eticas
Foundation

2026

Guide to Community-Led **AI Audits**

About This Guide

This guide builds directly on Eticas Foundation's 2023 Adversarial Algorithmic Auditing Guide, which introduced the methodology, vocabulary, and conceptual foundations of impact-focused, adversarial AI auditing. Where that guide focused on the framework, this one focuses on practice: what community-led auditing looks like across six real cases, conducted in four countries, over nearly a decade.

Readers unfamiliar with the broader landscape of AI and algorithmic system types should read the taxonomy in Annex A before proceeding. It provides an overview of the main categories of algorithmic and AI systems, what they do in practice, the contexts in which they are frequently deployed, and their documented or reasonably anticipated negative impact; crucial context for understanding the reasoning and impact of Eticas' community-led AI audits.

About Eticas and Eticas Foundation

Eticas was founded in 2012 by Gemma Galdon Clavell, a political scientist and public policy specialist who had spent years studying the social consequences of surveillance and digital technology. The premise was simple and unusual at the time: that the harms caused by AI and algorithmic systems could be measured, documented, and used to drive accountability, but only if the people measuring the algorithmic systems were willing to work directly with the communities most affected.

Eticas is now a recognized pioneer in independent AI auditing, operating as both a research and private practice (Eticas.ai) and a nonprofit foundation (Eticas Foundation). The Foundation exists to advance the practice of community-led auditing as a public good: producing open methodologies, building field capacity, and conducting audits in the public interest when commercial incentives are absent.

Gemma has served on expert bodies advising the European Commission, the Council of Europe, several UN bodies and the Spanish government on AI policy, and is a founding member of the International Association of Algorithmic Auditors (IAAA). But the work that defines Eticas is not advisory. It is the audits themselves. They are investigations conducted without permission, without access to source code, and often without institutional cooperation, that have nonetheless produced and documented findings on AI harms, built capacity among those impacted by such systems and, in several cases, changed law.

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Acknowlegements

"This guide is the result of many years of work with communities in the many community-led audits Eticas has conducted over the last decade. We would like to thank Mireia Orra, Luis Gonzalez, Evren Yalaz, Toni Lorente, Yung-Hsuan Wu, Matteo Mastracci, Mariana Carvajal, Katsiaryna Vladziochyk, Oliver Smith, Catalina Bernal and Melissa Robles for their contributions to this work at Eticas over the years.

This guide would also not have been possible without the communities who have sought and trusted us with their time and data, including Fundación Ana Bella, Elite Taxi, Iridia, A11, Amnesty International, Cedown Jerez, Taxi Project 2.0, Observatorio TAS, Fundación Secretariado Gitano, and the Reframing Migrants in European Media program.

It is also worth mentioning pioneering women like Julia Angwin, Joy Buolamwini, Timnit Gebru and Cathy O'Neil, whose work in AI auditing and accountability inspired this line of inquire in the first place and continues to inspire us.

Finally, I would like to thank Jose Miguel Calatayud and Alexandra Magaard, who have been fundamental to turning all our experience and past work into the guide you now have in your hands.

Thank you all for trusting Eticas and partnering with us to bring the voices of those impacted by AI to the forefront, and in doing so helping us bring real accountability to the tech sector and a clear path for the effective regulation of innovation."

— Gemma Galdon Clavell, Founder and President of the Eticas Foundation

Table of Contents

Executive summary

In today's global economy, artificial intelligence is driving record amounts of investment and disruption, but also a new industry: AI evaluation, or "auditing" in this case. As AI enters our workplaces, relationships, and governments, it is remarkable how little data there is on AI impacts. In a public debate that seems stuck between celebration and doom, robust and independent evaluation of how AI works and what it does to people are remarkably scarce.

This guide shows how AI evaluation and auditing are at the heart of the possibility of a new conversation around power, agency and impact in the age of automation. It shows how some of the impacts of AI that audits can unearth have never even been covered by media nor considered by regulators. Fear and fascination, and not data, have guided the conversation in ways that have taken society away from building effective controls and protections.

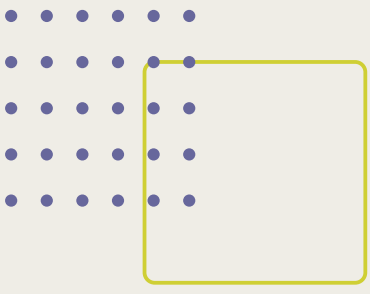
This guide seeks to change this dynamic by showcasing community-led AI auditing as a verifiable, reproducible methodology for meaningful AI governance and demonstrates how via six case studies. Our work in community-led AI auditing spans four countries, 10 years, and highlights multiple gaps in AI governance today.

Findings include **a statistically inconsistent algorithm that decided 7,713 Spanish prisoners' parole,** one of Europe's **largest health insurance companies' algorithms that classifies 50 percent of women with down syndrome falsely as children**, and how **Uber is unavailable for 27% of trip requests if they come from a Roma neighborhood**.

This report highlights the need for end-users (i.e. communities) to be integrated into AI system evaluation and provides practical guidance for how to effectively do so.

In the absence of robust international law, federal US regulation, strict EU regulation, and patchy US-state regulation of artificial intelligence, we encourage philanthropists, activists, investors, non-governmental organizations, non-profit organizations, think-tanks, policymakers, companies, and mission-driven venture capital firms to consider community-led audits as their chosen strategy to impact AI governance, support communities and protect people and rights.

This guide is not only a report but a wakeup call: if every community impacted by AI had audit data on each system working around them, the accountability and liability conversations would be different to what they are now. With independent metrics on AI impacts, communities can open the door to establishing benchmarks, making demands and holding developers accountable.



List of Acronymns

This list provides an overview of the key acronyms and terms used throughout the CLA Guide.

AA . . . . . Adversarial Audit

AI . . . . . Artificial Intelligence

BMI . . . . . Body Mass Index

CNMC . . . . . *Comisión Nacional de los Mercados y la Competencia*
Spain's National Markets and Competition Commission

CLA . . . . . Community-Led AI Audit

CSO . . . . . Civil Society Organization

DSA . . . . . Digital Services Act

DS . . . . . Down Syndrome

ECHR . . . . . European Convention on Human Rights

EU . . . . . European Union

FEDETAXI . . *Federación Profesional del Taxi*
Spanish Taxi Drivers' Federation

FR . . . . . Facial Recognition

FTC . . . . . Federal Trade Commission

GDPR . . . . . General Data Protection Regulation

IP . . . . . Intellectual Property

ML . . . . . Machine Learning

MINRZS . . . *Ministarstvo za Rad, Zapošljavanje, Boračka i Socijalna Pitanja*
Serbia's Ministry of Labour

NGO . . . . . Non-Governmental Organization

PSA . . . . . Pre-trial Services Agency

PHV . . . . . Private Hire Vehicle

ROC . . . . . Receiver Operating Characteristic

SCR . . . . . Social Card Registry Serbia

TAS . . . . . *Observatorio de la Tarifa de Aplicaciones de Servicios*
Platform Fare Observatory

VPR . . . . . *Valoración Policial del Riesgo*
Police Risk Assessment

Introduction and Background

Algorithmic and AI systems now shape some of the most consequential decisions in people's lives. They determine who receives housing assistance, parole, or protective police supervision. They set the price of a taxi ride, health insurance premium and rent. They rank the information people see about health, politics, and their own communities. Often, the people most directly affected by these systems have no idea they exist, let alone any meaningful way to question or contest their outputs.

The question of what to do about it, how to study these systems, hold them accountable, and protect the people they affect most, is one of the defining governance challenges of this decade.

Auditing an AI system is not new. Developers audit their own systems before deployment. Regulators commission evaluations. Researchers publish independent assessments. Each of these serves a purpose. But they share a structural limitation: they examine AI systems primarily as technical objects and from the outside, without sustained engagement with the people, the end-users whose lives the systems influence.

> *Most audits of AI systems examine them from the outside: what they were designed to do, whether they meet technical standards, how they perform in controlled tests. Community-led auditing starts from a different question: what actually happens to real people when these systems are deployed in the real world?*

About this Guide

Community-led AI auditing starts from the opposite premise. The people most affected by an algorithmic system (the women who answered VioGén's questionnaire in a police station, the Roma families who lost welfare benefits through Serbia's Social Card Registry, the taxi drivers whose income was reshaped by Uber's pricing algorithm) possess knowledge about how these systems actually operate in practice that no technical team can replicate. They know which questions feeding the algorithm are ambiguous, which outcomes made no sense for the situation at hand, and which forms of harm the system's designers never anticipated or simply never interrogated.

In practice, this means that in a community-led audit, communities are not consulted at the end to verify findings. They are involved from the beginning: helping us define the research questions, identify which outcomes matter, collect and interpret data, and decide how the audit's findings should be used. The audit is designed from inside the social context in which the system operates, not applied to it afterward.

This participatory approach also reshapes what an audit can find. Internal reviews may identify technical errors; community-led audits surface patterns of harm that only become visible over time and across many individuals. A system may perform exactly as designed while producing outcomes that are systematically unfair and irrelevant for the situation. Measuring that requires the kind of granular, contextual evidence that only affected communities can provide.

However, a note of warning. It is not always sunshine and roses partnering with communities. It can be messy. Community partners have forced our team to rethink our entire approach, modify some of our research questions, consider the implicit bias and unfairness even in careful research questions, or cancel our partnership late in the project. But as messy and complex as the partnership can be, the benefits are worth it.

Many of the AI platforms referenced in this guide are extremely complicated, cutting-edge technologies. The owners want to protect their innovations. The developers resist disclosure. Their operators have institutional incentives against transparency. The people scored, priced or classified by these systems have no formal channels to contest the outputs, and often no knowledge that an algorithm was involved at all. Therefore, the AI systems are difficult to audit.

Community-led auditing is designed to function in exactly these conditions. It does not assume cooperation but relies on observable behavior (what the system produces, for whom, and in what circumstances) to draw conclusions about what the system is doing. Where internal data is unavailable, auditors creatively identify alternatives: public records, freedom of information requests, structured testing, sock puppet accounts, and direct testimony from affected individuals.

> *Communities are not the recipients of our audits. They are the reason the audits exist, the source of the evidence that makes them credible, and the protagonists of every outcome that matters.*

This adversarial approach (also referred to as "black-box auditing") is not about antagonism but pragmatism. The Eticas approach to an audit is to reach out to the institution and partner with them first.If the offer to cooperate is refused, then we conduct an adversarial audit. . Accountability cannot depend on the goodwill of the institution being scrutinized. When that goodwill is absent, as it was in most of the cases documented here, the audit proceeds anyway.

Throughout this guide, the term algorithmic and AI system refers to socio-technical systems in which automated or semi-automated processes play a meaningful role in shaping outcomes that affect people's lives. This includes predictive and classification-based systems that assign scores, categories, or rankings, as well as more recent generative systems whose outputs influence judgment and behavior.

Useful References

For readers who want a systematic overview of the landscape before or after reading the case studies, a reference taxonomy of the main types of algorithmic and AI systems is provided in **Annex A.**

For readers interested in research techniques behind community led audits, a reference taxonomy of the main research techniques Eticas uses for a Community Led Audit is provided in **Annex B**.

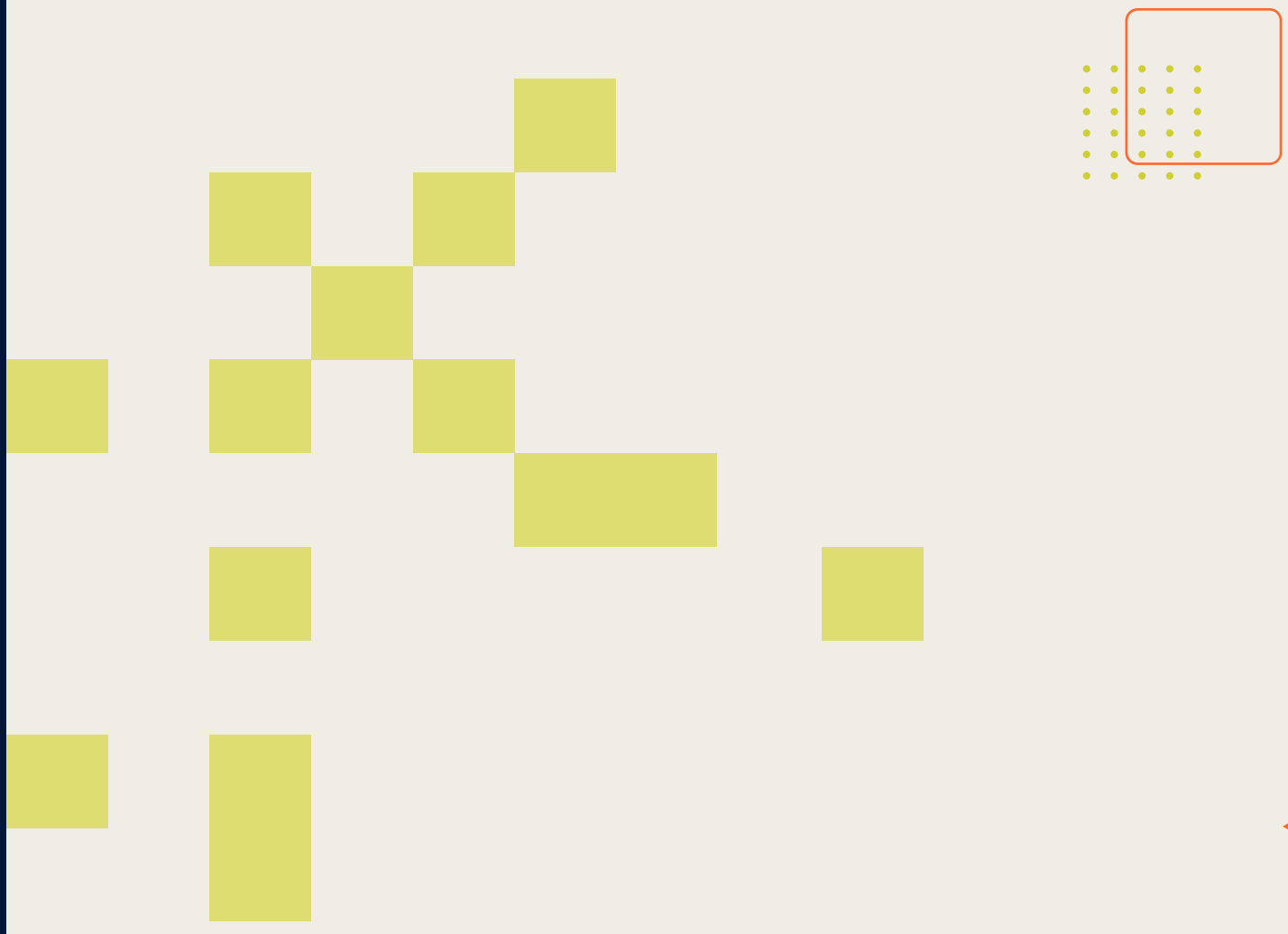
How to Read This Guide

The guide opens with the audit that started it all: Eticas Foundation's community-led audit of VioGén. This first effort started in 2017 and was published in 2021; it reverses engineered the Spanish government's domestic violence risk-assessment system, identifying specific problems within the algorithm and how it was introduced to users. From there, it moves through six further case studies conducted between 2022 and 2024. Each is structured consistently: context, methodology, findings, and recommendations based on how the model performed to observed end-users as a result.

