

# COMMUNITY-LED AUDITS: A GUIDE

*A comprehensive guide to uncovering and addressing the real- world impacts of AI on communities.*





## Preamble

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 or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

The guide is intended for independent auditors of algorithms, civil-society organisations, and the communities they serve. Its purpose is to provide these groups with practical tools to measure and assess the real-world impacts of AI systems on communities. By uncovering harmful or discriminatory outcomes, the guide enables those most affected to hold developers and deployers to account and to press for fairer, more transparent, and more trustworthy AI.

## About Eticas

Eticas.ai is a global leader in responsible artificial intelligence, working at the intersection of technology, regulation, and social impact. Since 2012, the organisation has been at the forefront of addressing the risks of AI—well before global regulation or mainstream debate caught up.

The company was first to conceive the concept of AI auditing and has built a recognised track record in AI assurance and accountability. Its work equips governments, regulators, corporations, and civil society organisations across Europe, the Americas, and Africa with the expertise and tools needed to ensure AI systems are not only effective but also transparent, compliant, and resilient under scrutiny.

A central figure in shaping the field of AI accountability, founder and CEO Gemma Galdon-Clavell has advised leading international institutions, contributed to key policy debates, and been widely recognised as a pioneer in responsible AI.

In 2018, Eticas expanded its auditing practice to include community-led approaches, working directly with groups most affected by algorithmic decision-making. A pioneering audit with women survivors of domestic violence was published in 2022, followed by the first guide to adversarial auditing. Since then, Eticas has conducted dozens of community-driven audits worldwide.



# CONTENT

Preamble .....	2
About Eticas .....	2
1. Executive Summary .....	5
2. Introduction .....	6
3. Algorithmic auditing as an accountability instrument .....	8
4. Placing communities at the centre .....	10
4.1. The importance of community leadership.....	10
4.2. Participatory approaches to community leadership.....	11
4.3. Eticas's model .....	12
5. How to audit algorithms .....	13
5.1. Steps for conducting community-led audits .....	13
5.1.1. Planning.....	14
5.1.1.1. Choosing a system to audit .....	14
5.1.1.2. Contextual analysis .....	15
5.1.1.3. Stakeholder mapping .....	17
5.1.1.4. Feasibility assessment .....	17
5.1.1.5. Alliance building.....	18
5.1.1.6. Methodology design .....	19
5.1.2. Execution .....	21
5.1.2.1. Data collection .....	21
5.1.2.2. Data analysis.....	22
5.1.2.3. Mitigation and recommendations.....	22
5.1.2.4. Managing limitations and setbacks .....	23
5.2. Methods for conducting adversarial audits .....	24
5.2.1. Open-source code audit .....	24
5.2.2. Scraping .....	26
5.2.3. Sock puppet .....	27
5.2.4. Crowdsourcing.....	28
5.2.5. Experimental user audit .....	29
5.2.6. Comparative output audit .....	30
5.2.7. Ethnographic audit.....	31
5.3. Audit Report Index.....	33
6. Insights from use cases .....	34



7. Conclusion .....	38
8. Acknowledgements.....	39
9. Glossary .....	40
10. Appendix: Case studies.....	43
10.1. Auditing risk assessment algorithms.....	43
10.1.1. VioGén Audit.....	43
10.1.2. RisCanvi Audit.....	45
10.2. Auditing social media.....	47
10.2.1. YouTube Audit .....	47
10.2.2. TikTok Audit.....	50
10.3. Auditing facial recognition.....	52
10.3.1. Use of facial recognition on people with disabilities.....	52
10.4. Auditing consumer platforms .....	54
10.4.1. Audit of ride-hailing platforms .....	54
10.4.2. Ride-hailing platforms for Roma people .....	56
10.5.1. Vape Detector .....	58
11. Bibliography.....	61

# 1. Executive Summary

Artificial intelligence increasingly shapes daily life, from risk assessment in justice systems to ride-hailing platforms, social media, and facial recognition. These systems influence decisions and opportunities in ways that can deepen inequalities or create new ones. The risks are most acute for vulnerable communities, who are rarely consulted in the design and deployment of AI. Without accountability, the spread of AI threatens to entrench opacity, bias, and unfairness.

This guide, based on Eticas's pioneering work since 2018, sets out how community-led audits (CLAs) provide a vital counterweight. CLAs empower those most affected to measure real-world impacts, challenge harmful practices, and demand accountability from the organisations that design and deploy AI systems.

The guide explains what CLAs are, why they matter, who can lead them, and how they can be conducted. It sets out a socio-technical methodology and a toolkit of techniques — such as open-source code audits, scraping, sock puppets, crowdsourcing, comparative output audits, and ethnographic methods. Case studies illustrate how these methods have been applied in practice.

Three key conclusions emerge from this work:

- **Accountability requires community involvement.** Vulnerable groups bear the brunt of AI's risks, yet they are rarely included in oversight. CLAs put communities at the centre.
- **CLAs make hidden harms visible.** By combining lived experience with technical methods, they reveal biases and unfair practices missed by internal audits.
- **CLAs provide a framework for change.** This guide translates principles into actionable steps, equipping communities and civil society with a robust toolset to demand remedies.

By equipping communities with practical tools to investigate AI systems, CLAs ensure that fairness, transparency, and accountability become enforceable standards. The methods are tested, the tools are available — what remains is for more communities to use them, reclaiming agency over the technologies shaping their lives.

This guide offers the key to unlock that power — transforming community knowledge and technical scrutiny into a force for safe, trusted, and accountable AI.

## 2. Introduction

As algorithmic and AI systems<sup>1</sup> proliferate worldwide, we are just starting to learn about their impact. Not enough is known about the ways in which vulnerable and marginalised communities feel the disparate impacts of these systems, nor how communities and oversight authorities might prevent risks and harms that accompany the use of AI tools and systems on the ground. Eticas has years of experience working with civil society organisations (CSOs) and communities in investigating algorithmic and AI injustice, with the aim of increasing transparency and accountability in these systems while empowering affected communities to voice their concerns and take action to demand accountability and redress. We have learned that even seasoned organisers, advocates, and grassroots leaders are often unfamiliar with the specifics of AI technology, which creates a barrier for them to describe, evaluate, and challenge the negative impacts these systems impose in their expert domains.

Eticas recognised this conundrum back in 2018 and began exploring work with communities to reverse-engineer AI systems. A pioneering audit was undertaken with the Ana Bella Foundation to assess VioGén, the gender-violence risk assessment tool used by the Spanish Ministry of the Interior. Our report, published in 2022,<sup>2</sup> found a lack of transparency and accountability, as well as significant concerns with the accuracy of the risk scores outputted by the system. A year later we formalised this method in our Adversarial Audit Guide.<sup>3</sup> This document is an update on that, bringing in the lessons from the many third-party audits that we have subsequently undertaken. Importantly, we are also putting greater emphasis on the role of communities in third-party audits, hence the change of name to community-led audits.

