.2 Threats of suicide by the abuser	**Yes**	**No**	
5.3 Death threats by the abuser to the victim	**Yes**	**No**	
Indicator 6: In the last six months there has been an escalation in attacks or threats	**Yes**	**No**	**No reply**

2. CHARACTERISTICS OF THE ABUSER			
Indicator 7: The abuser has shown excessive jealousy or suspicion of infidelity in the last six months	**Yes**	**No**	**No reply**
Indicator 8: The abuser has exhibited controlling behavaiours in the last six months	**Yes**	**No**	**No reply**
Indicator 9: The abuser has committed harassment in the last six months	**Yes**	**No**	**No reply**
Indicator 10: Problems in the life of the abuser in the last six months	**Yes**	**No**	**No reply**
10.1 Employment or economic difficulties	**Yes**	**No**	
10.2 Problems with the justice system	**Yes**	**No**	
Indicator 11: The abuser has caused substantial injury in the last year	**Yes**	**No**	**No reply**
Indicator 12: The abuser has shown disrespect for the authorities or their officers in the last year.	**Yes**	**No**	**No reply**
Indicator 13: The abuser has physically assaulted third persons and/or animals in the last year	**Yes**	**No**	**No reply**
Indicator 14: There have been threats or contempt shown to third parties in the last year	**Yes**	**No**	**No reply**

Indicator 15: The abuser has a criminal and/or police record	**Yes**	**No**	**No reply**
Indicator 16: There are previous or current violations (precautionary or criminal)	**Yes**	**No**	**No reply**
Indicator 17: There is a history of physical and/or sexual assault	**Yes**	**No**	**No reply**
Indicator 18: There is a history of gender-based violence against another partner(s)	**Yes**	**No**	**No reply**
Indicator 19: Presents mental problems and/or psychiatric disorder	**Yes**	**No**	**No reply**
Indicator 20: Suicidal ideation or attempts	**Yes**	**No**	**No reply**
Indicator 21: Has some form of addiction or drug abuse (alcohol, drugs, pharmaceuticals)	**Yes**	**No**	**No reply**
Indicator 22: Has a family history of gender or domestic violence	**Yes**	**No**	**No reply**
Indicator 23: The abuser is under the age of 24	**Yes**	**No**	**No reply**

3. RISK FACTORS / VULNERABILITY OF THE VICTIM

Indicator 24: Existence of any kind of disability or serious physical or mental illness	**Yes**	**No**	**No reply**
Indicator 25: Suicidal ideation or attempts by the victim	**Yes**	**No**	**No reply**
Indicator 26: Has some form of addiction or drug abuse (alcohol, drugs, pharmaceuticals)	**Yes**	**No**	**No reply**
Indicator 27: Lacks positive family or social support	**Yes**	**No**	
Indicator 28: Foreign victim	**Yes**	**No**	

4. CIRCUMSTANCES RELATED TO CHILDREN

Indicator 29: The victim has dependent children	**Yes**	**No**	**No reply**
Indicator 30: Existence of threats to the physical safety of children	**Yes**	**No**	**No reply**
Indicator 31: The victim fears for the safety of children	**Yes**	**No**	**No reply**

5. AGGRAVATING CIRCUMSTANCES

Indicator 32: The victim has reported other abusers in the past	**Yes**	**No**	
Indicator 33: There have been episodes of reciprocal lateral violence	**Yes**	**No**	**No reply**
Indicator 34: The victim expressed to the abuser their intention to end the relationship less than six months ago	**Yes**	**No**	**No reply**
Indicator 35: The victim thinks the abuser is capable of assaulting her with great violence or even killing her	**Yes**	**No**	**No reply**

References

AI Now Institute (2018). Litigating algorithms: challenging government use of algorithmic decision systems. [PDF] Available at: https://ainowinstitute.org/litigatingalgorithms.pdf (Accessed: November 25, 2022)

Alston, P. (2019). Report of the special rapporteur on extreme poverty and human rights. [online] Available at: https://www.ohchr.org/en/press-releases/2019/10/world-stumbling-zombie-digital-welfare-dystopia-warns-un-human-rights-expert (Accessed: November 29, 2022)