After the case studies, the guide draws cross-cutting insights from the full portfolio: what these audits reveal about how AI systems affect people in practice, what the methodology requires, and what community leadership makes possible that technical analysis alone cannot.

Next, the guide provides a detailed account of Eticas Foundation's ten-step model for community-led auditing, grounded in the cases that preceded it, and concludes with a section on the current opportunity for this work in 2026. Right now, regulatory momentum and a growing funder ecosystem are creating conditions that did not exist even two years ago.

This guide is not a theoretical argument for why community-led auditing matters. It is a record of what it looks like when you actually do it — and what it can change.



I. Where It All Began: Our Community-Led Audit of VioGén

CONTEXT

Spain, 2007. A government under pressure to protect women from gender-based violence, under budget constraints deploys an algorithmic solution. The Integral Monitoring System in Cases of Gender Violence, known as VioGén, is a web application built by the Spanish Ministry of the Interior. When a woman reports gender-based violence to the Spanish police, her case is activated. A police officer administers a 35-question risk questionnaire (the VPR) asking her to confirm which risk indicators apply to her situation. The system processes the answers and assigns a risk score. That score determines how much protection she receives: whether police will check in on her, whether she gets a restraining order reinforced, whether she is placed under 24-hour supervision.

> *Eticas first approached the Spanish Ministry of the Interior in 2018 not as adversaries, but partners. We offered a pro bono, confidential audit. The Ministry met with us only after we enlisted the help of a Member of Congress. The meetings were cordial but no action was taken.*

By January 2022, VioGén had performed over three million risk evaluations. It was, at the time, the largest risk-assessment system in the world by volume of cases. It had nearly 674,000 active cases, including over 69,000 classified as extreme risk requiring active police oversight, and almost all of these cases were decided by a entirely opaque system.

The questions were public, but the algorithm's internal logic was not. No one outside the Ministry knew how the algorithm weighted each answer, why certain responses raised the risk score, or whether the scoring reflected any empirical relationship to actual future harm. No independent audit had ever been conducted. And the Ministry was rumored to be considering adding machine learning to a system that no one outside government had ever been allowed to examine.

Three years later, we decided to act without them.

THE COMMUNITY-LED AUDIT

The decision to proceed without the Ministry's cooperation shaped everything that followed. Without access to the source code, the training data, or the internal weights of the algorithm, Eticas had to build the audit from the outside in: we started from what was observable and worked backward to what could be inferred. This adversarial approach, which has since become

3M+	673,912	69,391	7
risk evaluations performed since 2007	active cases as of January 2022	extreme-risk cases requiring police supervision	recommendations, all adopted into Spanish law

the backbone of Eticas Foundation's methodology, requires two things that no technical team can supply alone: creative data strategy and the trust of affected communities.

For the community partnership, Eticas turned to the Ana Bella Foundation, a Spanish nonprofit working with survivors of gender-based violence, many of whom had personally answered the VioGén questionnaire under police questioning. Ana Bella became the bridge between Eticas and the women the algorithm was designed to protect. Without that relationship, the qualitative dimension of this audit and the things about VioGén that no quantitative dataset could have revealed would not have been possible.

PLANNING PHASE

- Eticas defined four audit dimensions: algorithmic transparency, independent oversight and accountability, end-user treatment, and the implications of potential machine learning integration.
- The Ana Bella Foundation facilitated trust-building with survivors of gender-based violence, the primary affected community, and co-designed the interview framework to protect participant wellbeing.

> *"Half the women interviewed by Eticas evaluated their experience with VioGén negatively. Another third highlighted negative aspects alongside positive ones. All seven lawyers interviewed reported low trust in the system."*

- Because no internal data was accessible, Eticas's team identified a substitute dataset: a public record of over 1,100 intimate partner homicide victims, which contained demographic information and data on prior police reporting and protection status. Because VioGén only operated on women who had reported their partner, and women classified as very low risk received no police protection, this dataset allowed the team to draw inferences about which kinds of cases VioGén had — and had not — flagged for intervention.
- Interview guides were developed for three groups: survivors who had been through VioGén, lawyers specializing in gender-based violence, and former Ana Bella Foundation staff. Protocols prioritized psychological safety and were administered by professionals trained in gender-based violence, not by researchers alone.

EXECUTION PHASE

The audit operated under significant constraints. Notably, the complete inaccessibility of VioGén's source code, training data, or internal weighting logic. Eticas treats this kind of constraint not as a limitation but as a finding in itself: a system that makes life-altering decisions for hundreds of thousands of women, funded by public money, should be auditable. The fact that it was not is a governance failure, not a methodological problem.

Method	Purpose and approach
Quantitative: Homicide dataset analysis	Analysis of 1,100+ intimate partner homicide records cross-referenced against VioGén coverage and police protection status. Statistical examination of which victim profiles were and were not captured by the system, and how demographic factors, including whether the victim had children that was correlated with protection assignment.
Qualitative: Survivor interviews	31 in-depth interviews with women who had gone through VioGén's questioning process. All interviews conducted by gender-based violence specialists. Questions probed the experience of being questioned, the clarity and fairness of the process, and the treatment received from police officers.
Qualitative: Expert interviews	7 interviews with lawyers specializing in gender-based violence; 2 interviews with former Ana Bella Foundation staff. These provided systemic and institutional context that survivors alone could not supply.
Structural / regulatory analysis	Review of VioGén's design documentation, public information about the questionnaire, Ministry communications, and applicable Spanish law on algorithmic decision-making to assess governance and compliance gaps.

FINDINGS

The sum of qualitative and quantitative research was rich enough that, even without access to source code, Eticas was able to draw meaningful conclusions about how VioGén functioned in practice and what it consistently failed to do.

Risk Category Tested	Finding	Scale	Impact
System Performance	VioGén misses the majority of women at risk	73% of intimate partner homicide victims had never reported to VioGén	The system produces precise risk scores for a small, self-selected group — women who had already taken the significant step of filing a police report — while the majority of women killed by partners never entered the system at all. The algorithm's apparent rigor masks a fundamental coverage failure.
Accuracy	Only 1 in 7 women who sought protection received it	2021 data	Among women who did report and were assessed by VioGén, only one in seven received any police protection. The bar for intervention was set so high that the system functioned more as a filter for inaction than as a protection mechanism.

Fairness	Childless women systematically scored lower risk	Consistent pattern across the dataset	Women without children were disproportionately assigned low or no-risk classifications, independent of other factors. No public explanation exists for why the presence of children should be a dominant factor in a domestic violence risk score. This suggests either a flawed design assumption baked into the algorithm or a pattern that no one had ever investigated.
Process integrity	80% of women reported problems during questioning	31 of 31 interview subjects raised systemic issues	Women answered VioGén's 35-question assessment in acute distress, without legal counsel, without psychological support, and with police officers who had received little or no training in trauma-informed practice. Questions were described as 'ambiguous,' 'rigid,' and 'generic.' In 95% of cases, officers simply accepted the algorithm's output without override while remaining formally accountable for a decision that, in practice, the machine had already made.
Accountability	No one was responsible for the decision	System-wide governance failure	Police officers could increase but not decrease the algorithm's risk score. Yet almost never did. Survivors, lawyers, and even officers themselves reported deep ambiguity about who bore responsibility for the outcome: the algorithm or the person. This diffusion of accountability is not a side effect of the system. It is embedded in its design.

IMPACT AND RECOMMENDATIONS

Eticas closed the audit with seven recommendations, structured for three audiences: the Spanish Ministry of the Interior (the funder of the system), law enforcement agencies (the overseers of the system), and the designers of the VioGén system itself (the specific psychologists and developers who created the system).

Audience	Recommendation
Ministry of the Interior	Remove barriers preventing more women from reporting their aggressors, including the procedural, institutional, and social conditions that make reporting inaccessible for most women at risk.
Law enforcement	Require trauma-informed training for all police officers who administer VioGén, and mandate legal and psychological support for women before and during the questioning process.
Law enforcement	Require officers to provide written justification when they accept or modify the algorithm's risk score, ending the pattern of automatic deference to the machine.

System design	Eliminate the 'no risk' classification entirely, so that every woman who enters VioGén is considered at some level of risk.
System design	Publish transparent documentation of how the questionnaire is calibrated and how answers are weighted in the final risk score.
Governance	Establish regular feedback mechanisms involving survivors of gender-based violence and legal professionals in VioGén's ongoing development.
Governance	Subject VioGén to regular, independent audits, including adversarial audits by civil society organizations.

LEGISLATIVE OUTCOME

In January 2025, the Spanish government announced a major overhaul of VioGén. Two of Eticas Foundation's core demands were enacted into law:

- The "no risk" classification was eliminated. Every woman who enters VioGén is now considered at some level of risk.
- Police officers are now required to enter more information about each victim, which officials said would lead to more accurate risk predictions and reduce automatic deference to the algorithm's output.

> *The audit that prompted these changes worked because community members were at the center of the process from the beginning (and because Eticas refused to wait for permission).*

The VioGén audit demonstrated something important: that a credible, evidence-based, community-centered audit is possible even without institutional cooperation. Eticas changed a national law by interviewing 31 women, analyzing a publicly available homicide dataset, and asking the questions that no one inside the system was willing to ask.

Our Transparency Commitment and Open Invitation for Future Audits

All data, code, and quantitative analyses underlying this audit are publicly available. Eticas Foundation publishes its full methodology so that findings can be independently verified and replicated.

View the open repository for this audit -> https://github.com/Eticas-Foundation/VIOGEN-Audit

II. From 2018 to today: Six Community-Led Audits Since VioGén

A. The First Adversarial Audit of an AI Criminal Justice System in Europe

CONTEXT

RisCanvi — risk change in Catalan — is a predictive algorithm that has determined access to parole, the day of release, and length of a sentence for every person incarcerated in Catalonia's prison system since 2009. Developed by a university research group at the request of the Catalan Department of Justice, *RisCanvi* assesses five behavioural risk categories: violent recidivism 2, self directed violence, intra-institutional violence, general recidivism, and breach of sentence. As of 2022, this algorithm directly shaped the futures of 7,713 individuals.

The system works through a dual assessment protocol. On entry to prison, every inmate receives an initial screening (**RisCanvi-S**) using 10 risk factors, producing a low or high classification. Those flagged high risk proceed to a second assessment (**RisCanvi-C**) that is more comprehensive encompassing 43 factors from criminal history, biographical background, family and social context to clinical indicators.

A multidisciplinary team of psychologists, criminologists, and social workers inputs the evidence, which is then processed by a statistical model to generate colour-coded risk scores: red (high), yellow (medium), green (low). Assessments are repeated every six months. Despite its far-reaching impact, *RisCanvi* had never been independently audited, a violation of Spanish regulations requiring audits of automated decision-making systems affecting individual rights since 2016. The system's designers had published studies claiming strong predictive accuracy, but a growing body of critical literature, combined with testimony from frontline staff and lawyers, suggested the system was neither fair nor reliable.

In 2023, Eticas and Iridia (The Center of Defense of Human Rights), a civil society organisation defending the rights of incarcerated people, partnered to conduct the first adversarial audit of a criminal justice AI system in Europe. Concerns about fairness, reliability, transparency, and regulatory compliance motivated the CLA. The project aimed not only to evaluate *RisCanvi*'s technical validity but also to surface its social impacts, drawing attention to the profound asymmetries of information and power faced by those most affected.