This guide includes actionable guidelines for conducting community-led audits. The guide is addressed to social science researchers, journalists, data scientists, members of civil society organisations, and members of affected groups and end users. It presents a methodology to reverse engineer and evaluate algorithmic and AI systems without the cooperation of their developers, including social media recommender systems, computer vision, risk assessment algorithms, pricing algorithms and others. The goal of this guide is to empower communities to measure the potentially negative impacts of algorithms and AI through a set of rigorous steps and methods.

---

<sup>1</sup> AI system here refers to software which generates outputs for a given set of objectives such as content, predictions, recommendations, or decisions influencing the environments they interact with. The term AI system in this guide refers to the entire technology. For a mobility service, it could be the app that integrates a Machine Learning (ML) model to predict demand and adjust pricing, including, for example, the data pipelines and protocols. In the rest of the paper, algorithmic audits and AI audits are used interchangeably, both referring to audits on algorithmic or AI systems.

<sup>2</sup> Eticas. (2022). The External Audit of the VioGén System. Eticas Research and Innovation.

<sup>3</sup> Eticas, "Adversarial Algorithmic Auditing Guide 2023" (Eticas Research and Innovation, 2023).

In addition to building on Eticas' Adversarial Audit Guide, this document consolidates the knowledge and experience of Eticas in conducting algorithmic audits more generally. Eticas has built a track record as a global leader in practical and applied AI ethics since 2012. We have developed and applied a methodology for conducting internal (second-party) and external (third-party) socio-technical algorithmic audits of risk assessment tools, social media, facial recognition, consumer platforms and other systems.<sup>4</sup> In addition to our own experience, this guide is informed by an extensive review of previous external audits, and it summarises the best practices for external algorithmic auditing.

---

<sup>4</sup> Eticas, "Adversarial Algorithmic Auditing Guide 2023" (Eticas Research and Innovation, 2023); Eticas, "Guide to Algorithmic Auditing" (Eticas Research and Innovation, 2021).

# 3. Algorithmic auditing as an accountability instrument

Algorithmic auditing is an instrument for the appraisal and systematic inspection of AI systems with regard to their performance and external impacts. Auditing as a practice is well-established in aviation, finance, accounting, and information security industries where evidence in the security and functionality of complex systems is much needed; AI researchers and practitioners have built on and adapted the lessons and auditing practices from these industries to the context of algorithmic and AI systems.<sup>5</sup> Conducting algorithmic audits can promote procedural regularity, increase transparency, and inspire proactivity in harm prevention and mitigation during the design of systems.<sup>6</sup>

Algorithmic audits should not be confined to auditing only the algorithm or model,<sup>7</sup> but should encompass entire AI systems.<sup>8</sup> Depending on their scope, algorithmic audits can inspect and evaluate one or more algorithms within an AI system.<sup>9</sup> This guide encompasses a wide range of AI systems, including predictive systems, recommender systems, and biometric systems.

---

<sup>5</sup> Ryan C. LaBrie and G. Steinke, "Towards a Framework for Ethical Audits of AI Algorithms," 2019, <https://www.semanticscholar.org/paper/Towards-a-Framework-for-Ethical-Audits-of-AI-LaBrie-Steinke/c103601dbf79c05c7f72b865ce05e6f82048c1ca>; Miles Brundage et al., "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims" (arXiv, April 20, 2020), <https://doi.org/10.48550/arXiv.2004.07213>; Jakob Mökander et al., "Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations," *Science and Engineering Ethics* 27, no. 4 (July 6, 2021): 44, <https://doi.org/10.1007/s11948-021-00319-4>; Jakob Mökander and Luciano Floridi, "Ethics-Based Auditing to Develop Trustworthy AI," *Minds and Machines* 31, no. 2 (June 1, 2021): 323–27, <https://doi.org/10.1007/s11023-021-09557-8>; Adriano Koshiyama et al., "Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms," SSRN Scholarly Paper (Rochester, NY, January 1, 2021), <https://doi.org/10.2139/ssrn.3778998>.

<sup>6</sup> Shahar Avin et al., "Filling Gaps in Trustworthy Development of AI," *Science* 374, no. 6573 (December 10, 2021): 1327–29, <https://doi.org/10.1126/science.abi7176>; Jakob Mökander and Luciano Floridi, "Operationalising AI Governance through Ethics-Based Auditing: An Industry Case Study," *AI and Ethics* 3, no. 2 (May 1, 2023): 451–68, <https://doi.org/10.1007/s43681-022-00171-7>; Jakob Mökander et al., "Auditing Large Language Models: A Three-Layered Approach," SSRN Scholarly Paper (Rochester, NY, February 16, 2023), <https://doi.org/10.2139/ssrn.4361607>.

<sup>7</sup> An algorithm here refers to a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer. An AI model refers to the trained algorithm where the process or the rules are adapted to a particular domain. In this guide, an algorithm is used interchangeably with AI model.

<sup>8</sup> AI system here refers to software which generates outputs for a given set of objectives such as content, predictions, recommendations, or decisions influencing the environments they interact with. The term AI system in this guide refers to the entire technology. For a mobility service, it could be the app that integrates a Machine Learning (ML) model to predict demand and adjust pricing, including, for example, the data pipelines and protocols.

<sup>9</sup> For the sake of simplicity, this guide refers to "AI system" in the latter sections, but depending on each case this may apply to the entire system or a specific algorithm within that system.

The importance of focusing on AI systems as a whole is that it enables the audit to address the overall impact of an algorithm, on the users of the AI system and those communities that are more broadly affected by it. True accountability to a community can only be delivered by considering the overall impact of a model. The failure to do so is a failing of many current approaches to algorithmic auditing.

Algorithmic audits can be broadly classified into three types depending on the auditors' distance from the developers or implementers of an algorithm. A first-party audit is conducted by the developer of an AI system. A second-party audit is also conducted internally, but by an independent auditor in collaboration with the developers of an AI system. It is an iterative process of interaction between the auditor(s) and the development team(s) who provide the data inputs necessary for auditors to complete the assessment and validate results.

A third-party audit, also known as an external, adversarial, or community-led algorithmic audit, is a process by which an independent external party or community examines the impact and, to the extent possible, the functioning of an algorithmic system. The auditor's independent position helps remove misaligned incentives in developer self-reporting, establish accountability, and increase overall transparency and public trust in the inspected system. External audits further hold the promise of detecting biases, inefficiencies, anomalies, and other hidden practices that could be unfair or harmful towards vulnerable communities within the society.<sup>10</sup>

The key distinguishing feature of a third-party audit is the restricted access to the algorithm and its associated databases used for design, development, testing and validation. For this reason, such audits can be conducted only when the algorithm's social impacts can be observed, i.e. in the post-deployment or the post-processing stage of the AI system lifecycle, unlike first or second-party audits that can encompass the entire end-to-end process.

---

<sup>10</sup> Eticas, "Adversarial Algorithmic Auditing Guide 2023" (Eticas Research and Innovation, 2023).