Álvarez, P. (2019). 'Denuncié, me dijeron que no pasaría nada y mis hijas ya no están'. *El País*. [online] Available at: https://elpais.com/sociedad/2019/03/23/actualidad/1553369290_857804.html (Accessed: November 29, 2022)

Álvarez, P. (2021). In Spain, gender violence claims more lives in last 30 days than in first four months of 2021. El País. [online] Available at: https://english.elpais.com/society/2021-06-17/in-spain-gender-violence-claims-more-lives-in-last-30-days-than-in-first-four-months-of-2021.html/ (Accessed: November 25, 2022)

Angwin J., Larson J., Mattu S., et al. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica. [online] Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Accessed: November 25, 2022)

Barbería, J.L., (2015) Why Valencia is paying the high price of rampant political corruption. El País. [online] Available at: https://english.elpais.com/elpais/2015/05/06/inenglish/1430932717_848440.html. (Accessed: January 10, 2023)

Big Brother Watch (2020). Big Brother Watch briefing on Algorithmic Decision-Making in the Criminal Justice System. [PDF] Available at: https://bigbrotherwatch.org.uk/wp-content/uploads/2020/02/Big-Brother-Watch-Briefing-on-Algorithmic-Decision-Making-in-the-Criminal-Justice-System-February-2020.pdf (Accessed: November 25, 2022)

Capdeferro Villagrasa, O. (2018). El análisis de riesgos como mecanismo central de un sistema efectivo de prevención de la corrupción. En particular, el sistema de alertas para la prevención de la corrupción basado en inteligencia artificial. Revista Internacional Transparencia e Integridad. [PDF] Available at: http://www.encuentros-multidisciplinares.org/revista-65/oscar_capdeferro-1.pdf (Accessed: December 14, 2022)

Capdevila Capdevila, M., Blanch Serentill, M., Ferrer Puig, M., Andrés-Pueyo, A., Framis Ferrer, B., Comas López, N., Garrigós Bou, A., Boldú Pedro, A., Batlle Manonelles, A., Mora Encinas, J. (2015). Tasa de reincidencia penitenciaria 2014. Centro de Estudios Jurídicos y Formación Especializada, Generalitat de Catalunya. [PDF] Available at: https://cejfe.gencat.cat/web/.content/home/recerca/cataleg/crono/2015/taxa_reincidencia_2014/tasa_reincidencia_2014_cast.pdf (Accessed: December 14, 2022)

Cardoso, T. (2020). Bias behind bars: A Globe investigation finds a prison system stacked against Black and Indigenous inmates. [online] Available at: https://www.theglobeandmail.com/canada/article-investigation-racial-bias-in-canadian-prison-risk-assessments/ (Accessed: November 24, 2022)

Catanzaro, M. (2020). In Spain, the VioGén algorithm attempts to forecast gender violence. AlgorithmWatch. [online] Available at: https://algorithmwatch.org/en/viogen-algorithm-gender-violence/ (Accessed: November 24, 2022)

Cedefop (2021). Digital, greener and more resilient. Insights from Cedefop's European skills forecast. [online] Available at: http://data.europa.eu/doi/10.2801/154094/ (Accessed: December 7, 2022)

Cid, G. (2018). Ingenieros valencianos crean 'Satan' un 'software' para cazar corruptos: así funciona. Elconfindencial.com. [online] Available at: https://www.elconfidencial.com/tecnologia/2018-10-22/algoritmo-anticorrupcion-valencia-satan_1632428/ (Accessed: November 25, 2022)

Coronado, N. (2022). Una Guardia Civil de Viogén revictimiza a una madre en su llamada de socorro al denunciar que el maltratador no le entrega a su pequeña. La Hora Digital. [online] Available at: https://www.lahoradigital.com/noticia/32054/igualdad/una-guardia-civil-de-viogen-revictimiza-a-una-madre-en-su-llamada-de-socorro-al-denunciar-que-el-maltratador-no-le-entrega-a-su-pequena.aspx/ (Accessed: November 25, 2022)