7,713
people affected as of 2022

15 yrs
deployed without an independent audit

<5%
of scores ever overridden by Prison staff

13%
positive predictive value in some studies

THE COMMUNITY LED AUDIT

The CLA combined two distinct approaches for a socio-technical design: ethnographic research and comparative output. The partnership with Iridia determined, in part, our methodology: Iridia's relationships with incarcerated people, corrections staff, and legal professionals provided the access, trust, and contextual knowledge of the prison system that no external research team could have replicated.

PLANNING PHASE

- **Community partnership:** Eticas partnered with Iridia to build trust with affected communities, facilitate access to participants, and strengthen the legitimacy of findings.
- **Adversarial audit approach:** Because no direct access to the algorithm or its datasets was available, an adversarial methodology was designed from the beginning: we worked backwards by examining observable outputs to infer internal logic.
- **Mixed-method framework:** A methodology was designed combining quantitative analysis of public recidivism data with qualitative ethnographic research to capture both technical patterns and lived experience.

> *"RisCanvi is like a house that has so many structural defects that it will not be worth rehabilitating. It has to be demolished and rebuilt." — Prison Psychologist interviewed during the CLA*

- **Dataset identification:** A publicly available dataset of 3,651 individuals released from Catalan prisons in 2015 and tracked through 2019 was identified for comparative statistical analysis, yielding 1,889 usable observations after data cleaning.
- **Interview protocol development:** Interview guides were developed for former inmates, families, lawyers, psychologists, educators, and activists, with Iridia facilitating access to the more vulnerable participants.
- **Research question design:** Audit questions were structured around four dimensions: fairness, transparency, accountability, and the social impact of RisCanvi on incarcerated people and their families.

EXECUTION PHASE

The audit was carried out through two complementary methods (qualitative and quantitative), run concurrently and integrated at the analysis stage.

The audit operated under real constraints. *RisCanvi's* source code and internal data were inaccessible. The team relied on publicly available datasets that were incomplete and required substantial variable substitution, and we acknowledged these limitations upfront, which are themselves a finding: a taxpayer funded system affecting thousands of people should be auditable, and the fact that it is not is a governance failure, not a methodological limitation of the auditors.

Method	Purpose and approach
Ethnographic Audit	18 interviews conducted July–October 2023 with former inmates, prison psychologists, social educators, lawyers, a RisCanvi validator, activists, and family advocates. Iridia facilitated access and conducted interviews with vulnerable participants.
Comparative Output Audit	Quantitative analysis of publicly available data on 3,651 individuals released from Catalan prisons in 2015, tracked through 2019. Using 1,889 usable observations, the team applied logistic regression, Venn diagram intersection analysis, factor prevalence analysis, hierarchical clustering, and spectral clustering to reverse-engineer RisCanvi's internal logic.
Reverse Engineering	Because RisCanvi's weights, variables, and calculations are not disclosed, the team worked backwards from outputs — testing whether risk scores could be explained by the stated risk factors and behaviours, as one might reasonably estimate a recipe's ingredients from tasting a cake.

FINDINGS

The audit's core technical question was whether *RisCanvi's* risk scores reflected a coherent, consistent internal logic. Do the factors and behaviours defined in the system actually explain who received high, medium, or low risk classifications? Across multiple statistical approaches, the answer was no.

Risk Category Tested	Finding	Impact
Fairness	No robust relationship between the 43 factors and risk scores	Logistic regressions found no statistically significant relationship between RisCanvi's 43 risk factors and its outcomes. The system could not be reverse engineered because there was no consistent pattern to find, meaning that the system's outcomes are essentially random.
Social Impact	Static factors dominate high-risk scores	"Poor childhood adjustment" is a factor that cannot change regardless of an inmate's behavior, rehabilitation, or growth and was significantly overrepresented among those classified as high risk.
Transparency	No staff training or oversight of the system	Staff did not know what underpinned the risk scores. For example, psychologists did not know how their behavior exam was being used by the algorithm, and lawyers could not challenge classifications they couldn't see. A RisCanvi educator: "We were not explained how it works, and we do not know it."
Social Impact	The system penalizes socially disadvantaged individuals and/or marginalized groups	74% of inmates classified as high risk for violent recidivism had only elementary-level education, a factor prevalent in the prison population generally (76%) but weighted heavily in risk scores. Factors like criminal family history and problematic upbringing were also heavily weighted, factors inmates have no control over.

Fairness	Risk categories do not cluster meaningfully	Hierarchical clustering and spectral clustering both failed to identify distinct groupings in the inmate population based on RisCanvi factors, displaying another example of RisCanvi's inconsistent outcomes. Silhouette scores near zero indicated high overlap between supposedly different risk categories.
Accountability	Regulatory non-compliance	Spanish law has required audits of automated decision-making systems affecting individual rights since 2016. RisCanvi had never been audited in 15 years of operation. The EU AI Act's classification of criminal justice as high-risk AI adds additional legal exposure going forward.

The audit's conclusion was direct: based on available data, RisCanvi is not fair, not transparent, not accountable and in violation of at least two regulations. It is a good example of how a bad introduction of AI can make decisions worse, because AI ends up combining the worst of human intervention with the worst of data.

RECOMMENDATIONS

Our community-led audit was the first independent scrutiny of RisCanvi in 15 years, generating media coverage in Catalonia and Spain, legal and policy attention to the system's compliance failures, and provided Iridia with documented, statistically grounded evidence to support advocacy and potential legal challenges on behalf of incarcerated people. The refusal of authorities to grant access to the algorithm underscored the importance of civil society for filling the gaps in democratic oversight.

To the Catalan Department of Justice

- Immediately disclose indicators, weights, and scoring methodology to legal professionals and the courts.
- Establish robust and understandable protocols guaranteeing inmates' access to legal support and meaningful mechanisms to contest their risk classifications.
- Commission the external, independent, recurrent audits with public reporting required by law since 2016.
- Consider suspension of RisCanvi's use in binding parole decisions pending a full independent review.

To European Policymakers and Regulators

- Treat "human in the loop" as a prioritized design requirement. RisCanvi demonstrates that formal override rights without procedural support and explainability are meaningless in practice.
- Apply the EU AI Act's high-risk requirements proactively to current criminal justice systems, not only to new deployments.
- Mandate that AI systems in high-stakes public sector contexts are auditable as a condition of procurement and continued operation

To Other Auditors

- Reverse engineering is possible even without system access but must be combined with qualitative research.
- Community partnership is crucial in high-stakes settings: Iridia's relationships and trust made the qualitative findings possible.
- Opacity is itself a finding. When a system cannot be reverse engineered, that is evidence of a problem, not a limitation of the audit.

Our Transparency Commitment and Open Invitation for Future Audits

All data, code, and quantitative analyses underlying this audit are publicly available. Eticas Foundation publishes its full methodology so that findings can be independently verified and replicated.

View the open repository for this audit: https://github.com/Eticas-Foundation/RisCanvi-Audit

B. Serbia's Social Card Registry (SCR)

CONTEXT

The Serbian Social Card Registry is a **centralised information system** established under the Law on Social Card (2021) and managed by Serbia's Ministry of Labour, Employment, Veterans and Social Affairs (MINRZS). The Social Card Registry integrates data from multiple registries: populations, taxes, pensions, employment, property, and internal affairs, and creates assessments about whether the Serbian applicants (the beneficiaries) are eligible for government assistance.

Designed to reduce human error, prevent fraud and ensure fairer distribution of government assistance, the system introduces **semi-automated decision-making** for MINRZS social workers, who previously determined eligibility with full discretion. Serbian citizens quickly raised concerns about semi-automating an already limited and underfunded social assistance system.

A11 – Initiative for Economic and Social Rights, our eventual CLA partner, began receiving complaints from beneficiaries who lost assistance after the system was established. This body of evidence led to an eventual **complaint with the Serbian Constitution Court**, an authority comparable to the US Supreme Court. Amnesty International (who has a chapter based in Serbia) and partner organizations submitted detailed legal opinions, and an in-depth investigation into how the Social Card Registry exacerbated poverty, inequality, and barriers in Serbia's welfare system, particularly for marginalized groups such as the Roma community.

The community-led audit was launched in response.

THE COMMUNITY LED AUDIT

The audit combined limited access to anonymized beneficiary-level data from the system, and by combined it with qualitative data from those affected, and contextual data on socioeconomics in Serbia.

- **Contextual Analysis:** Consolidated information about the registry's operation, data flows, and governance framework, identifying where automation occurs, and where human oversight remains.
- **Feasibility Assessment:** Recognised limited system access and suggested mitigating this by combining quantitative and qualitative evidence from multiple sources. The logic was that data from multiple sources is more valid than basing our audit on one source with potential sample bias.
- **Community Partnership:** Established partnership with on-the-ground advocacy organization (A11), who helped us access qualitative data from beneficiary communities and social workers and ensure valid data and legitimate findings.
- **Stakeholder Mapping:** Mapped direct operators (MINRZS, social work centres) and non-operational actors (A11 Initiative, World Bank, NGOs) with indirect influence.

176,000
registered SCR recipients of financial social assistance

47%
of cases falsely excluded Serbian citizens from welfare assistance

40%
of cases misclassified income

20%
of beneficiares excluded without due process

- **Methodology design:** a mixed-methods approach was selected by researchers, combining: a quantitative micro-dataset analysis (N=15 case studies); qualitative analysis of interviews and focus groups with beneficiaries and social workers; and contextual statistical assessment of national data on poverty and social assistance coverage.

Hypothesis	Description
Accuracy	The SCR system misclassifies irregular or non-income transactions (donations, scrapped vehicles, informal earnings) as disqualifying income or assets.
Discrimination	Reliance on inaccurate or incomplete data results in disproportionate harms against marginalized groups, like Roma communities over-represented in social assistance and engaged in informal labor.
Transparency	Beneficiaries receive notifications with no clear explanation for termination; appeals are blocked by lack of access to the algorithm or underlying input data. In multiple cases, social workers themselves reported being unable to override system-generated instructions.
Compliance	The system's functioning may violate domestic law and international human rights standards.[16]

Four research hypotheses were tested:

FINDINGS

Our team found a pattern of statistically significant, consistent inaccuracy, discrimination, almost non-existent transparency measures, and a violation of at least two European international human rights standards.

Risk Category Tested	Finding	Scale	Impact
Accuracy	Wrongful exclusion via misclassification	47% of cases	Funeral donations, irregular transfers, and scrapped vehicles registered as 'significant assets' triggered benefit terminations for households with no meaningful change in economic circumstances.

Compliance	Illegal treatment of seasonal income	40% of cases	Seasonal earnings treated as disqualifying income in direct breach of Serbia's Law on Seasonal Workers, a legal compliance failure with concentrated impact on informal and agricultural workers.
Compliance	Exclusion without due process	~20% of cases	Beneficiaries excluded without a written decision, eliminating any meaningful appeals pathway and violating basic due process protections under domestic and international law.
Discrimination	Disproportionate harm to Roma communities	Structural	Over-representation of Roma in informal labor and social assistance, combined with systemic data gaps, made them disproportionately vulnerable to misclassification and wrongful exclusion.
Transparency + Accountability	Erosion of Human Oversight	Systemic	Social workers reported being unable to override system outputs and facing disciplinary pressure to comply.

Overall, the system did not fulfil the MINRZS's intended goals. Instead of reducing human error, it simply created more non-human-error; instead of creating a fairer distribution of government benefits, it made Serbia's welfare arguably less fair given the discrimination and higher levels of exclusion. In terms of fulfilling MINRZS' goal of reducing fraud, it remains to be seen.

RECOMMENDATIONS

The audit findings directly supported A11's constitutional challenge and provided evidence for advocacy with the Serbian government, institutional donors (who provide a significant portion of the original funding for Serbian's government assistance programs), algorithm developers, and future auditors.

To the Serbian Government

- Make it easier for social workers to override algorithmic outputs and remove all disciplinary pressure to comply
- Redesign beneficiaries' appeals pathway
- Publish plain-language documentation of the system that makes it easier for citizens to understand and engage with
- Introduce periodic, independent audits based on outcomes
- Strengthen the quality of data governance, particularly for informal economy households

[1] Because of the Constitutional Court case that granted us access to a limited set of data, this is a unique CLA for us since it is not entirely adversarial.
[2] The registry is frequently described by A11 and Amnesty International's Serbia office as failing to comply with the European Convention on Human Rights (ECHR) and the International Covenant on Economic, Social and Cultural Rights (ICESCR).

To Institutional Donors (including the World Bank)

- Embed human-rights due diligence in financing frameworks
- Condition disbursements on measurable system transparency benchmarks
- Prioritize on-the-ground capacity-building

To Algorithm Developers

- Adopt user-centred design from the outset
- Implement audit-readiness and explainability features as the standard for all centralized government information systems
- Test systems every year for risk sensitivity

To Future Auditors

- Prioritize community participation and center their feedback in research questions
- Advocate for permanent oversight mechanisms, not one-time audits
- Develop open-data frameworks that enable independent replication

Our Transparency Commitment and Open Invitation for Future Audits

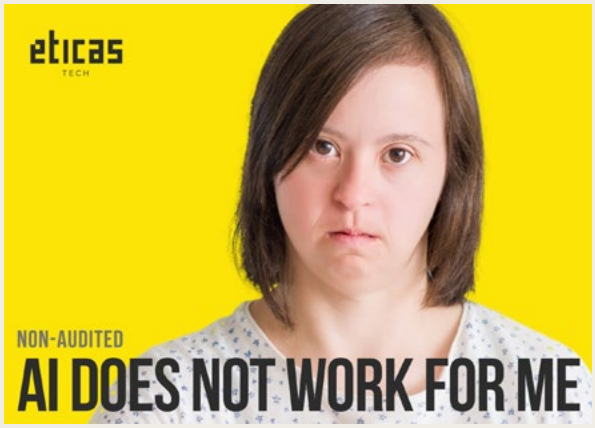
All data, code, and quantitative analyses underlying this audit are publicly available. Eticas Foundation publishes its full methodology so that findings can be independently verified and replicated.