## 4. Placing communities at the centre

Due to the limited access to internal data and information about an algorithm, third-party audits do not aim to provide a comprehensive, conclusive assessment of the entire system at hand. Rather, they help to identify instances of bias and inefficiencies in algorithms and AI, prompt developers to address them, and inform regulators and the public to ask the right questions. One of the canonical third-party audits is ProPublica's audit of the COMPAS recidivism prediction algorithm, which found that black defendants were significantly more likely to be assigned a higher risk score than white defendants even after controlling for factors such as prior crimes, age, and gender.<sup>11</sup>

Third-party audits are particularly powerful when the audit is led by the affected community, and this chapter emphasises why this is important, and briefly explains the participatory approach that is at the core of meaningful community leadership.

### 4.1. The importance of community leadership

Algorithmic audit studies have gradually acknowledged the value of involving those most affected by an AI technology in one way or another to complement an auditor's analysis. In the case of first and second-party audits, development teams often do not have the mandate, incentives, or tools to properly engage with affected communities, even if the developers personally may wish too. For second-party audits, these challenges also limit the scope of work of the external auditors, often against the recommendations of the auditor.

Even in the case of third-party audits, independent experts who are not steeped in the context of a community will struggle to fully capture the richness of community members' lived experiences and how these should inform the audits hypotheses. Therefore, only by fully involving communities can a third-party audit effectively assess the impact of an AI system on that community.

Another advantage of fully involving a community in a third-party audit is that it supports the community to build its own capacity to directly engage with systems that impact them. This helps to create the 'muscle' such that community is better able to challenge such systems in the future.

---

<sup>11</sup> Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, "Machine Bias: Risk Assessments in Criminal Sentencing" (ProPublica, May 23, 2016).

## 4.2. Participatory approaches to community leadership

In an attempt to include community perspectives, some researchers proposed end-user audits or crowdsourced audits, such as tech companies' bounty challenges.<sup>12</sup> However, these approaches often transfer the onus of auditing algorithmic and AI systems onto individual end users, who may not hold the technical knowledge to investigate and build a case around their experiences. They can also be inaccessible to vulnerable communities who have been historically excluded from digital access that would have enabled them to participate meaningfully in critical response.

Participatory approaches address this gap by bringing together auditors and community members as collaborators while mutually learning about each other's knowledge and perspectives.<sup>13</sup> It draws on the social audits methods implemented in social science research, especially those driven by explicit concerns of social justice and racial equity, which have a strong convention of requiring the direct participation of the impacted communities and are oriented around establishing accountability.<sup>14</sup> These participatory audits stress the importance of conducting research and analysis *with* participants, not *on* or *for* them, and serving their needs and goals.<sup>15</sup>

Community participation requires time and respect on the part of auditors. It takes time to earn the trust of the community, and to demonstrate that their views and experiences will be genuinely represented in the work and not twisted to fit someone else's narrative. By respect we don't just mean respecting the views of community members, although this is clearly vital, but also respecting that their time is valuable and to compensate them for it, not simply assume that people will participate because they are experiencing adverse impacts.

---

<sup>12</sup> Lam et al.; Wesley Hanwen Deng et al., "Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23 (New York, NY, USA: Association for Computing Machinery, 2023), 1–18, <https://doi.org/10.1145/3544548.3581026>.

<sup>13</sup> William F. Whyte, "Advancing Scientific Knowledge through Participatory Action Research," *Sociological Forum* 4, no. 3 (September 1, 1989): 367–85, <https://doi.org/10.1007/BF01115015>.

<sup>14</sup> Diana Auret and Stephanie Barrientos, "Participatory Social Auditing: A Practical Guide to Developing a Gender-Sensitive Approach," IDS Working Papers 237 (Brighton, the UK: Institute of Development Studies (IDS), 2004), [https://opendocs.ids.ac.uk/articles/report/Participatory\\_social\\_auditing\\_a\\_practical\\_guide\\_to\\_developing\\_a\\_gender-sensitive\\_approach/26480614?file=48231346](https://opendocs.ids.ac.uk/articles/report/Participatory_social_auditing_a_practical_guide_to_developing_a_gender-sensitive_approach/26480614?file=48231346); Diana Auret and Stephanie Barrientos, "Participatory Social Auditing: Developing a Worker-Focused Approach," in *Ethical Sourcing in the Global Food System* (Routledge, 2006); Briana Vecchione, Karen Levy, and Solon Barocas, "Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies," in *Equity and Access in Algorithms, Mechanisms, and Optimization* (EAAMO '21: Equity and Access in Algorithms, Mechanisms, and Optimization, -- NY USA: ACM, 2021), 1–9, <https://doi.org/10.1145/3465416.3483294>.

<sup>15</sup> Vecchione, Levy, and Barocas, "Algorithmic Auditing and Social Justice"; Alice McIntyre, *Participatory Action Research* (SAGE Publications, Inc., 2008), <https://doi.org/10.4135/9781483385679>.

### 4.3. Eticas's model

Eticas began introducing participatory approaches because it felt that traditional models of community engagement, such as technology training, were insufficient. This was particularly so in the face of AI eroding the rights of some communities. Community-led audits were developed as a tool that combines hands-on training with awareness and empowerment. As an example of the latter, Eticas often hires community members and directly integrates them into an audit team.

This guide to CLA balances the technical expertise of auditors and the lived, contextual, and strategic expertise of community members, allowing both to co-facilitate different steps of the way, from setting research agenda to providing on-the-ground data, reports, and metrics. The outcome of a CLA is guided by the needs of the communities to support them to take action against unfair or unjust algorithmic and AI systems. Moreover, the process of the CLA also builds the capacity of an affected community to better engage with the developers of AI systems.

Community-led audits allow communities to have a seat at the table, rather than being passive recipients of the impact of AI systems. The case studies demonstrate this across a wide range of issues, and also show that the CLA is not an abstract process but one which has real world impact. For instance, the results of the CLA on ride-sharing was submitted as evidence to Spain's competition authority. The intention of this guide is that it helps you to achieve the real-world impact that is important to you and the community you are part of, or working alongside.

# 5. How to audit algorithms

## 5.1. Steps for conducting community-led audits

This section describes the steps for conducting CLAs. Based on our experience in the field, we have found that the sequence of steps outlined below is commonly followed and convenient. However, it is important to note that this order is not always 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. In the following subsections, we divide the adversarial audit process into two main phases: planning and execution. We provide a description of how each step can be applied in practice.