Digital Future Society (2020a). Governing algorithms. [PDF] Available at: https://digitalfuturesociety.com/report/governing-algorithms/ (Accessed: November 29, 2022)

Digital Future Society (2020b). Towards gender equality in digital welfare. [PDF] Available at: https://digitalfuturesociety.com/app/uploads/sites/9/2020/09/Gender_equality_in_digital_welfare_digital-1.pdf (Accessed: November 29, 2022)

Digital Future Society (2022a). Capítulo 1. RisCanvi (I): el algoritmo de la cárcel. [podcast] Available at: https://digitalfuturesociety.com/es/podcasts/capitulo-1-riscanvi-i-el-algoritmo-de-la-carcel/%20 (Accessed: January 10, 2023)

Digital Future Society (2022b). Capítulo 2. Riscanvi (II): ¿Se puede predecir el próximo delito? [podcast] Available at: https://digitalfuturesociety.com/podcasts/capitulo-2-riscanvi-ii-se-puede-predecir-el-proximo-delito/ (Accessed: January 10, 2023)

Digital Future Society (2022c). Capítulo 3: BOSCO and the social bonus to pay for the electricity bill. [podcast] Available at: https://digitalfuturesociety.com/podcasts/capitulo-3-bosco-y-el-bono-para-pagar-la-luz/ (Accessed: January 10, 2023)

Digital Future Society (2022d). *Capítulo 4. VioGén, el software contra la violencia machista.* [podcast] Available at: https://digitalfuturesociety.com/podcasts/chapter-4-viogen-the-software-against-gender-based-violence/ (Accessed: January 10, 2023)

Eticas Foundation (2022). VioGén External Audit. [PDF] Available at: https://eticasfoundation.org/wp-content/uploads/2022/03/ETICAS-FND-The-External-Audit-of-the-VioGen-System.pdf (Accessed: November 29, 2022)

European Commission. (n.d.a). ARACHNE risk scoring tool. [online] Available at: https://ec.europa.eu/social/main.jsp?catId=325&intPageId=3587&langId.it (Accessed: December 7, 2022)

European Commission. (n.d.b). SALER Rapid Alert System. [online] Available at: https://antifraud-knowledge-centre.ec.europa.eu/library-good-practices-and-case-studies/good-practices/saler-rapid-alert-system_en (Accessed: January 10, 2023)

European Digital Rights (2021). Civil society calls on the EU to put fundamental rights first in the AI Act. [online] Available at: https://edri.org/our-work/civil-society-calls-on-the-eu-to-put-fundamental-rights-first-in-the-ai-act/ (Accessed: November 29, 2022)

EuroEFE (2021). Spain approves €70 billion recovery plan to help transform economy. [Online] Available at: https://www.euractiv.com/section/economy-jobs/news/spain-approves-e70-billion-recovery-plan-to-help-transform-economy/ (Accessed: December 7, 2022)

European Institute for Gender Equality (n.d.). Risk assessment and risk management by police. [online] Available at: https://eige.europa.eu/gender-based-violence/risk-assessment-risk-management/areas-improvement/ (Accessed: November 25, 2022)

Garrett, B. and Monahan, J. (2019). Judging Risk. [PDF] Available at: https://judicature.duke.edu/articles/assessing-risk-the-use-of-risk-assessment-in-sentencing/ (Accessed November 29, 2022)

Geiger, G. (2021). How a Discriminatory Algorithm Wrongly Accused Thousands of Families of Fraud. Vice. [online] Available at: https://www.vice.com/en/article/jgq35d/how-a-discriminatory-algorithm-wrongly-accused-thousands-of-families-of-fraud/ (Accessed: November 25, 2022)

González-Álvarez, J.L., López-Ossorio, J.J., Urruela, C., et al. (2018). Integral Monitoring System in Cases of Gender Violence. VioGén System. Behavior & Law Journal. [PDF] Available at: https://behaviorandlawjournal.com/BLJ/article/download/56/65/299/ (Accessed: January 10, 2023)