View the open repository for this audit: https://github.com/Eticas-Foundation/Serbia-CLA-Audit

C. Invisible No More: Facial Recognition Systems on People with Down Syndrome

CONTEXT

Inspired by Joy Buolamwini and Timnit Gebru's landmark Gender Shades research that exposed how commercial facial recognition systems fail darker-skinned and female faces, Eticas extended that inquiry into an even less-studied blind spot: people with disabilities. *Invisible No More* takes its name from that lineage, and its methods from the same adversarial auditing tradition.



An estimated around one in one thousand live with Down Syndrome (DS), a genetic condition that involves distinctive facial features that standard FR models are not trained to recognize. In Europe, only half of people with significant disabilities including DS are employed, leaving the other half at elevated risk of poverty and social exclusion, and as access to public and private services becomes algorithmically mediated, these systems leave people with disabilities at further risk of exclusion. The issue is also paradoxical: often, those with significant disabilities are the ones who need public and private healthcare services the most, yet they are excluded from the healthcare industry's new technologies.

Eticas launched a community-led audit of two facial recognition systems to examine how they perform with people with Down Syndrome (DS). **People with DS are among the most underrepresented groups in AI training data**, making them acutely vulnerable to bias in algorithmically mediated systems.

The investigation focused on two systems. The first was Azul, a facial recognition tool developed by Zurich Insurance Group, the world's 98th largest public company, and launched in 2019 as a virtual assistant to estimate a person's age, body mass index (BMI), and smoking status via camera feed, in order to generate a personalized life insurance premium.

The second was DeepFace, a freely available open-source FR framework widely used in commercial applications for demographic and emotional analysis, including age, gender, ethnicity, and emotion classification.

The audit asked whether these systems reproduce or amplify discrimination against people with

40
participants tested

50%
women with Down Syndrome misclassified by DeepFace's gender detection

6
recommendations for Facial Recognition developers, regulators, and auditors

Down Syndrome and whether the use of FR in contexts like insurance pricing is justified at all. Eticas partnered with Cedown Jerez, a Spanish nonprofit organization dedicated to supporting and advocating for the rights of people with Down Syndrome.

THE COMMUNITY LED AUDIT

The audit combined experimental testing with expert input to capture both the technical and social dimensions of bias in facial recognition systems. The methodology integrated qualitative expert knowledge with controlled quantitative testing across two facial recognition systems and two participant groups.

PLANNING PHASE

- **Choosing the systems:** The audit focused on Azul, Zurich Insurance's commercial FR tool for risk assessment and insurance pricing, and DeepFace, the best available open-source dataset of faces for testing facial recognition. Examining both a proprietary commercial system and a publicly available model allowed the audit to measure both immediate real-world outcomes and structural biases embedded in the broader FR ecosystem.
- **Contextual analysis:** Background research confirmed a significant gap in the existing literature: while evidence of FR bias related to gender and race is well documented, bias affecting people with disabilities had been almost entirely neglected. This justified centering the audit on Down Syndrome as a case where exclusion is both measurable and acute.
- **Stakeholder selection:** Semi-structured interviews were conducted with five specialists: a big data engineer and social psychologist, a disability rights activist, a prosecutor with experience protecting people with disabilities, and a member of the European Commission's expert group on responsibility and technologies. Interviewees were selected for their combined technical, legal, social, and policy expertise.
- **Methodology design:** A two-part plan was established. First, an experimental user audit with 40 participants (20 with DS and 20 without) to test Azul's age, BMI, and smoking status estimation. Second, image-based pilot tests using two curated datasets of 60 images each. one of people with DS and one of well-known individuals without DS, to evaluate DeepFace's performance across age, gender, ethnicity, and emotion classification.

EXECUTION PHASE

The audit was carried out through a combination of controlled participant testing and structured expert consultation. Unlike others, this audit is notably structured around comparative performance testing (DS vs. no-DS groups) rather than hypothesis-driven questions, so the research questions are implicit rather than explicitly stated.

- **Expert interviews:** Five semi-structured interviews were held with specialists. Their testimony connected the technical findings to their social and legal consequences.
- **Azul testing:** 40 participants — 20 with Down Syndrome (12 men, 8 women; 1 smoker, 19 non-smokers) and 20 without DS (9 men, 11 women; 8 smokers, 12 non-smokers) — each used Zurich's Azul system while researchers observed and recorded its outputs for age, BMI, and smoking status. Results were compared against participants' actual characteristics to assess accuracy and fairness across groups.
- **DeepFace testing:** Two image datasets were compiled: 60 images of individuals with DS (ages 4–57; 30 women, 30 men; 10 individuals each from six ethnicity categories: Asian, Black, Indian, Latino Hispanic, Middle Eastern, and White) and 60 images of well-known individuals without DS (ages 17–73, with matched gender and ethnicity distribution). Ethnicity, gender, and emotion were manually labelled by the research team. DeepFace was then run across both datasets to identify disparities in classification performance.
- **Analysis:** Statistical measures including error rates, mean absolute error, and recall were applied to compare results across the DS and non-DS groups and to identify systematic patterns of bias.

FINDINGS

The audit found consistent inconsistences when participants with Down Syndrome were tested, across both systems and all measured attributes. The pattern of failures was concentrated on the group least represented in training data.

Risk Category Tested	Finding	Scale	Impact
Accuracy	Age estimation failures	Both systems	Both Azul and DeepFace produced highly inaccurate age predictions for participants with DS, with errors significantly larger than for the non-DS control group.
Fairness + Discrimination	Gender bias in age estimation (Azul)	All women with DS	Azul drastically underestimated women's ages. In extreme cases classifying adults in their twenties as children aged 5 or 8 while consistently overestimating men's ages. Beyond fairness concerns, this creates a direct legal risk: under Spain's Civil Code and the UN Convention on the Rights of the Child, individuals predicted to be under 18 could complete age-restricted insurance contracts without triggering legal safeguards.
Accuracy	BMI estimation errors	Higher for DS group	Azul's BMI error rate was lower than its age error rate but remained significantly higher for participants with DS (mean error 3.82) than for those without (mean error 2.98).

Discrimination	Gender misclassification (DeepFace)	>50% of women with DS	DeepFace correctly classified 100% of men with DS as male, but fewer than half of women with DS were correctly identified as female. In the non-DS dataset, male accuracy remained 100% and female accuracy was 80%.
Accuracy	Emotion misclassification (DeepFace)	Both groups; lower confidence for DS	Overall emotion classification accuracy was similar across groups (0.567 for DS; 0.583 for no DS), but mean confidence in correct classifications was markedly lower for DS participants (8.05 vs. 13.19), indicating the model was guessing rather than recognizing.
Accuracy and Discrimination	Ethnicity misclassification (DeepFace)	Structural	DeepFace performed poorly on ethnicity classification in the DS dataset, frequently misclassifying Asian participants and producing mixed or poor results for Indian and Middle Eastern participants. In the non-DS dataset, accuracy was 100% for Asian and White participants and near perfect for Black participants.
Compliance	Inadequate consent mechanism (Azul)	Systemic	Consent to biometric data collection was buried in small print and triggered by clicking a generic "more information" button, with no meaningful disclosure at the point of interaction. The system operated only in Spanish and directed users to seek detail in a separate privacy policy, placing the burden of comprehension on the participant. This falls short of GDPR's requirement that consent be freely given, specific, informed, and unambiguous, a standard the audit notes is especially difficult to meet for users with cognitive disabilities.

Expert interviews reinforced these technical findings. Exclusion, they noted, happens by design: algorithmic systems group individuals around shared characteristics, and smaller communities who do not fit majority patterns are systematically excluded from the model. Even a system trained on perfectly unbiased data would reproduce bias through underrepresentation in its user base. The use of FR in insurance pricing was also questioned in terms of practicality: asking a person directly about their age, BMI, and smoking status would be faster, more accurate, and far less invasive than a five-minute biometric scan.

RECOMMENDATIONS

As of 2024, Azul was no longer in use, though it is not clear approximately when Zurich discontinued the system. The audit's findings point to systemic failures, and its recommendations are more holistic than other CLA recommendations, addressed to the computer vision industry as a whole, to policymakers, and to future auditors.

1. All developers and implementers of facial recognition technologies must adopt comprehensive and transparent bias mitigation strategies, with specific attention to underrepresented groups including people with disabilities.
2. Accessibility must be prioritised: systems deployed in public or commercial contexts should be required to demonstrate inclusive performance across diverse user groups before deployment.
3. Training datasets must be made significantly more diverse, including meaningful representation of people with Down Syndrome and other disabilities across age, gender, and ethnicity.
4. Meaningful recalibration processes and validation processes should be implemented and documented following GDPR and equivalent frameworks, with results published before and after deployment.
5. Consent procedures for biometric data collection must be made clear, explicit, and genuinely informative that is in full compliance with GDPR and equivalent frameworks.
6. Facial recognition models deployed in high-stakes contexts must be subject to regular independent audits.

Our Transparency Commitment and Open Invitation for Future Audits

All data, code, and quantitative analyses underlying this audit are publicly available. Eticas Foundation publishes its full methodology so that findings can be independently verified and replicated.

View the open repository for this audit: https://github.com/Eticas-Foundation/FacialRecognition_DownSyndrome_Audit

D. Uber, Bolt, and Cabify's Pricing Algorithms

CONTEXT

Ride-hailing platforms like Uber, Bolt, and Cabify have reshaped transportation across the world, promising consumers more choice and lower prices. But the pricing algorithms that underpin these platforms are almost unknown and that creates conditions for predatory practices that harm both consumers and workers.

Surge pricing algorithms adjust standard fares using complex calculations that apply "surge multipliers" when demand outstrips supply. When multiple platforms operating in the same market use similar algorithmic logic, the result can be de facto price collusion, even without explicit communication between firms. Research has shown that platforms with similar algorithmic architectures tend to converge on similar prices, producing outcomes that function like price-fixing regardless of intent.

In Spain, the issue surfaced through formal complaints. In 2018 and 2019, FEDETAXI, a Spanish professional taxi association, filed complaints with Spain's Competition Markets Authority (CNMC), alleging that Uber and Cabify were engaging in price-fixing. The CNMC concluded that the platforms set prices independently, but a dissenting opinion from one board member, combined with parallel allegations of Uber–Lyft price collusion in the United States, left the question unresolved. FEDETAXI approached Eticas, and together they concluded the issue warranted a community-led audit.

Eticas quickly recognized that the audit's scope extended beyond competition law. The same algorithmic systems that were allegedly fixing prices for consumers were also shaping working conditions for drivers through automated performance penalties, opaque earnings calculations, and restrictions on how and when drivers could be paid. This made the audit as much about labor rights as about consumer protection.

THE COMMUNITY LED AUDIT

The audit used a mixed-method approach that combined automated fare collection, statistical analysis, and ethnographic research with platform workers.

PLANNING PHASE

- **Scope definition:** The audit was framed around three central concerns rather than narrow research questions: (1) whether the platforms' pricing algorithms amounted to illegal price collusion (2) whether those algorithms discriminated against lower-income neighborhoods; and (3) whether the platforms were complying with labor protections for drivers.
- **Community partnership:** Eticas partnered with Taxi Project 2.0, representing traditional taxi

3
ride-hailing platforms audited across 15 routes in two cities

3 months
of automated fare data collected every 10 minutes

Higher
fares in low-income vs. high-income neighborhoods for equivalent trips

3
labor violations uncovered: pay, penalties, and gratuities

drivers who filed the original competition complaints, and Observatorio TAS, an organization defending platform economy workers, to ground the audit in affected communities and access hard-to-reach participants.

- **Regulatory context:** Spain's 2021 Rider Law that requires platforms in the food delivery sector to treat riders as employees and guarantees workers' right to algorithmic transparency was used as a legal benchmark, even though it did not technically apply to ride-hailing. The law's requirement that ride-hailing apps operate through licensed Private Hire Vehicle (PHV) companies with separately employed drivers shaped the labor rights dimension of the audit.
- Route selection: Eight routes in Madrid and seven routes in Andalusia were selected for the price-fixing analysis. For the socioeconomic discrimination study, 20 routes were selected across four low-income, two medium-income, and four high-income neighborhoods in Madrid and Málaga.
- Crowdsourcing attempt: Eticas first attempted a crowdsourcing campaign to gather real-user fare data for the discrimination analysis. The tactic failed because the ride-hailing platforms made the data-sharing process prohibitively complicated for users, a finding itself.

EXECUTION PHASE

The audit was carried out through three parallel methods, run concurrently across the audit period from October 2021 to January 2022.

Method	Purpose and approach
Automated fare scraping	Sock puppet accounts submitted automated trip requests every 10 minutes across 15 routes in Madrid and Andalusia over a three-month period (October 2021 – January 2022), scraping the fares offered by Uber, Bolt, and Cabify for each route simultaneously. This enabled direct, time-controlled comparison of prices across competing platforms.
Socioeconomic discrimination analysis	Fares were collected for 20 routes across neighborhoods segmented by median income in Madrid and Málaga. Price per kilometer was calculated for each route and a linear regression analysis was run to identify correlations between fare levels and neighborhood income, testing for systematic price discrimination against lower-income areas.
Ethnographic research with drivers	Semi-structured interviews were conducted with PHV drivers working with Uber and Cabify, and with a PHV fleet company manager. Interviews explored working conditions, algorithmic penalty systems, payment practices, and the structural barriers drivers faced in understanding or challenging platform decisions.