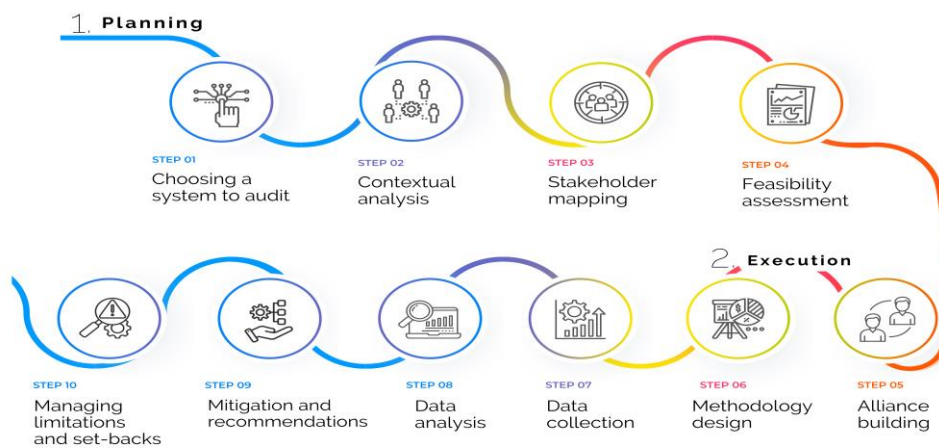


Figure 1: Steps for conducting Community-Led Audits

Planning	
<b>Choosing a system to audit</b>	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
<b>Contextual analysis</b>	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
<b>Stakeholder mapping</b>	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
<b>Feasibility assessment</b>	Data mapping to determine if the auditor can obtain sufficient information about an AI system via legal means within the relevant jurisdiction
<b>Alliance building</b>	Participatory research design with communities and civil society organizations to ensure that the perspectives of affected groups are incorporated in the auditing process
<b>Methodology design</b>	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
<b>Execution</b>	
<b>Data collection</b>	Safe and consentful <sup>16</sup> 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
<b>Data analysis</b>	Translating raw data into meaningful insights via quantitative and qualitative data analysis
<b>Mitigation and recommendations</b>	Providing actionable audit outputs, including reports, metrics, visualisations, and recommendations that serve community leadership in demanding accountability and improvement from developers, implementers, and policymakers
<b>Managing limitations and set-backs</b>	Strategies for how to cope when a CLA doesn't go to plan, or simply has to work within a more limited scope than is desirable given the community's goals.

### 5.1.1. Planning

The planning phase involves a series of steps aimed at ensuring that the audit has a clear goal and that it is well-prepared and organised. This involves the following steps: choosing a system to audit, contextual analysis, stakeholder mapping, feasibility assessment, alliance building, and methodology design.

#### 5.1.1.1. Choosing a system to audit

**Overview of step 1:** 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.

<sup>16</sup> The term "consentful" is inspired by a Design Justice practice. See more: <https://alliedmedia.org/projects/consentful-tech-project>

**Key questions it addresses:** What AI system should be audited? What technology does it utilise and in what field? What do we know about its impact? What are possible ways to access the AI system for an audit?

**Core considerations:** Either communities or auditors may identify an AI system of concern. Regardless of the initiator, being deciding whether to proceed with an audit, the following considerations are crucial:

- **Is (a part of) the AI system accessible to the auditor?** For example, a system in a web-based platform such as social media recommender systems.
- **Is there any open-source or public record information about the algorithm?** For instance, in our audit of the VioGén gender-based violence risk assessment tool, we were unable to obtain the original database for intimate partner violence in Spain, but we identified a public record of homicide victims, including victims of intimate partner violence (a subset of the database we sought to obtain).
- **Does the auditor have access to affected communities?** In our audit of the VioGén system, we worked with the civil society organization the Ana Bella Foundation to access women victims of gender-based violence.
- **Can the auditor directly or indirectly observe the inputs or the outputs of an AI system?** For example, in our external audits of social media platforms YouTube and TikTok, we were able to observe the outputs (suggested content) of YouTube's and TikTok's recommender systems.

While this checklist is non-exhaustive, it serves as an initial feasibility check for conducting an adversarial audit. If the answer to one or more of the questions above is positive, the auditor can proceed on to the next steps. If there are no possibilities to access any part of the AI system directly or indirectly through user and stakeholder experiences, it is recommended to consider alternative ways to obtain information such as requests for information access to public authorities or contact with other organizations who can facilitate access to affected communities.

A more comprehensive feasibility assessment is suggested in Step 4.

**Further resources:** Previous academic research, the work of civil society organisations and journalists, user feedback, public data or the experiences of affected communities are good starting points in choosing an algorithm for an adversarial audit. There are also specialised resources and tools which can help auditors identify AI systems of interest, such as the Observatory of Algorithms with Social Impact ([OASI](#)), which gathers and classifies information about algorithms searchable and is regularly updated with new content.

#### 5.1.1.2. Contextual analysis

**Overview of step 2:** 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.

**Key questions it addresses:** How does the AI system work? What is it trying to do? Where and in what context does it operate? What are the biases and inefficiencies expected to occur? Could there be other unexpected biases, inefficiencies or other anomalies?

**Core considerations:** Contextual analysis involves an extensive literature review and interviews with technical and subject matter experts in the domain in which the AI system operates. The goal of this step is to form initial hypotheses for the presence of algorithmic harm or inefficiencies in a given AI system within its broader social, legal and economic context.

An important step in hypothesis generation is determining what biases to check for in a system. Below is a non-exhaustive list of possible biases.

<b>Techno-solutionist bias</b>	Failure to consider no-tech or low-tech options, to perform a proportionality assessment or to consider social and environmental issues before deciding to develop or implement an algorithmic system.
<b>Population bias</b>	Population bias arises from differences between the actual usage population and the design target population of a system. This means that the target population defined during the design and development phases is not representative of the population that will use the system after it is deployed. Population bias results in non-representative data and results that fit only the most salient groups while harming all minority groups.
<b>Omitted variable</b>	When one or more important variables are not included in the model, resulting in biased regression coefficients and inaccurate statistical results.
<b>Historical bias</b>	Existing bias in the world percolates into the data used for training, validation, and testing. Even if data is accurate and well measured and sampled, the world "as it is" may lead to a model that produces harmful outcomes. Historical bias stems from societal inequalities, cultural differences, stereotyping, etc.
<b>Aggregation bias</b>	When a given model is not optimal for any group or is skewed towards the dominant population. This type of bias is also known as ecological fallacy, for it occurs when incorrect or false conclusions are drawn about individuals by observing the population.
<b>Accessibility bias</b>	This bias occurs when the AI system or parts of it are the best fit for the greatest average of the majority, but leave out marginalised groups, in particular people with disabilities. For this reason, accessibility bias affects a smaller portion of the population.

An important aspect of adversarial algorithmic auditing is checking for previously untested or otherwise unexpected biases or inefficiencies. For example, in our audit of facial recognition (FR) in the insurance sector, we found that FR has been shown to be biased against women and people of colour, but its performance had not been tested on individuals with disabilities with physical manifestations.

**Further resources:** For a more comprehensive list of sources and moments of bias in AI systems, see Eticas' Guide to Algorithmic Auditing.

#### 5.1.1.3. Stakeholder mapping

**Overview of step 3:** 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.

**Key questions it addresses:** Who developed and implements the algorithm? Do they have previous experience in using automated solutions? Who is promoting this system? Which communities are impacted directly or indirectly? Which groups are at risk?