Heilbrun, K., Dvoskin, J., Hart, S., et al. (1999). Violence risk communication: Implications for research, policy, and practice. *Health, Risk and Society* 1(1): 91–105. DOI: 10.1080/13698579908407009

Helmus, L. and Bourgon, G. (2011). Taking stock of 15 years of research on the spousal assault risk assessment guide (SARA): A critical review. *International Journal of Forensic Mental Health*. DOI: 10.1080/14999013.2010.551709

Holligan, A. (2021) Dutch Rutte government resigns over child welfare fraud scandal. *BBC*. [online] Available at: https://www.bbc.com/news/world-europe-55674146/ (Accessed: November 24, 2022)

Kayser-Bril. N. (2019). Spain: Legal fight over an algorithm's code. [online] Available at: https://algorithmwatch.org/en/spain-legal-fight-over-an-algorithms-code/ (Accessed: November 11, 2022)

La Vanguardia (2019). Portugal reforzará la protección a las víctimas de violencia machista. [online] Available at: https://www.lavanguardia.com/politica/20190207/46286941071/portugal-reforzara-la-proteccion-a-las-victimas-de-violencia-machista.html (Accessed: November 25, 2022)

López-Ossorio, J.J., González-Álvareza, J.L., Loinaz, I., et al. (2020). Intimate partner homicide risk assessment by police in Spain: The dual protocol VPR5.0-H. *Psychosocial Intervention* 30(1). Colegio Oficial de la Psicologia de Madrid: 47–55. DOI: 10.5093/PI2020A16

Martinez Garay, L. (2016). Errores conceptuales en la estimación de riesgo de reincidencia. Revista Española de Investigación Criminológica, 14, 1-31. [PDF] Available at: https://reic.criminologia.net/index.php/journal/article/view/97/94/ (Accessed: November 29, 2022)

Mileusnic, M. (2022). Spain's National Recovery and Resilience Plan. European Parliamentary Research Service. [PDF] Available at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/698878/EPRS_BRI(2022)698878_EN.pdf (Accessed: January 10, 2023)

Minder, R. (2020). Ana Orantes, la mujer cuyo asesinato atroz hizo que España cambiara sus leyes. *The New York times*. [online] Available at: https://www.nytimes.com/es/2020/01/17/espanol/ana-orantes-times.html/ (Accessed: November 29, 2022)

Ministerio de Asuntos Economicos y Transición Digital (n.d.) *Digital Spain 2025*. [PDF] Available at: https://espanadigital.gob.es/sites/agendadigital/files/2022-01/Digital-Spain-2025.pdf (Accessed: November 29, 2022)

Ministerio de Justicia (2009), Organic Act of Protection Measures against Gender Violence. [PDF] Available at: https://violenciagenero.igualdad.gob.es/definicion/pdf/Ley_integral_ingles.pdf (Accessed: November 29, 2022)

Ministerio del Interior (2022). VioGén cumple 15 años con más 700.000 casos y 5,4 millones de valoraciones de riesgo realizadas. [online] Available at: https://www.interior.gob.es/opencms/eu/detalle/articulo/VioGén-cumple-15-anos-con-mas-700.000-casos-analizados-y-54-millones-de-valoraciones-de-riesgo-realizadas/ (Accessed: November 29, 2022)

Moreno Yuste, L. (2015). Sistemas de seguridad en los Centros Penitenciarios. [PDF] Available at: http://derechop-cp62.wordpresstemporal.com/wp-content/uploads/2019/09/SISTEMAS-DE-SEGURIDAD-EN-LOS-CENTROS-PENITENCIARIOS.pdf (Accessed: November 29, 2022)

NL Times (2021). Politicians, parents respond to Dutch Cabinet collapse. [online] Available at: https://nltimes.nl/2021/01/15/politicians-parents-respond-dutch-cabinet-collapse/ (Accessed: November 24, 2022)

OECD (2019). Fraud and corruption in European structural and investment funds a spotlight on common schemes and preventive actions. [PDF] Available at: https://www.oecd.org/gov/ethics/prevention-fraud-corruption-european-funds.pdf (Accessed: December 7, 2022)