The methodology carried limitations. The fare data provided a snapshot covering a limited number of routes in two cities over three months, and the socioeconomic discrimination study was exploratory rather than conclusive. However, the combination of statistical patterns and qualitative testimony produced findings that were mutually reinforcing and sufficient to support targeted regulatory referrals.

FINDINGS

The audit found evidence of price collusion between platforms, systematic price discrimination against lower-income neighborhoods, and three categories of previously unreported labor violations.

Risk Category Tested	Finding	Impact
Price Coordination	Strong price correlation suggesting algorithmic collusion	A statistically significant positive correlation was found between Uber and Cabify fares, and between Uber and Bolt fares, particularly on Madrid routes. The weaker correlation between Cabify and Bolt, combined with the stronger correlations with Uber, suggests that Bolt and Cabify's algorithms are converging on Uber's pricing as an anchor, producing market outcomes consistent with coordinated pricing.
Discrimination	Higher fares in lower-income neighborhoods	Under particular conditions, customers in lower-income neighborhoods paid higher fares for equivalent trips compared to customers in higher-income areas. The pattern was consistent with what prior research predicted: surge pricing algorithms are sensitive to supply–demand dynamics that vary by neighborhood demographics, producing discriminatory outcomes without any explicit discriminatory intent.
Transparency	Platform design blocked audit crowdsourcing	Eticas's attempt to gather real-user fare data was defeated by platform design choices that made data sharing prohibitively difficult. This is itself a finding: the opacity that prevents users from easily exporting their own data is the same opacity that shields discriminatory pricing from independent scrutiny.
Compliance with labor laws	Automated arbitrary penalties for drivers	Platforms penalized drivers for "excessive and unjustified ride cancellations" without defining what constituted a violation. Drivers could not determine whether a given cancellation had triggered algorithmic restrictions on their ability to work, a penalty system with no transparency, no notice, and no meaningful right of appeal.
Compliance with labor laws	Tip suppression through platform design	Drivers reported difficulties receiving gratuities. Cabify explicitly prohibited drivers from accepting cash tips: a design choice that reduced driver earnings while keeping the platform's take unaffected.

Compliance with labor laws	Poor working conditions and structural opacity	Interviews documented poor working conditions more broadly and revealed a complex platform–PHV structure that made it difficult for fleet managers to manage their operations and for drivers to understand or exercise their rights. The layered corporate structure appeared to function as a mechanism for diffusing accountability.

RECOMMENDATIONS

To Competition Regulators (CNMC / FTC)

- Investigate potential indirect price-fixing by ride-hailing platforms, with specific attention to the use of shared algorithmic pricing architectures that produce convergent market outcomes without requiring direct communication between firms.
- Examine geographic and socioeconomic discrimination in platform pricing as a distinct regulatory concern separate from competition law since our CLA found that lower-income users systematically pay more for equivalent services.
- Treat platform opacity as a regulatory problem: design choices that prevent users from accessing or sharing their own fare data impede both individual rights and independent oversight.

To Labor Regulators and Legislators

- Extend the employment protections of Spain's Rider Law and equivalent gig worker protections elsewhere to self-employed PHV drivers working for ride-hailing platforms, closing the loophole that currently excludes them.
- Mandate algorithmic transparency for platform workers: platforms must be required to disclose how their systems make decisions, assess performance, profile workers, and determine pay and drivers must have a meaningful right to contest algorithmic penalties.
- Prohibit platform design choices that suppress driver compensation, including platform-enforced restrictions on receiving gratuities.

To Platform Companies

- Disclose the logic of surge pricing algorithms to regulators and, in accessible form, to consumers and drivers.
- Establish clear, published definitions of the behaviors that trigger algorithmic penalties, with a transparent process for drivers to understand and appeal decisions.

Our Transparency Commitment and Open Invitation for Future Audits

All data, code, and quantitative analyses underlying this audit are publicly available. Eticas Foundation publishes its full methodology so that findings can be independently verified and replicated.

View the open repository for this audit: https://github.com/Eticas-Foundation/TAXI-COMP-AI

E. Uber's Pricing Algorithms in Roma Madrid Neighborhoods

CONTEXT

In a previous audit, Eticas had uncovered evidence that the largest ride-hailing companies in Europe were subtly discriminating against poorer neighborhoods, where customers were offered more expensive prices for equivalent rides. Three years later, Eticas launched a follow-up. That earlier inquiry covered multiple platforms including Uber, Bolt, and Cabify; this follow-up focused specifically on Uber.

The Roma are the largest ethnic minority in Europe, and Spain hosts one of the biggest Roma communities, totaling 1.6 percent of the Spanish population. The Roma face persistent structural barriers, including poverty, poor infrastructure, spatial segregation, stigmatization, and limited access to public transport. Critically, Roma settlements are typically located on urban peripheries, often only partially covered by public transport routes, meaning ride-hailing is not a convenience for Roma communities but frequently their primary means of accessing the city. Notably, Roma communities in Spain became permanently settled in urban areas in part through EU integration projects, making their continued exclusion from algorithmically mediated services not a failure of integration but a new layer of discrimination imposed on top of it.

This follow-up audit focused on just Uber and sought to measure whether Uber offers different levels of service reliability for trips originating from Roma neighborhoods as compared to those from non-Roma neighborhoods. Given the Roma's high dependence on affordable and reliable mobility, the audit set out to test whether three machine learning algorithms underpinning the Uber platform:

1. supply–demand prediction,
2. surge pricing,
3. driver–rider matching,

provide services of equal reliability in Roma and non-Roma neighborhoods.

THE COMMUNITY LED AUDIT

The planning phase focused on identifying the right system to audit, grounding the work in the lived realities of Roma communities, and setting up a robust yet practical methodology. This stage ensured that the audit was both scientifically rigorous and socially relevant.

PLANNING PHASE

- **Choosing the system:** Uber was selected because of its central role in urban mobility in Spain and its reliance on multiple algorithmic systems that directly impact access to rides.

27%
of trip requests in Roma neighborhoods, Uber was unavailable

1.4x
Longer wait time for Roma riders

- **Contextual analysis:** Desk research and prior Eticas audits highlighted concerns about fairness in ride-hailing platforms. Input from Fundación Secretariado Gitano, a Spanish organization defending and promoting the rights of the Roma people, helped frame Roma communities' vulnerability to mobility barriers.
- **Stakeholder mapping:** Stakeholders included Roma community members, Uber riders, drivers, regulators, and consumer protection bodies.
- **Methodology design:** Researchers selected 10 Roma settlements in Madrid, which has the second-highest Roma population in Spain, and 10 nearby control neighborhoods (7 socioeconomically well-off municipalities near Roma settlements, and 3 were neighborhoods from the previous audit). For each settlement, three destination types were chosen (city center, nearest commercial center, nearest hospital), creating 60 routes for comparative testing.

EXECUTION PHASE

The execution phase translated the audit plan into concrete testing, combining controlled data collection with comparative analysis. The goal was to detect whether Uber's service delivery differed systematically between Roma and non-Roma neighborhoods.

> *the findings point to the role of Uber's algorithmic systems in reinforcing unequal access to mobility resources and raise serious concerns about Roma people's right to mobility as a public service.*

- **Data collection setup:** Researchers created sock puppet accounts to simulate rider behavior and submitted 240 trip requests across 20 neighborhoods at four times of day (rush and off-peak).
- **Metrics recorded:** Ride availability ("ride offered or not") and estimated wait times for driver pickup were logged for every request.
- **Comparative testing:** Results were disaggregated by neighborhood type, destination, and time of day. Chi-squared and Welch's t-tests were used to measure statistical significance.

FINDINGS

The audit revealed clear disparities in Uber's service between Roma and non-Roma neighborhoods in Madrid. In Roma areas, Uber was unavailable for 27 percent of trip requests, while in non-Roma areas rides were always available. Even when service was provided, Roma riders had to wait 1.4 times longer than riders in other neighborhoods.

These findings carry weight given the absence of alternative transport options for many Roma communities. An important methodological caveat applies: without access to the platforms' internal algorithmic code and data, the research could only approximate the inner workings of those systems, a limitation that itself underscores the audit's central policy recommendation. Taken together, the findings point to the role of Uber's algorithmic systems in reinforcing unequal

access to mobility resources and raise serious concerns about Roma people's right to mobility as a public service.

RECOMMENDATIONS

1. Greater algorithmic transparency to unveil the full-scale impact of these algorithmic systems on urban mobility, especially from the point of view of the most marginalized communities.
2. Legal frameworks must be established to empower experts and researchers to conduct independent audits of these platforms.

Our Transparency Commitment and Open Invitation for Future Audits
All data, code, and quantitative analyses underlying this audit are publicly available. Eticas Foundation publishes its full methodology so that findings can be independently verified and replicated.

View the open repository for this audit: https://github.com/Eticas-Foundation/Ridehailing-Roma

F. YouTube and TikTok's Immigration Recommendations during the US Midterm Elections

CONTEXT

YouTube and TikTok are the world's two most popular video platforms, with an estimated 2.58 billion and 1.99 billion users respectively. Because video content generates stronger emotional responses than text or audio, both platforms are unusually powerful in shaping users' attitudes toward complex social and political questions, including immigration.

YouTube's recommendation algorithm suggests videos **based on user viewing history, engagement, and demographic**.

TikTok's recommendation algorithm suggests videos **using a hybrid method combining user-based and collaborative filtering**. Both are designed primarily to maximize engagement, a core objective for their business model that research has consistently linked to the amplification of emotionally charged, divisive, or misleading content.

Eticas audited both platforms to study how their recommendation systems portray immigration and how that might shape user attitudes. Without access to either platform's internal code or data, the adversarial audit focused on observable outputs, combining automated data collection with qualitative community research.

THE COMMUNITY LED AUDIT

The audit was conducted in two parts: a study of YouTube's recommendation and search algorithms, undertaken as part of the EU-funded Reframing Migrants in European Media project (2022–2023), and a study of TikTok's recommendation algorithm, conducted during the November 2022 U.S. midterm elections. Both used a socio-technical approach combining automated data collection with community-grounded qualitative research.

PLANNING PHASE

- **Feasibility assessment:** Because neither platform provides access to its algorithmic code or data, the audit was designed around what could be directly observed: the outputs served to different types of users under controlled conditions.
- **Community collaboration:** Eticas collaborated with migrant community members and the Reframing Migrants consortium, a coalition of media and civil society organizations, to shape the audit's research questions and ground the technical findings in lived experience.
- **Research question design — YouTube:** Four questions were defined: (1) How are migrants and refugees represented in top-watched YouTube search results? (2) Do algorithms suggest

4.57B
combined users across YouTube and TikTok

2 of 3
top YouTube search result framings depicted immigrants negatively or as a burden

0
significant variation by user location or imimmigrant background on YouTube

~0
political immigration content served by TikTok regardless of user profile

differently framed content across national settings? (3) Do algorithms suggest differently framed content to migrant versus non-migrant accounts? (4) How do immigrants themselves perceive the portrayal of immigrants on YouTube?

- **Research question design — TikTok:** Three questions were defined: (1) Does recommended content vary by users' attitude toward immigration? (2) Does it vary by the political leaning of the user's location? (3) Does it vary over time across the course of the U.S. midterm election cycle?
- **Methodology design:** A sock puppet methodology was selected for both platforms, supplemented by content analysis and ethnographic research with migrant community members. Sock puppet accounts allow controlled comparisons of what different types of users are shown.

EXECUTION PHASE

The two platform audits were executed using parallel yet customized methods, reflecting the different architectures of YouTube and TikTok.

Method	Purpose and approach
YouTube: Sock puppet & scraping	Sock puppet accounts were created representing migrant and non-migrant users, located in Toronto (Canada) and London (UK). Scrapers collected thumbnails, titles, and descriptions for the top-watched and recommended videos when each account searched for "migrants" and "refugees." This allowed direct comparison of what different users were shown.
YouTube: Content analysis	Collected thumbnails, titles, and video descriptions were coded using content analysis to study how migrants and refugees were represented visually and textually across the platform's outputs.
YouTube: Ethnographic research	A roundtable at the "Decolonizing the Newsroom" event in Madrid (July 2022) brought together four participants with immigrant backgrounds, one moderator, and two researchers. Participants discussed media representation of immigrants and reacted to preliminary audit findings.
TikTok: Sock puppet & training	Nine sock puppet accounts were created across three U.S. locations. Accounts were trained over six days by watching, liking, and sharing content, and following creators with different attitudes toward immigration, to simulate users with distinct immigration-related preferences.

TikTok: Longitudinal scraping	The first 20 recommended videos on each account's "For You" feed were scraped at three points: before, during, and after the November 2022 U.S. midterm elections, enabling temporal as well as cross-profile comparison.
TikTok: Content analysis	Video content was coded to study how immigration was treated in political discourse, whether content varied by user profile or location, and how recommendations evolved across the election period.

Both audits acknowledged significant methodological limitations.

The YouTube audit was restricted to thumbnails, titles, and descriptions rather than video content itself; used only two single-word search queries in English; and collected data during an atypical political moment — June and July 2022 — when Ukrainian refugees were being welcomed in Western Europe on markedly different terms than other refugee populations, likely skewing the "refugees" results.