**Core considerations:** Stakeholder mapping involves both identifying a comprehensive list of relevant parties and how they are positioned with respect to the operation of the AI system.

- Relevant parties include, but are not limited to: the developers and implementers of the system, operators, pertinent public authorities and regulators, target population, users and communities affected directly or indirectly.
- The mapping should seek to identify where groups overlap. For example, the developers, implementers and operators of an algorithm may comprise a single stakeholder group (e.g., a company) or three distinct groups depending on the specific system at hand.
- In terms of stakeholder positioning, it is important to understand (to the extent possible) the objectives and motivations for – and previous experience in – creating or utilizing automated solutions, referring back to the techno-solutionist bias among others.
- Stakeholder mapping is also a useful tool to identify which groups may be at risk of bias, discrimination or other harm.

#### 5.1.1.4. Feasibility assessment

**Overview of step 4:** Data mapping to determine if the auditor can obtain sufficient information about an AI system via legal means within the relevant jurisdiction.

**Key questions it addresses:** Is there existing literature on the topic? Where can we get data from? Can we access affected communities? Is the audit legally feasible?

**Core considerations:** Feasibility assessments determine whether an auditor can obtain sufficient information about an algorithm via legal means. Informed by the knowledge about the algorithm, the environment in which it operates, and the relevant stakeholders acquired in the previous steps, feasibility assessments entail two major components:

- Data mapping: identifying relevant literature on the topic, specific access points to the AI system and the means to contact affected communities, and evaluating whether those sources can provide sufficient information about the functioning or the impacts of an AI system.
- Legal feasibility assessment: examining applicable legislation in the relevant jurisdiction of the auditor(s), as well as the terms of service of the platform which implements an algorithm in the case of auditing web- and app-based systems.

The next step, Alliance building, can help to address challenges identified during the feasibility assessment, such as difficulties with access to affected populations. However, if no access points have been successfully identified or if those access points do not comply with applicable laws, a community-led algorithmic audit of the given AI system may not be feasible at this time.

#### 5.1.1.5. Alliance building

**Overview of step 5:** Participatory research design with communities and civil society organizations to ensure that the perspectives of affected groups are incorporated in the auditing process.

**Key questions it addresses:** How to collaborate with affected communities? Which are the civil society organisations working with them? How can we collaborate and partner with them? What do we expect from them and what can we offer in exchange?

**Core considerations:** Alliance building with communities or civil society organisations representing them enables communication and facilitates trust between the auditors and those at risk of algorithmic harm. Critical aspects of this include:

- Mapping and reaching out to members of affected communities and relevant civil society organisations in the field. In cases where communicating with groups at risk is difficult or concerns sensitive issues, civil society organizations can facilitate contacts and conduct ethnographic research on behalf of the auditors. For example, in our adversarial audit of the VioGén system, we partnered with the Ana Bella Foundation who interviewed victims of gender-based violence.
- Early establishment of the obligations of each partner within the collaboration. It is recommended to set this out in a contract and that proper attribution is given to each depending on their role. Whenever possible, civil society organisations should

be compensated for their time once the terms of the audit have been agreed to by all parties.

#### 5.1.1.6. Methodology design

**Overview of step 6:** 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.

**Key questions it addresses:** How to look for bias and inefficiencies in an algorithmic system? What is the most appropriate method to use? How to approach it in a systematic way? With limited access to internal data, how can we gather information about an algorithmic system?

**Core considerations:** The research questions for the audit will reflect its scope, as identified in the previous steps, but crystallised to set out key parameters such as: the AI system(s); the relevant affected group(s); the geographical location; and the time period. When formulated into a verifiable proposition about the presence of bias, harm or inefficiencies within an AI system, the research question then guides the choice of auditing methods.

Community-led algorithmic auditing combines traditional social science methods from a socio-technical perspective and specialised methods for algorithmic auditing. The table below provides an overview of the main methods.

Method	Description	Community role
<b>Experimental user</b>	A systematic method for observing and recording system responses to real user behaviours under different conditions, including real-life conditions.	Community members should be invited to co-design experiments based on their lived experiences and replicate them in experimental conditions for further analysis.
<b>Crowdsourcing</b>	Collecting users' regular interactions with a system.	Community members offer their lived experiences in the form of data collected by algorithmic or AI systems to be the center of analysis.
<b>Ethnography</b>	A qualitative method for data collection through observation, interviews and surveys to understand and analyze how	Community members are participants alongside ethnographers. The goal is to make visible community

	end users, particularly vulnerable groups, interact with a system.	members' real-life practices and perspectives.
<b>Comparative output</b>	Compares an algorithm's predicted outcomes with the actual outcomes, or the performance of one system against another, a benchmark, or a statistical measure for accuracy.	Communities and CSOs may be potential data sources, especially if they have been recording data on a long timescale.
<b>Sock puppet</b>	A systematic method for simulating real user behaviour which involves the use of impersonation through (sock puppet accounts) and recording the system's response to different user characteristics and behaviour(s).	Community members act as advisors and observers. They may also co-design sock puppet.
<b>Scraping</b>	A systematic method of issuing repeated queries to a platform under different conditions and collecting the results.	Community members act as advisors and observers.
<b>Open-source code audit</b>	If white-box access to the system is possible, the auditor can perform a thorough review of the system's source code, training data, and other inputs to understand the algorithm's intentions and objectives.	Community members act as advisors and observers.

(Table 1. Brief description of each audit method.)

The figure below serves as guidance on the selection of the most appropriate auditing method depending on the availability of data. Importantly, it is often appropriate to use multiple methods, depending on the scope of the audit, as demonstrated in the case studies (see later section).

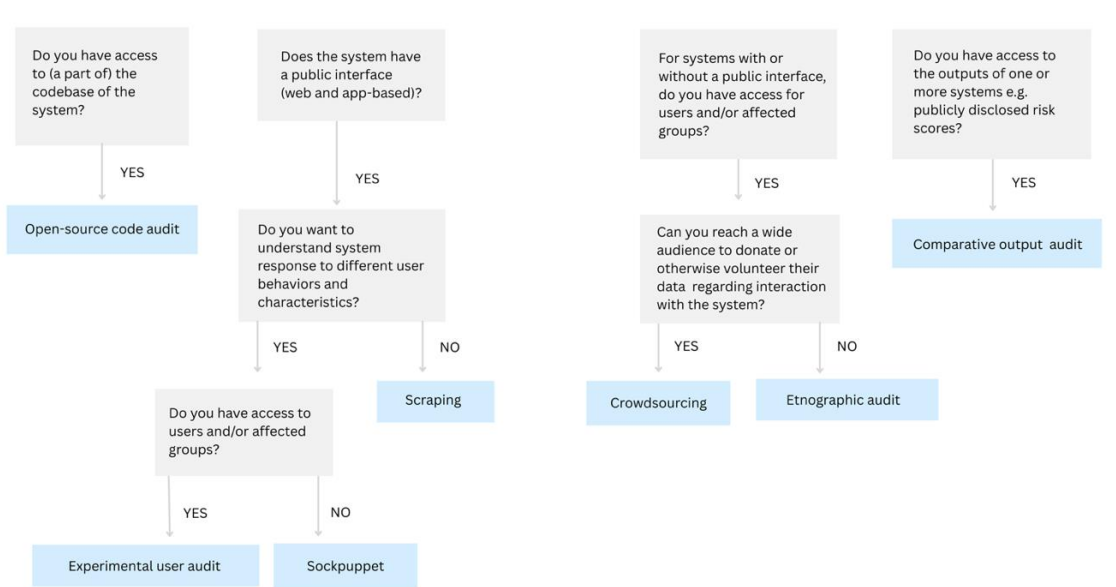


Figure 2. Selection of method(s) for conducting community-led audits

**Further resources:** More details on each of the above methods can be found within the appendix

## 5.1.2. Execution

The execution phase involves carrying out the audit according to the previously designed methodology, starting with data collection, analysing and interpreting results, presenting findings and providing recommendations or mitigation measures.