Marco Francia, M.P., (2020). Evaluando la peligrosidad criminal. [online] Available at: https://www.researchgate.net/publication/346734133/ (Accessed: November 29, 2022)

Portela, M., Castillo, C., Tolan, S., et al. (2022). A Comparative User Study of Human Predictions in Algorithm-Supported Recidivism Risk Assessment. [PDF] Available at: https://arxiv.org/pdf/2201.11080v2.pdf (Accessed: November 29, 2022)

Savary, P. (2009). Francis Evrard condamné à 30 ans dont 20 ans de sûreté. Reuters. [online] Available at: https://www.reuters.com/article/ofrtp-france-justice-evrard-verdict-urge-idFRPAE59T0ON20091030 (Accessed: November 24, 2022)

Selbst, A. and Powels, J. (2017). Meaningful information and the right to explanation. [PDF] Available at: https://academic.oup.com/idpl/article/7/4/233/4762325 (Accessed: January 10, 2023)

Silver, E. and Miller, L.L. (2002). A cautionary note on the use of actuarial risk assessment tools for social control. *Crime and Delinquency*. DOI: 10.1177/0011128702048001006

Smee, B. (2021). Queensland police to trial AI tool designed to predict and prevent domestic violence incidents. *The Guardian* [online] Available at: https://www.theguardian.com/australia-news/2021/sep/14/queensland-police-to-trial-ai-tool-designed-to-predict-and-prevent-domestic-violence-incidents/ (Accessed: November 29, 2022)

Tiry, E. and Kim, K. (2021). Measuring Risk Assessment Tool Performance. [PDF] Available at: https://www.urban.org/sites/default/files/publication/103863/measuring-risk-assessment-tool-performance.pdf (Accessed: November 29, 2022)

UN Women (2022). Facts and figures: Ending violence against women. [online] Available at: https://www.unwomen.org/en/what-we-do/ending-violence-against-women/facts-and-figures/ (Accessed: November 25, 2022)

Digital Future Society

Acknowledgements

Lead author

Tanya Álvarez leads the Digital Future Society Think tank research on digital divides and digitalisation of the public sector. She advocates for an interdisciplinary perspective of how technology impacts society. She has a degree in art history from Swarthmore College and a master's degree in cultural heritage management from the University of Barcelona.

Interviewer

Pablo Jiménez Arandia is a freelance journalist. He researches and writes about the social impact of technology. He has also produced and directed several journalistic projects, including two podcasts for Digital Future Society. During his career he has also covered a wide array of topics including economics, politics, and migration.

Interviewees (organised by case)
These case studies were based on interviews of the following experts and users:

BOSCO

- Eva Belmonte, journalist and co-director, Civio Foundation
- David Cabo, computer engineer and co-director, Civio Foundation
- Sergio Carrasco, IT lawyer and engineer
- Nerea Caballero, Insercoop
- Mercedes, beneficiary of energy bill subsidy

RisCanvi

- Antonio Andrés Pueyo, Universitat de Barcelona
- Griselda Barris, Department of Justice Generalitat de Catalunya
- Manuel Portela, researcher, Universitat de Pompeu Fabra
- Nuria Monfort, lawyer, IACTA

VioGén

- Pilar Álvarez, journalist, El País
- Sonia Márquez, lawyer, Fundación Ana Bella
- Belén Méndez, survivor of gender violence, Fundación Ana Bella
- Gemma Galdón, founder, Eticas Foundation

SALER

- Alfons Puncel, public servant and former undersecretary of the Ministry of Transparency, Generalitat Valenciana
- Pedro Giménez, Service Inspector, Generalitat Valenciana
- Oscar Capdeferro, professor lector (adjunct professor), Universitat de Barcelona, Department of Administrative Law

Think Tank team

Thank you to the following Digital Future Society Think Tank colleague for their input and support in the production of this report:

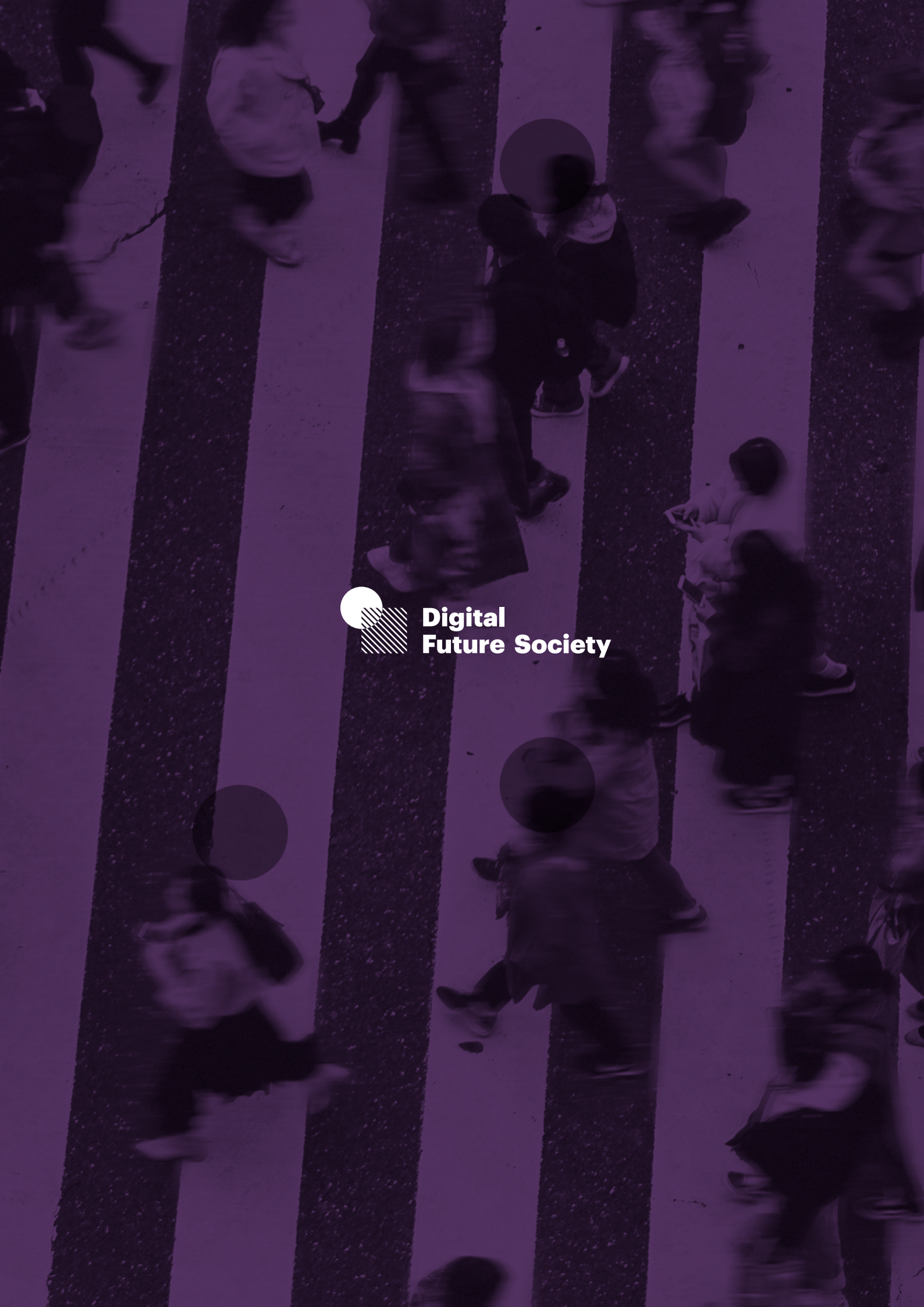
- **Olivia Blanchard**, Researcher, Digital Future Society Think Tank.

Please cite this report as:

- Digital Future Society. 2022. Algorithms in the public sector: four case studies of ADMS in Spain. Barcelona, Spain.

Contact details

To contact the Digital Future Society Think Tank team, please email:
thinktank@digitalfuturesociety.com


Digital
Future Society