The TikTok audit was limited to three U.S. locations, three data collection points, and a six-day account training period; and used the desktop web version of TikTok rather than the more widely used smartphone app, which may behave differently. These limitations are acknowledged transparently as constraints inherent to adversarial auditing without platform access.

FINDINGS

Across both platforms, the audit found that recommendation algorithms consistently produce and reinforce negative, dehumanizing representations of immigrants but through different

Risk Category Tested	Finding	Impact
Discrimination	Immigrants systematically dehumanised on YouTube	Top-watched videos consistently portrayed migrants negatively and in dehumanised ways, predominantly as non-white individuals crossing borders. Refugees, by contrast, were shown as small groups of white individuals with visible faces, a racialised distinction in how the platform frames the two groups.
Discrimination	Racialised framing is consistent, not incidental	The distinction between how migrants and refugees are visually represented is not random variation. It appeared consistently across both personalised and non-personalised feeds: migrants shown as anonymous non-white groups in motion; refugees shown as identifiable white individuals. This pattern reflects a structural framing choice embedded in the platform's outputs, not a one-off result.
Fairness	Immigration portrayed as victimisation or national burden	Content frequently framed migrants either as victims or as a problem for receiving countries, with little representation of migrant agency, contribution, or perspective. Neither framing reflects the complexity of immigration as lived experience.

Fairness	No meaningful personalization by location or user background (YouTube)	Results were near-identical for users based in Toronto and London, and for migrant and non-migrant accounts. This refutes the hypothesis that geographical location and user imimmigrant background are significant factors in YouTube's algorithm, and suggests the platform's baseline framing choices drive outcomes more than personalisation.
Transparency	Political immigration content almost absent on TikTok	Regardless of user attitude toward immigration, location's political leaning, or timing relative to the midterm elections, political discourse on immigration was virtually absent from TikTok's "For You" feeds. The platform's algorithm systematically prioritises entertainment over political content.
Accuracy	No variation by user profile or election timing (TikTok)	Content recommendations showed little variation across the nine accounts despite their different trained profiles and locations. The algorithm did not amplify pro- or anti-immigration content based on user characteristics. Rather it simply did not surface it at all.
Compliance	Algorithmic opacity prevents accountability	Neither platform disclosed how its recommendation algorithm functions. Without access to internal code, weights, or training data, independent verification of these systems remains impossible. The outputs are observable; the mechanisms driving them are not.
Transparency	TikTok's algorithm is not purely algorithmic	A Forbes investigation cited in the audit revealed that TikTok employees manually push selected videos to guaranteed view thresholds through a practice called "heating." This means the "For You" page that is presented as a personalized, algorithmically ranked feed is also subject to undisclosed editorial intervention, further limiting transparency and independent accountability.

mechanisms. YouTube's algorithm surfaces a uniform landscape of negative framing that does not vary meaningfully by user profile. TikTok's algorithm suppresses political content on immigration almost entirely, regardless of user characteristics.

RECOMMENDATIONS

The audit produced findings applicable to the platforms, policymakers, and civil society organizations in both the European and U.S. contexts. Given that both YouTube and TikTok are globally deployed platforms that operate outside meaningful independent oversight, the case for structural reform is not jurisdiction-specific.

To YouTube and TikTok

- Publish clear, accessible documentation of how recommendation algorithms make decisions, including what signals are weighted, how content is classified, and how edge cases are handled.
- Grant access to independent researchers to conduct second party and third party, impact-

focused audits of recommendation systems, particularly in high-stakes domains such as immigration, health, and electoral politics.
- Actively address the racialized framing patterns identified in this audit by reviewing content labelling, search result ranking, and the training data underpinning recommendation models.
- Co-design community guidelines and their enforcement principles with immigrants and other affected communities — not merely consult them — to address documented patterns of selective suppression that disproportionately affect the groups at the center of political debates.

To Policymakers and Regulators

- Establish legal frameworks that mandate platform transparency and enable independent algorithmic audits in the United States, where no such framework currently exists.
- Require that platforms operating in public interest contexts, particularly around elections and immigration, demonstrate that their recommendation systems do not systematically amplify harmful or dehumanizing content.
- Establish collaborative efforts among public bodies, platform researchers, and minority communities to define the standards that recommendation algorithms should meet in the treatment of vulnerable groups.

To Civil Society and Researchers

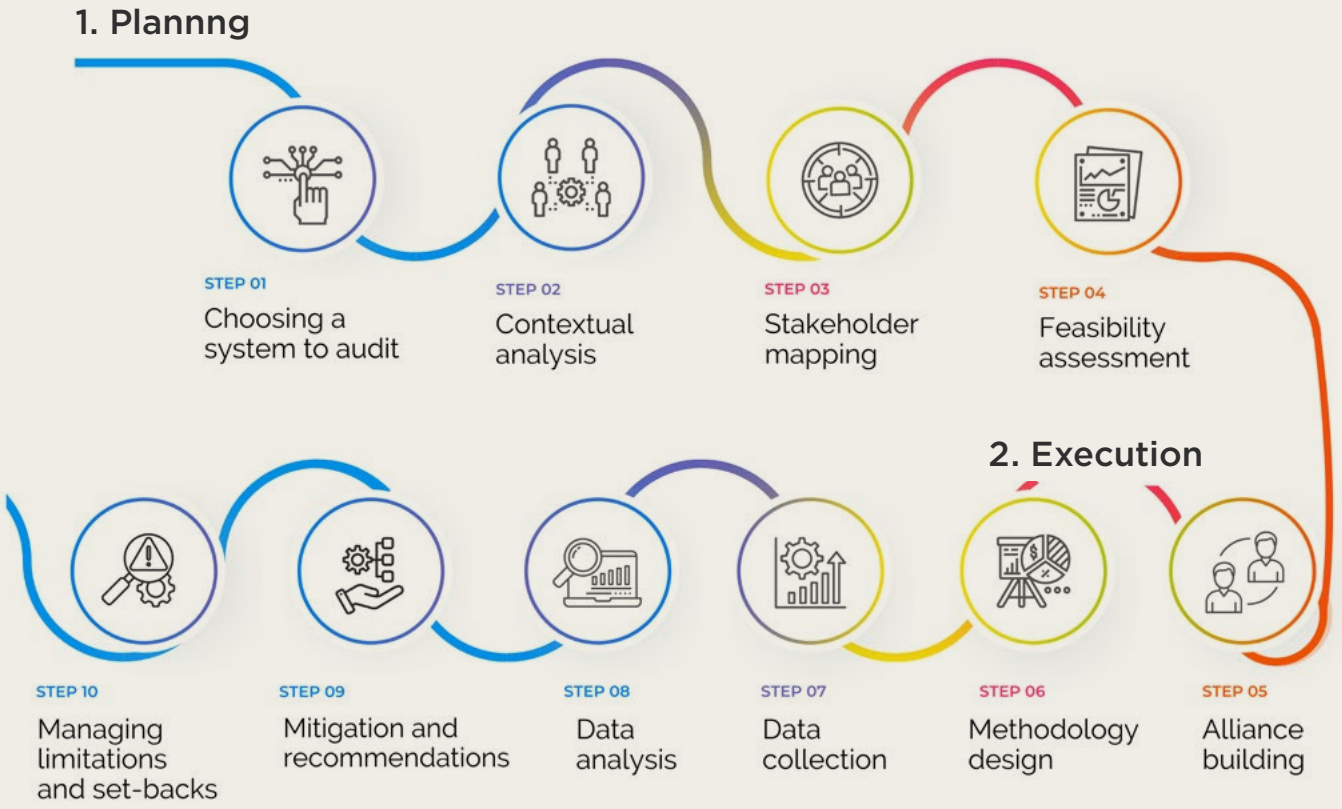
- Develop shared, open methodologies for auditing recommendation algorithms across platforms and jurisdictions, reducing the duplication of effort that adversarial auditing currently requires.
- Integrate community members into audit design from the outset, not as subjects but as co-investigators.

Our Transparency Commitment and Open Invitation for Future Audits

All data, code, and quantitative analyses underlying this audit are publicly available. Eticas Foundation publishes its full methodology so that findings can be independently verified and replicated.

View the open repository for this audit:
https://github.com/Eticas-Foundation/USElections2024-Audit

III. Eticas' 10-Step Model for Creating a Community-Led Audit



Eticas has developed a 10-step model for community-led audits, based on its experience in the field. However, it is important to note that the order of the steps is not mandatory and should be considered as a guide rather than a set of strict instructions. Depending on the specifics of each case, the order of steps may vary, and some steps may even be unnecessary.

For ease, we divide the 10 steps into two main phases: planning and execution. The table below each step within both phases.

A. Planning

The planning phase is the foundation of every community-led audit. Done well, it determines whether the audit will produce findings that are credible, significant, and actionable. Done poorly, it produces a study that may be methodologically sound but disconnected from the realities it was meant to examine. The six steps below are not a rigid sequence — in practice, they overlap and feed back onto each other — but each one must be addressed before the audit moves into execution.

Step 1: Choose the System to Audit

Start by taking stock of the algorithmic and AI systems operating within the communities you work with. In some cases, this will be obvious: a particular system is already generating complaints, legal challenges, or press coverage. In others, it requires active investigation, consulting experts,

PLANNING	
Choosing a system to audit	Listening to communities to identify an AI system with social impact and an initial feasibility check for identifying possible access points to the algorithm(s) for an audit
Contextual analysis	Building understanding about the AI system and the community's experience of it, the context in which it operates and the possible negative impacts it may lead to, through discussions and interviews as well as technical and policy research
Stakeholder mapping	Identifying all relevant parties to an AI system, such as the developers and implementers of the system and the communities affected directly or indirectly by it
Feasibility assessment	Data mapping to determine if the auditor can obtain sufficient information about an AI system via legal means within the relevant jurisdiction
Alliance building	Participatory research design with communities and civil society organizations to ensure that the perspectives of affected groups are incorporated in the auditing process
Methodology design	In consultation with communities, defining the scope of the audit, the research questions, the methods to investigate them, the utility of the results, and the timeline of the project

EXECUTION	
Data collection	Safe and consentful11 qualitative and quantitative data gathering about the inputs, outputs and societal impact of an AI system via specialized techniques for adversarial algorithmic auditing and social science research methods
Data analysis	Translating raw data into meaningful insights via quantitative and qualitative data analysis
Mitigation and recommendations	Providing actionable audit outputs, including reports, metrics, visualisations, and recommendations that serve community leadership in demanding accountability and improvement from developers, implementers, and policymakers
Managing limitations and set backs	Strategies for how to cope when a CLA doesn't go according to plan or simply has to work within a more limited scope than is desirable given the community's goals.

reviewing public registries of algorithmic systems (particularly government registries), and listening carefully to how community members describe the decisions being made about them.

Useful orienting questions at this stage:

4. What algorithmic systems have been introduced in this context, and for what purpose?
5. Are particular services that people rely on — housing, benefits, transport, credit — being run or influenced by an algorithm?
6. Is there a government-deployed system that community members interact with but know little about?

7. Is a current public debate connected to the use of an AI or algorithmic system?

Once candidate systems are identified, prioritize based on two criteria: the severity of the known or suspected negative impacts, and the practical feasibility of conducting the audit. If developers or deployers deny access to the system's code, a third-party adversarial audit is the only option, and in most of the cases documented in this guide, that is precisely what was required. Denial of access is not a reason to stop. It is a reason to be methodologically creative.

Step 2: Contextual Analysis and Stakeholder Mapping

Before any technical work begins, build a detailed understanding of the environment in which the algorithm operates. This means researching the social, political, economic, legal, and institutional context and understanding how the algorithm is embedded within it. The contextual analysis and stakeholder mapping inform each other and should be developed together.

The contextual analysis enables auditors to move from general concern to specific hypotheses. For a recidivism risk-scoring system, a hypothesis might be that the algorithm systematically overestimates risk for people from demographic groups. For a pricing algorithm, it might be that fares or rates are higher in lower-income neighborhoods independent of demand. Hypotheses should be grounded in what is already known about the social context, not imported from other cases without adjustment.

The stakeholder mapping identifies everyone with a relevant relationship to the system:

8. Developers and deployers of the algorithm
9. Those who operate it day-to-day (e.g. police officers, social workers, platform workers, loan officers)
10. Relevant public authorities, regulators, and oversight bodies
11. The intended target populations the system was designed to serve
12. Communities directly or indirectly affected, including groups who may have been overlooked in the system's design

The stakeholder map will evolve as the audit progresses. What matters at this stage is identifying who knows what, who has what kind of power over the system, and who bears the consequences of its outputs.

Step 3: Technical Pre-Audit Research and Feasibility Assessment

This step addresses the algorithm's inner workings or, more precisely, what can be understood about them without internal access. The goal is a data-mapping exercise that identifies every available source of information about how the system functions.

Key questions for the data-mapping exercise:

13. Has previous research been published on the technical side of this system?
14. Can inputs and outputs be systematically observed (e.g. through interactions with community members who use it)?
15. Is any part of the source code, model documentation, or training data publicly available?