### 5.1.2.1. Data collection

**Overview of step 7:** Safe and consentful<sup>17</sup> qualitative and quantitative data gathering about the inputs, outputs and societal impact of an AI system via specialized techniques for adversarial.

**Key questions it addresses:** Have we gathered sufficient data? Is our data sufficiently representative? What insights can the collected data provide? How are those insights limited?

**Core considerations:** The goal of the data collection step is to gather raw information that enables auditors to address the research questions identified in the previous step. Depending on the chosen methodology, this step can include qualitative fieldwork such as surveys and interviews (ethnographic audit), manual or automated quantitative data collection (sock puppet and scraping audits), conducting tests with users (experimental user audit), or organizing data donation campaigns for users (crowdsourcing audit).

<sup>17</sup> The term "consentful" is inspired by a Design Justice practice. See more: <https://alliedmedia.org/projects/consentful-tech-project>

It is critical for the auditor to recognize and acknowledge the limitations of the data collection process, as these limitations can impact the applicability of the findings to different contexts. This involves addressing questions such as:

- Does the audit focus on a specific geographic area or time period?
- To what extent does the data reflect the experiences of all stakeholder groups?
- In cases where automated techniques are used for data collection, how accurately does the data represent the experiences of real users?

For handling qualitative and quantitative data from users or affected groups, participants in the study should sign an informed consent form. The consent form should outline data management principles, including anonymisation where possible and secure storage. Additionally, it should communicate, risks (if any) and conditions of participation.

#### 5.1.2.2. Data analysis

**Overview of step 8:** Translating raw data into meaningful insights via quantitative and qualitative data analysis.

**Key questions it addresses:** Have we observed the biases we initially suspected? Have we identified any additional instances of bias that were not identified in the previous steps? If we did not detect any anomalies or bias – how can we refine our methodology?

**Core considerations:** The goal of the data analysis is to translate raw data into meaningful insights that address the research questions formulated during the audit planning stage and identify bias or inefficiencies in an AI system. It can entail both quantitative and qualitative analysis.

- The methods used for quantitative analysis may vary depending on the collected data and research questions. They can include techniques such as confusion matrix, accuracy metrics, statistical significance testing, difference testing, ROC curve analysis, and endogeneity testing.
- Qualitative analysis methods may involve thematic analysis, content analysis, discourse analysis, and others.

To ensure the robustness of the findings, it is important to include validation of the results whenever possible.

#### 5.1.2.3. Mitigation and recommendations

**Overview of step 9:** Providing actionable audit outputs, including reports, metrics, visualisations, and recommendations that serve community leadership in demanding accountability and improvement from developers, implementers, and policymakers.

**Key questions it addresses:** What are the wider implications of the biases and inefficiencies we have identified? What can be done to address them? What can developers do to mitigate bias?

**Core considerations:** Once the analysis has been completed, the auditor should consider the social, legal and economic implications of the findings, and ways to address the biases, inefficiencies and other negative impacts.

The auditor should provide concrete and actionable mitigation measures for the developers or the implementers of the AI system. From a policy standpoint, the auditor should provide recommendations that go beyond existing regulations and empower the regulators with the knowledge of pertinent questions to ask. Examples of such mitigation measures and policy recommendations are described in the case studies (see later section).

#### 5.1.2.4. Managing limitations and setbacks

**Overview of step 10:** Strategies for how to cope when a CLA doesn't go to plan, or simply has to work within a more limited scope than is desirable given the community's goals.

**Key questions it addresses:** What to do when data access is limited? How to cope when circumstances mean that the original scope is not feasible?

**Core considerations:** The feasibility assessment set out in step 4 won't always lead to a clear answer. In particular it can be difficult to judge whether it is worth proceeding. Two scenarios are considered here:

- First, it was clear that data was limited from the outset. Whilst a full audit may not be possible, it may still be worth undertaking more of a scoping audit, in which the limited data available is used to highlight the challenges that the community is facing, and to help plan what a full audit would look like. In this circumstance, even a more limited audit can have value as part of a wider campaign to support that community and may even help to secure access to more data in the future that will allow a fuller audit.
- Second, it may be that an audit process has started but some data that was initially available may no longer be. The approach here is similar to the first scenario, with the addition of including a description of how data access was reduced, including capturing the views of different stakeholders on why it was reduced. Transparency about such a change can itself be a motivator to strengthen a campaign for change.

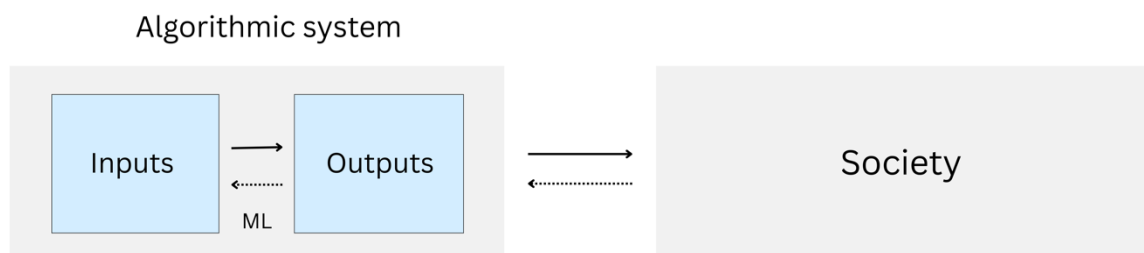
## 5.2. Methods for conducting adversarial audits

This section explores different methods of conducting adversarial audits to assess the impact of algorithmic systems. Previous guides to adversarial algorithmic auditing have focused on web- and app-based systems and as a result, they examine the interaction between platforms and users (Sandvig et al.).



*Figure 3. Visualization of the interaction between platforms and users*  
*Source: Sandvig et al.*

This guide presents a methodology for auditing various types of AI systems including but not limited to social media recommender systems, computer vision, risk assessment tools, chatbots and consumer platforms regarding their impact on affected communities and society. To accomplish this, we conceptualize the interaction between an AI system and society. In the graph below, the arrows represent the flow of information or the direction of the interaction.