16. Can additional information be obtained through Freedom of Information requests or human sources?

When examining training data (where accessible), auditors should be alert to the following well-documented bias types:

Bias type	What it means in practice
Omitted variable bias	Relevant factors are missing from the model because designers lacked knowledge or foresight about the population being scored.
Historical bias	Training data reflects past inequalities and discrimination, which the algorithm learns and reproduces. This often amplifies them.
Population bias	The system was designed or trained on data from one population and deployed on another (i.e. most commonly, a model built on majority-group data applied to everyone).
Aggregation bias	The model draws conclusions about individuals based on group-level patterns that do not apply to them.
Accessibility bias	The system fails to account for the specific needs of people with disabilities, elderly users, or those with limited digital literacy.
Techno-solutionist bias	The algorithm was deployed as a solution to a problem that could have been addressed more appropriately through non-technical means.

This list is a starting point, not a ceiling. Every system and context will reveal its own patterns. Document what you find, note what you cannot determine, and treat the limits of your access as a finding in themselves. Perhaps it is a publicly funded system whose logic cannot be independently examined is a governance failure, regardless of what the technical analysis ultimately reveals.

The final stage of the data-mapping exercise is the feasibility assessment. Before committing to the full audit, the team must honestly ask

17. Is there sufficient information and data — direct or indirect — to conduct a meaningful audit?
18. Does the team have the capacity, time, and expertise required for the planned scope?
19. Will the audit produce results that are informative, significant, and actionable? Put differently, will the findings be new, meaningful, and capable of supporting realistic recommendations?

If these questions cannot be answered affirmatively, return to the preliminary research phase or conclude that a quality audit is not achievable at this time. A constrained audit that is honest about its limitations is more valuable than a comprehensive audit that overreaches its evidence base.

Step 4: Build Community Partnerships

If community members have not already been involved in the audit by this point, that changes now. This step is not optional, and the quality of the entire audit depends on it.

Identify affected community members and reach out, either directly, or through civil society

organizations that already have established relationships with the community. In contexts where direct contact is difficult or potentially harmful, like when the audit concerns a system used in immigration enforcement, or where community members face retaliation risks, a trusted intermediary CSO can conduct interviews and facilitate participation while protecting individuals.

The terms of every collaboration should be formalized in writing: what each party contributes, what they receive, how findings will be attributed, and how decisions about publication and advocacy will be made. All contributors should be fairly compensated for their time, especially community members who participate in data collection.

Partnerships may also extend to academic researchers, domain experts, investigative journalists, and legal advocates. The more multidisciplinary the team, the more dimensions of a complex system can be examined. The constraint is always realistic scope: a partnership that overcommits will underdeliver.

Step 5: Design the Research Methodology

The methodology is where the audit's hypotheses are translated into a testable research design. It defines scope, research questions, methods, and timeline.

Scope decisions determine which population, geography, or time period the audit covers. A recommendation algorithm study might focus on users in a particular region during an election window. A pricing algorithm audit might focus on a specific set of neighborhoods. Defined scope makes findings specific and defensible.

Research questions should take the audit's hypotheses and make them concrete and verifiable. For each hypothesis, multiple questions should cover different dimensions, like statistical patterns, lived experience, and institutional or design choices. A well-designed hypothesis generates questions that can only be answered by combining quantitative and qualitative evidence:

Example hypothesis	Example research questions
A recommendation algorithm systematically underrepresents certain communities.	How does the visibility of those communities compare to majority groups in top-ranked outputs? How do users from those communities perceive the fairness of what they are shown? What role do engagement metrics play in determining visibility?
A pricing algorithm charges higher rates in lower-income or minority neighborhoods.	Do fare or rate patterns differ by neighborhood income level or demographic profile? Do affected riders report awareness of or experience with pricing disparities? What does the platform's documentation say about the factors that determine price?

Timeline: a typical community-led audit at Eticas takes between three and six months. The timeline should be realistic for the research, analysis, report production, and any planned follow-up activity.

Step 6: Plan the Audit Outcomes and Follow-Up

Before execution begins, decide what the audit will produce and who it is for. The primary output of most audits is a comprehensive report. However, a report is not the only outcome, and planning for follow-up activity from the outset makes the difference between an audit that generates attention and one that generates change.

A standard audit report includes: the social context of the system; the methodology used and its limitations; findings presented as a socio-technical analysis; and recommendations tailored to specific audiences. Eticas structures recommendations for seven distinct recipients:

Audience	What recommendations should address
Developers and deployers	Technical and operational changes to reduce harm and improve how the system affects people.
Policymakers and regulators	Political and policy measures governing how and whether the system should be deployed, and under what conditions.
Civil society organizations	Practical strategies to monitor the system's ongoing impacts and sustain community engagement.
Affected communities	How to recognize signs of harm, understand how the system operates, and engage with accountability processes.
Researchers and journalists	How to collaborate with affected communities in future audits and investigations.
The wider public	How algorithmic systems affect daily life and what practical steps people can take to protect their rights.
Funders of civil society	How to support community-led auditing and the organizations that conduct it.

Recommendations should be thoughtful, specific, and realistic; they must be adapted to what each audience actually has the power to do. Vague recommendations produce vague responses. The case studies in this guide consistently show that the most durable changes — changes to law, to system design, to institutional practice — came from recommendations that were precise enough to be implemented and attributed.

Beyond the report, consider: which findings warrant public communication? Are there datasets generated during the audit that can be made publicly available without violating privacy or applicable law? Are there follow-up audits, legal challenges, or advocacy campaigns that the findings could support? Planning these pathways before the execution phase begins means the audit's conclusions are ready to travel the moment they are reached.

B. Execution

Once the planning phase is complete, execution begins. This is where the research design meets the reality of the system and where the community partnership established in planning becomes the audit's operational backbone. Execution has three phases: data collection, data analysis, and producing the audit outputs.

Step 7: Data Collection

Data collection follows directly from the methodology designed in Step 5. What it looks like will vary: sock puppet accounts testing a recommendation algorithm, structured interviews with welfare recipients, a freedom of information request for a government risk-scoring system's design documentation, or a statistical analysis of a publicly available dataset. Most community-led audits use several of these techniques in combination.

Whatever the methods, two principles govern collection. First, all data must be gathered legally and ethically, with informed consent from research participants and appropriate protections for sensitive information. Second, community members are not just sources of data; wherever possible, they are participants in collecting it. Eticas regularly hires members of affected communities to conduct interviews and participate in structured testing. This is not a gesture toward inclusion; it is how audits capture dimensions of a system's behavior that external researchers cannot observe.

Data collection is also where the adversarial nature of the audit most directly shapes the methodology. When access to the system is denied, the team constructs alternative data sources: analyzing observable outputs, cross-referencing public records, and building the evidentiary record from the outside in. The absence of cooperation from system operators is documented as a finding in itself.

Step 8: Data Analysis

Analysis combines quantitative and qualitative methods, treating them as complementary rather than sequential. The goal is to develop a socio-technical account of how the system actually behaves — not just what it produces statistically, but what those patterns mean in context and for whom.

Qualitative analysis typically draws on three approaches:

20. **Thematic analysis:** reviewing interviews and observational notes to identify recurring patterns in how the system affects people.
21. **Content analysis:** systematically coding data to quantify how often particular issues or concerns appear, identifying which aspects of the context matter most.
22. **Discourse analysis:** examining the language and framing through which people describe their encounters with the system, to surface underlying assumptions and power dynamics.

Quantitative analysis applies statistical methods to observable system outputs. Common techniques include confusion matrices (visualizing where the system correctly or incorrectly classifies people), accuracy metrics, statistical significance testing, difference testing across demographic groups, ROC curve analysis, and endogeneity testing. The specific methods will depend on the system and the available data.

Throughout the quantitative analysis, auditors should assess the system's algorithmic fairness, whether it produces comparable outcomes across different demographic groups. Key fairness metrics include statistical parity (equal rates of positive outcomes across groups), equal opportunity (equal rates of correct positive predictions), and false positive rates (equal rates of incorrect positive predictions). Selecting among these metrics is not a purely technical decision. It depends on context, values, and what kind of harm the audit is most concerned with.

At the end of analysis, the quantitative and qualitative findings must be integrated into a unified account. This account should explain how the algorithm operates in the conditions it has actually been deployed in; identify any biases, inefficiencies, or shortcomings; and describe their consequences for the communities affected. The analysis must be explicit about its own limitations, particularly where the absence of internal access constrains what can be definitively concluded. A third-party community-led audit will rarely be able to fully explain exactly how an algorithm works. It can almost always document what it produces and for whom.

Step 9: Produce the Audit Report and Other Outputs

The audit report is the primary vehicle for the findings, but it is not the only one. The structure Eticas uses across its community-led audits includes:

23. **Executive summary:** a concise overview of the audit's aims, key findings, and recommendations written for a non-technical audience, including policymakers, funders, and journalists.
24. **Context and methodology:** a clear explanation of the socio-technical approach, the community partnership, the specific system being audited, and the methods used to study it.
25. **Findings:** the audit's results, presented as an integrated socio-technical analysis — not a list of statistical outputs, but an account of what those outputs mean in human terms.
26. **Recommendations:** tailored to each of the audiences identified in Step 6, grounded in evidence, and specific enough to be acted upon.
27. **Annexes:** detailed technical methodology notes, datasets (where legally and ethically permissible to publish), and supplementary documentation.

Beyond the report, consider what other forms the findings should take. A dataset made publicly available can enable follow-up research by others. A summary for affected community members — in accessible language, and in their language ensures that the people who made the audit possible can use its results. A press release or media briefing, timed to the report's release, extends the audit's reach.

Step 10: Manage Limitations and Setbacks

Not every audit goes according to plan. Access is denied. Key informants withdraw. A dataset turns out to be less useful than anticipated. A timeline slips. These are not failures; they are predictable features of adversarial auditing in real institutional environments.

The most important response to limitations is transparency: document them in the report, explain what they mean for the conclusions that can be drawn, and distinguish clearly between what the evidence shows and what it suggests. An audit that is honest about what it cannot determine is more credible, not less.

Where a full audit is not achievable, a partial audit with clearly scoped findings is still valuable. A study that documents 80 percent of what was planned and is transparent about the 20 percent it could not reach, is more useful than a study that overclaims. The goal is always findings that are informative, significant, and actionable, even if the path to them was not the one originally planned.

Conclusion: The Case for Community-Led AI Audits

Across our audits, we have seen algorithms that fail to protect those they were designed to safeguard, as in VioGén, where most women at risk never even enter the system and those who do are filtered through processes that are opaque, inconsistent, and often inadequate. We have seen systems like RisCanvi, embedded in the justice system for over a decade, producing outcomes that cannot be explained by their own logic, yet continue to determine people's freedom. We have documented welfare systems that exclude nearly half of eligible beneficiaries through misclassification, facial recognition tools that fail systematically for people with disabilities, and platform algorithms that reproduce and amplify inequality in access to work, mobility, and information.

These are not isolated failures. They are the predictable outcome of a model of AI governance that treats systems as technical artifacts to be optimized, rather than socio-technical infrastructures that redistribute power.

> *Eticas Foundation has been doing exactly that work — at scale, with documented outcomes — for nearly a decade. We have audited systems governments thought were above scrutiny. We have built methodologies that center the people most affected. And we have changed laws..*

This is where community-led auditing changes the equation. By grounding the audit process in the knowledge, experience, and priorities of affected communities, it becomes possible to surface forms of harm that would otherwise remain undetected, to generate evidence that reflects lived realities rather than abstract assumptions, and to produce findings that are both technically robust and socially meaningful. This is not simply a methodological refinement; it is a shift in where authority lies in the evaluation of AI systems, and in what counts as valid evidence in the first place.

The relevance of this approach is not confined to the contexts documented in this guide.

In the United States, systems such as COMPAS, PSA, or those developed by the Arnold Foundation continue to shape decisions around pre-trial detention, sentencing, and supervision, often without meaningful transparency or avenues for contestation. Similar dynamics can be observed in large-scale biometric and welfare infrastructures across Asia, in migration and humanitarian systems across the Middle East and North Africa, and in rapidly expanding platform economies in Latin America. The patterns identified in our audits (opacity, limited accountability, disproportionate impact on marginalized groups, and the erosion of meaningful human oversight) are not exceptions. They are defining features of how algorithmic systems are deployed globally.

At the same time, this work shows that these systems are not beyond scrutiny or change. Community-led audits have already contributed to regulatory reforms, influenced institutional practices, and provided communities with the tools and evidence needed to challenge harmful systems. They demonstrate that meaningful accountability is possible even in conditions of limited access, institutional resistance, and technical complexity.

However, the continued development and scaling of this work cannot rely on isolated efforts or short-term projects. If community-led auditing is to function as a durable component of AI governance, it requires sustained investment, institutional support, and the development of an ecosystem that connects civil society organizations, technical experts, legal practitioners, and affected communities. It requires recognizing that independent scrutiny is not a peripheral activity, but a core public function in an increasingly automated society.

For funders, this represents a clear and urgent opportunity. Supporting community-led audits is not only a way to address immediate harms, but also to build the infrastructure necessary for long-term, systemic accountability. It enables the production of independent, high-quality evidence in contexts where other forms of oversight are weak or absent and strengthens the capacity of communities to engage with technologies that are already shaping their lives. In doing so, it contributes to a model of innovation that is not only technically advanced, but also socially grounded and democratically accountable.

For communities, the message is equally important. The knowledge required to understand and challenge these systems does not reside exclusively within technical or institutional domains. It is already present in the experiences of those who interact with them daily. Community-led auditing provides a framework through which that knowledge can be mobilized, validated, and translated into change.