*Figure 4. Visualization of the interaction between an algorithmic system and society*  
*Source: Eticas*

In the following methods, we illustrate the direction of the interaction or data exchange among the AI system, society and the auditor. It is ideal for audit methods that rely on observations of the algorithmic system to be accompanied by methods that examine the impact on society, and vice versa, to enable a comprehensive assessment of the system's functioning within its context. When such a combination is not feasible, at a minimum, audit methods should be complemented by literature reviews and interviews with domain experts to bridge the gap between the two.

### 5.2.1. Open-source code audit

The open-source code audit entails a review of an algorithmic system's source code, training data, and other inputs to understand the algorithm's intentions and objectives.

Additionally, if feasible, statistical measures can be employed to assess bias and fairness. By gaining access to (a portion of) the source code, independent auditors can approximate the internal socio-technical audit process. For a comprehensive guide on conducting codebase reviews, please refer to Eticas' Guide to Algorithmic Auditing.<sup>18</sup>

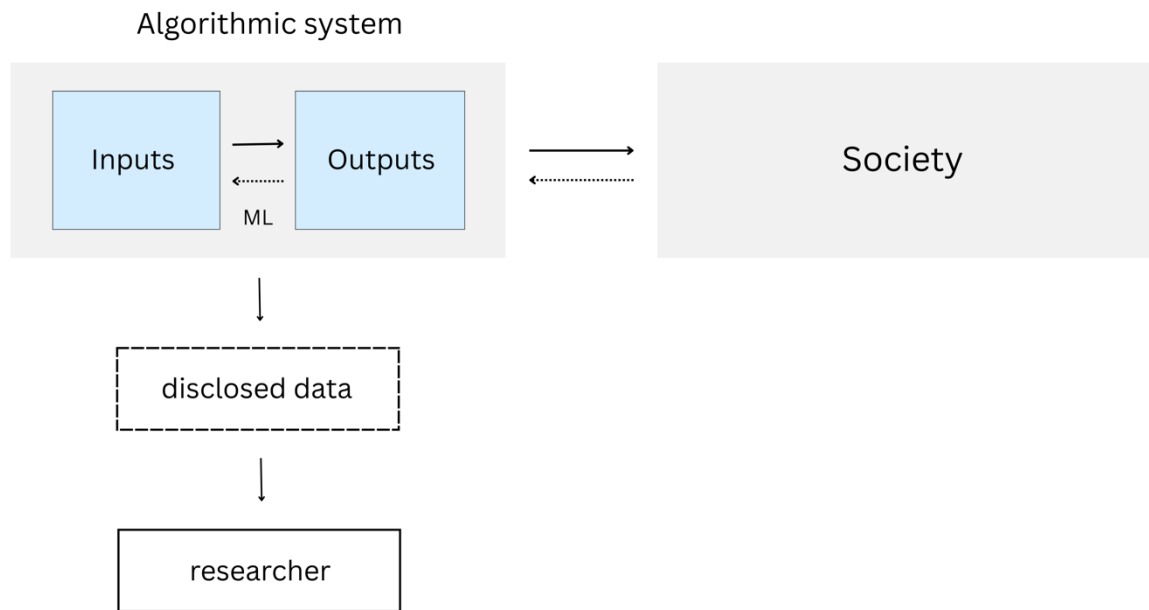


Figure 5. Visualization of the interaction between an algorithmic system and society  
Source: Eticas

When to use this method:

- When the source code of an algorithm is open-source or otherwise publicly available.
- When companies are required to disclose data, e.g., as part of legal proceedings.

Strengths:

- High level of accuracy in auditing the functioning of a system.
- Rich information about a system's design, intentions and objectives.
- Examining the codebases of an algorithm offers more conclusive findings.
- Possibility to compare, adjust and contrast with other hyperparameters, parameters or methods.

Limitations:

- Problematic machine behaviours may not be encoded within the system, and bias dynamics may only become evident when they manifest as impacts. Since open-source code audits do not examine the impact of an algorithmic system, conclusions solely based on this method have limitations in assessing harm and inefficiencies.

<sup>18</sup> Eticas, "Guide to Algorithmic Auditing" (Eticas Research and Innovation, 2021).

- A comprehensive algorithmic open-source code audit requires high-level access to all codebases and training data which can be challenging to access and time-consuming to review. This is especially the case for complex algorithmic systems comprising multiple algorithms such as social media platforms.
- Open-source code audits are difficult to perform due to a general lack of transparency in disclosing codebases: most algorithms remain inaccessible due to concerns about intellectual property, while open-source codebases may not disclose all relevant information for security reasons.

## 5.2.2. Scraping

A systematic method of issuing repeated queries to a platform under different conditions and collecting the results. Scraping can be done manually by the auditor, or automatically by using a custom script.

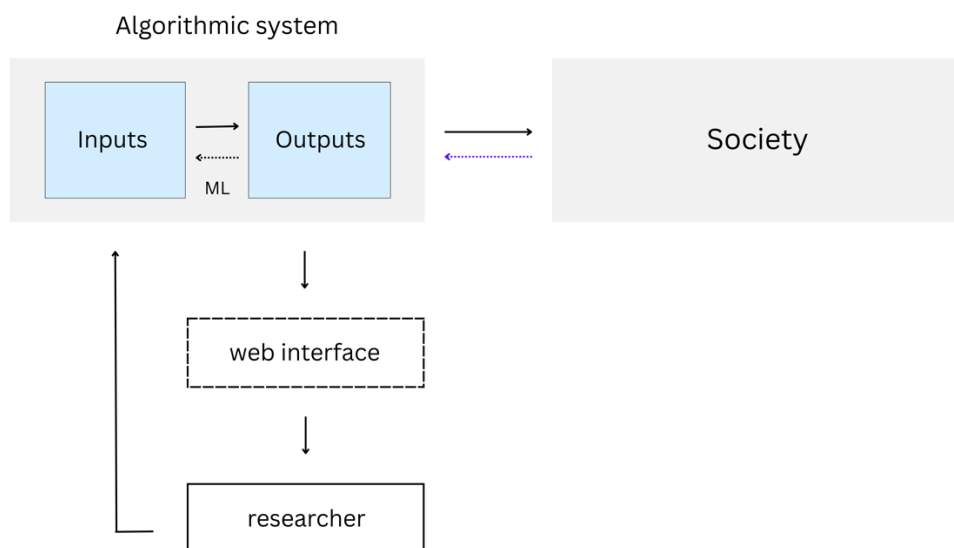


Figure 6. Visualization of the interaction between an algorithmic system and society  
Source: *Eticas*

When to use this method:

- When auditing web- and app- based systems which allow users to 'play' with the system including social media, search engines, e-commerce websites, online comparison tools, apps in the sharing economy.
- Suitable for large-scale audits.