As the pace of AI deployment continues to outstrip the development of effective regulatory frameworks, both in the United States and globally, and these systems become more deeply embedded in public and private infrastructures, the cost of inaction increases, and the space for meaningful intervention narrows. But with community-led audits, we have built a tested and replicable approach that demonstrates what effective, grounded AI governance can look like in practice.

CLAs are a powerful tool that show that if AI is to serve the public interest, then the public must be equipped to scrutinize it. And if accountability is to be more than a principle, it must be resourced, institutionalized, and led by those most affected by its absence.

The question, therefore, is not whether we can build accountability into AI systems, but who is willing to step forward to fund it, support it, and make community-led AI auditing a key pillar of how we govern technology.

Annex A:
Taxonomy of Algorithmic and AI Systems

The case studies in this guide involve a range of AI and algorithmic system types, from risk-scoring tools and pricing algorithms to recommendation engines and biometric recognition systems. The table below provides a reference overview of the main categories of algorithmic and AI systems, what they do in practice, the contexts in which they are deployed, and their documented or reasonably anticipated negative impacts. Readers may find it useful as orientation before or after reading the case studies.

Type of AI/ Algorithmic System	What do these algorithms do in practice?	Where are these systems used?	Known and potential negative impacts
Search	Retrieve, order, and present information based on inputs and criteria — often from indexes built according to particular interests and capacities, not the live internet.	Internet search engines (Google, Bing, Baidu); search within social media, shopping platforms, and news and entertainment services.	Societal and cultural: information degradation, stereotyping.
Recommendation	Make personalized recommendations of information, products, or services based on user data (including behavioral data), product data, and other parameters.	Content on news media and social media (Facebook, Instagram, TikTok, X/Twitter, YouTube, Spotify, Netflix); goods on Amazon; services on Airbnb, Booking; routes on Google Maps.	Psychological: addiction, alienation/ isolation, anxiety/depression, dehumanization, radicalization, sexualization. Societal/cultural: historical revisionism, information degradation, stereotyping. Political/ economic: electoral interference, institutional trust loss, political manipulation.
Allocation	Automate the distribution of supply and demand in real time and may execute transactions between parties.	Linking customers with service providers on Uber and delivery platforms; automated real-time ad bidding (Google, Meta); algorithmic trading in financial markets.	Financial and business: financial/ earnings loss, monopolization. Human rights and civil liberties: privacy loss.
Monitoring / Surveillance	Observe behavior to collect data and identify predefined or learned patterns and deviations.	User behavior databases for advertising and personalization; fraud detection; employee and student monitoring; automated market monitoring.	Autonomy: agency loss, impersonation, personality rights loss. Psychological: alienation, coercion/manipulation, harassment, dehumanization. Financial: confidentiality loss, earnings loss, livelihood loss. Human rights: dignity loss, discrimination, privacy loss. Societal: labor exploitation.

Type of AI/ Algorithmic System	What do these algorithms do in practice?	Where are these systems used?	Known and potential negative impacts
Facial and Biometric Recognition	Collect and identify patterns to identify an individual, or estimate personal features (gender, age, ethnicity, health status). Observe and track people using biological characteristics.	Identity-based access to devices, apps, payments; airport checkpoints; health condition estimation; identification in public and private spaces (stations, schools, hospitals, open spaces).	Autonomy: personality rights loss. Psychological: harassment/abuse. Reputational: defamation, trust loss. Human rights: dignity loss, discrimination, loss of freedom of assembly, liberty, and due process; privacy loss. Societal: chilling effect, public service deterioration. Political: institutional trust loss.
Pricing	Set or recommend prices for goods and services for particular customers using data on observable characteristics or market conditions.	Dynamic pricing on Uber, Bolt, Cabify, delivery apps; online marketplaces (Amazon); healthcare and insurance; plane tickets and hotel rooms.	Psychological: coercion/manipulation. Financial: business damage, earnings loss, monopolization, opportunity loss. Human rights: discrimination, loss of social rights and access to public services. Societal: breach of ethics/ values, societal inequality.
Aggregation	Collect, categorize, and reorder information or data from different sources.	Content feeds on social media (Facebook, Instagram, TikTok, X/Twitter, YouTube, Spotify, Netflix); internal business and public administration tools (data analytics, risk reporting).	Psychological: addiction, alienation, anxiety/depression, coercion, radicalization, sexualization. Reputational: defamation, trust loss. Human rights: dignity loss, privacy loss. Societal: cultural dispossession, historical revisionism, information degradation, loss of creativity/critical thinking, stereotyping. Political: electoral interference, institutional trust loss, political manipulation.
Communication	Automate communication between customers and businesses or public institutions.	Simple assistants (Siri, Alexa, Google Assistant); first-layer chatbots on digital services and platforms.	Psychological: alienation/isolation. Human rights: privacy loss. Societal: information degradation, loss of creativity/critical thinking.
Filters	Automate selection and filtering of information in the background before presenting it to the user; or allow users to filter content actively on the front end.	Content moderation on social media, media platforms, and marketplaces; content filtering by user data, geography, demographics; flagging of fraudulent operations in banking and social services.	Psychological: addiction, alienation, anxiety/depression, coercion, harassment, radicalization, sexualization. Reputational: trust loss. Human rights: dignity loss, privacy loss. Societal: cultural dispossession, damage to public health, historical revisionism, information degradation, loss of creativity, stereotyping. Political: electoral interference, institutional trust loss, political manipulation.

Type of AI/ Algorithmic System	What do these algorithms do in practice?	Where are these systems used?	Known and potential negative impacts
Information Generation	Generate text, audio, images, or video after being prompted, based on statistical analysis of large training datasets.	Large Language Models (ChatGPT, Gemini, Claude); software code generation; automated text translation (Google Translate, DeepL).	Autonomy: impersonation, IP/ copyright loss, personality rights loss. Psychological: addiction, alienation, anxiety/depression, coercion, radicalization, self-harm, trauma. Reputational: defamation, trust loss. Financial: business damage, confidentiality loss, earnings loss, livelihood loss. Human rights: dignity loss, discrimination, privacy loss. Societal: breach of ethics/values, cheating/plagiarism, cultural dispossession, damage to public health, job loss, labor exploitation, loss of creativity, stereotyping. Political: electoral interference, institutional trust loss.
Prediction	Predict future individual or systemic behavior or scenarios based on statistical calculations.	Fraud prediction in banking, financial, and welfare services; content preference prediction on media platforms; weather and supply chain forecasting; epidemiological forecasting; crime, threat, and security risk prediction.	Autonomy: agency loss, personality rights loss. Physical: bodily injury, loss of life, health deterioration, property damage. Psychological: dehumanization, harassment, over-reliance, trauma. Reputational: defamation, trust loss. Financial: earnings loss, opportunity loss. Human rights: benefits/entitlement loss, dignity loss, discrimination, loss of social rights and public services. Societal: breach of ethics/ values, stereotyping, public service deterioration, societal destabilization, societal inequality, violence/armed conflict. Political: institutional trust loss.

Type of AI/ Algorithmic System	What do these algorithms do in practice?	Where are these systems used?	Known and potential negative impacts
Scoring	Assign scores (relevance, reputation, credit, risk, etc.) and sort information accordingly, regarding people, products, businesses, or organizations.	Credit and financial risk scoring (banking); personal risk scoring in law enforcement and criminal justice (recidivism, aggression risk); insurance risk scoring; automated patient triage; relevance/reputation scoring on marketplaces.	Autonomy: agency loss, personality rights loss. Physical: health deterioration, property damage. Psychological: anxiety/depression, dehumanization, harassment, over-reliance. Reputational: defamation, trust loss. Financial: earnings loss, livelihood loss, opportunity loss. Human rights: benefits/entitlement loss, dignity loss, discrimination, loss of social rights and public services, loss of right to due process, privacy loss. Societal: breach of ethics/values, damage to public health, stereotyping, public service deterioration, societal inequality. Political: institutional trust loss.

Annex B:
CLA Research Techniques

The table below provides a comparative overview of the seven primary data-collection and research techniques used in community-led AI audits. Most audits will combine more than one technique to build a complete picture of how a system operates in practice.

Technique	Brief description	When it is used	Strengths	Limitations
Source Code Analysis	• Reviewing the internal source code, training data, and other relevant technical aspects of the algorithm's codebase to understand how the system works and what it is expected to achieve.	• Only usable when auditors have access to the actual source code, training datasets, and other technical internals — which is rare in adversarial third-party audits.	• High accuracy in auditing a system's functioning. • Rich information about design, intentions, and objectives. • More conclusive findings. • Enables comparison and contrast with other parameters or methods.	• Commercially or security-sensitive parts of the codebase may remain undisclosed even when portions are open. • Analysing significant portions of the codebase requires sophisticated expertise and is time-consuming. • Code analysis may not reveal biases or shortcomings that only emerge in real-life conditions.
Experimental User Testing	• Having real users interact with the algorithm in a predefined, systematic way to observe how it responds to particular inputs and circumstances.	• When the system allows for repeated, differentiated interactions with users — common in web- and app-based systems. • Particularly useful for systems requiring human participation, such as computer vision tools or risk-assessment algorithms.	• Allows affected community members to participate directly in the research. • Systematic observation of outputs helps identify patterns, biases, and shortcomings across conditions. • User-algorithm interactions are closer to real-life conditions than automated inputs.	• Difficult to find enough participants with the specific characteristics needed for significant output sets. • Difficult to execute at large scale — requiring many people to interact systematically and repeatedly.
Sock Puppet Testing	• Creating synthetic user accounts ('sock puppets') with defined characteristics and making them interact systematically with the algorithm. • Interactions can be automated via script or done manually. • Similar to experimental user testing, but with synthetic rather than real accounts.	• When auditing web- or app-based systems where new user profiles can be created easily and the algorithm responds quickly to different user profiles — such as recommendation algorithms, content curation, and online shopping platforms.	• Manual version requires no high-level technical expertise; automated version is accessible to non-specialist auditors. • Enables systematic output observation across conditions to detect biases and shortcomings. • May allow large-scale research in a shorter timeframe.	• Manual creation of many accounts is tedious and time-consuming. • Using sock puppets may violate applicable laws or platform terms of service. • The system may flag sock puppet behavior as suspicious and produce non-representative results. • Automated accounts generate only a limited approximation of how the system adapts to real users over time.

Technique	Brief description	When it is used	Strengths	Limitations
Scraping	• Issuing repeated queries to a digital platform under different conditions and collecting responses systematically. • Can be automated via script or done manually. • Can also be used to access and download existing datasets relevant to the system or its social context.	• When auditing web- or app-based systems where users can submit repeated, differentiated queries — such as search and recommendation algorithms, comparison tools, and gig-economy platforms. • When publicly available datasets contain relevant information about the system's social context.	• Manual version requires limited technical expertise; automated version remains accessible. • Automated scraping can generate large data volumes and support large-scale audits in reasonable timeframes. • Systematic output observation helps identify patterns across conditions. • Scraping existing datasets from the codebase can illuminate a system's inner workings.	• Manual scraping is tedious and time-consuming. • Scraping may be prohibited by applicable law or platform terms of service. • The system may flag repeated automated queries as suspicious and return non-representative results.
Crowdsourcing	• Collecting real-life user data from a number of users about their interactions with a platform or service. • Can be done through 'data donations' — users requesting their personal data under legal provisions (e.g. GDPR, DSA) and sharing it with auditors. • Can also be done in real time via a browser extension or other software installed on users' devices.	• Data donations: when the platform allows users to download their personal data — applicable to most social media, search engines, shopping services, and gig-economy platforms. • Real-time collection: when it is technically feasible to collect usage data via browser extension or similar software.	• Allows affected community members to participate directly in data collection. • Collected data may accurately reflect users' normal interactions with the algorithm.	• Participation samples are almost always self-selected and skewed toward highly motivated, technically literate users — not representative of the whole community. • Developing real-time collection software requires high expertise and ongoing maintenance across browsers and devices.
Ethnographic Research	• Qualitative research through observation, interviews, and surveys to understand the lived experiences of users or affected communities when they interact with the algorithmic system.	• When auditors have access to affected communities who use or are affected by the system. • Where direct access is limited, a civil society organisation with community relationships can facilitate or mediate the research. • Ideally, ethnographic research should be part of every impact-focused algorithmic audit.	• The only approach that directly captures the lived experiences of affected communities. • May allow for direct community involvement, giving members a voice in describing their own experiences. • Helps identify harmful effects and structural factors shaping how the algorithm operates	• Requires particular expertise to be conducted respectfully and to appropriate methodological standards. • Laborious, time-consuming, and difficult to execute at scale. • Experiences vary across user groups; a representative sample of different stakeholders is critical. • On its own, cannot provide definitive answers about why a system is biased or inefficient.

Technique	Brief description	When it is used	Strengths	Limitations
Comparative Output Analysis	• Comparing a system's expected or predicted outcomes with the actual outcomes it generates in real-life conditions. • Can also involve comparing performance against another system, a benchmark, or a statistical accuracy measure	• When reliable data about both the system's expected performance and its actual real-life outputs can be obtained — for example, with risk-assessment algorithms used by public administrations, or with testable systems such as facial recognition software.	• When disclosed data is sufficiently representative, this technique can yield accurate insights and highly informative comparisons that help identify biases, inefficiencies, or shortcomings.	• In practice, public disclosures of data that are representative enough to use this technique are rare. • When data is available, it may be unclear how reliable and representative it is, making it difficult to know whether results would be valid. • Even with representative data, analysis of outputs alone still lacks a complementary account of the system's real-life impacts on people.


eticas
Foundation