Strengths:

- Effective method to observe the outputs of a system and identify patterns.
- Accessible method to all auditors and communities regardless of level of technical expertise (for manual scraping) and resources.
- If automated, scraping can generate a high amount of data for testing and analysis.

Limitations:

- Depending on the jurisdiction and the terms of service of the platform, automated scraping may be illegal. If there are concerns about the legal feasibility of this method, auditors should seek legal counsel and ensure adequate safeguards are in place.
- The system under investigation may flag suspicious behaviour when using automated scraping via bots or scripts, producing results that are not representative real users' experience.
- Manual scraping can be time-consuming and laborious.

### 5.2.3. Sock puppet

A systematic method for simulating real user behaviour which involves the use of impersonation through (sock puppet accounts) and recording the system's response to different user characteristics and behaviour(s). The sock puppet method can be executed manually by the researcher, or automatically by using a custom script.

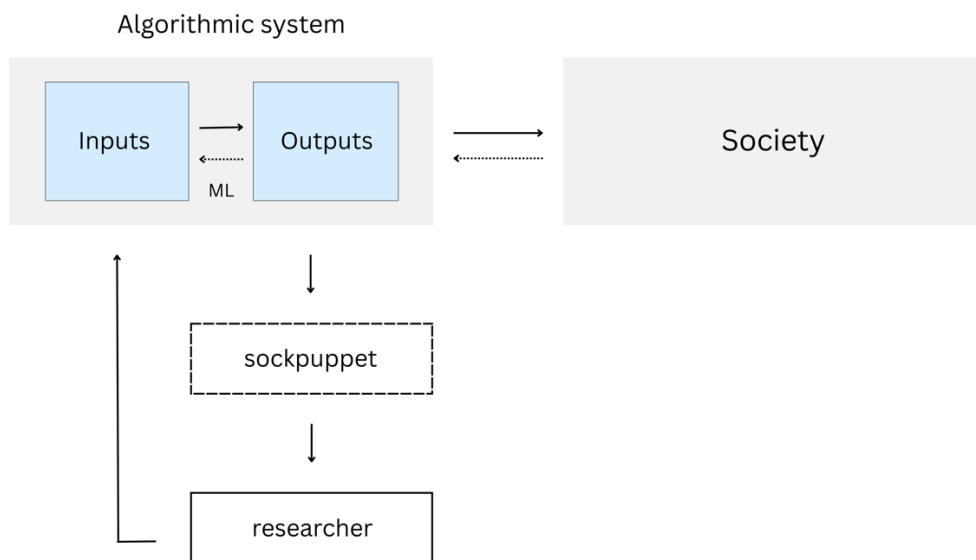


Figure 7. Visualization of the interaction between an algorithmic system and society

Source: Eticas

When to use this method:

1. When auditing web- and app-based systems where users can create profiles and 'play' with the system, particularly systems which employ personalization such as social media recommender systems, news curation services or e-commerce websites.
2. Large-scale audits.

Strengths:

- Effective method to observe the outputs of a system and identify patterns across different conditions, enabling comparison and more effective detection of biases.
- Accessible method to all auditors and communities irrespective of their level of technical expertise (for manual scraping) and available resources.

Limitations:

- Depending on the jurisdiction and the terms of service of the platform, using sock puppets may be illegal.
- The system under investigation may flag suspicious behaviour when using sock puppet accounts, producing results that are not representative of the experiences of real users.
- The manual creation of sock puppet accounts can be a time-consuming and laborious process.
- Sock puppets produce a limited approximation of system response to user behaviour since they lack embedded client-side information such as cookies.

## 5.2.4. Crowdsourcing

Method for collecting data of users' regular interactions with a platform, which can be done through voluntary data donations or automated collection using browser extensions or other software.

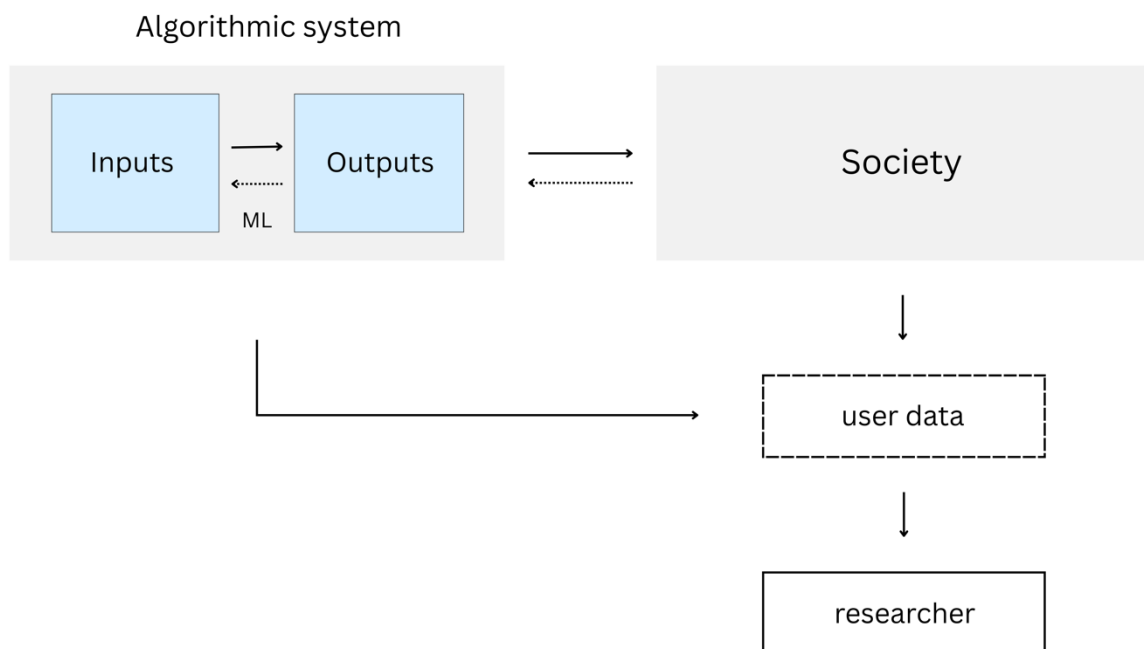


Figure 8. Visualization of the interaction between an algorithmic system and society  
Source: Eticas

When to use this method:

- When auditing a web- or app-based system which allows users to download their data such as social media platforms, search engines, e-commerce websites, online comparison tools or apps in the sharing economy.

Strengths:

- Reflective of real users' experience and the most accurate approximation between the interaction between an algorithmic system and society (via users).
- Direct involvement of the user community.

Limitations:

- Difficult to reach wide audiences and collect representative samples.
- Solutions for automated data collection require expertise and resources as they need to be custom-made for each platform and may require frequent maintenance.
- While they provide rich insight into user experience, crowdsourcing audits alone cannot determine the source of bias or inefficiency.

## 5.2.5. Experimental user audit

The experimental user audit is a systematic method for observing and recording system responses to real user behaviours under different conditions predetermined by the auditor. While the users are authentic, their interactions with the system are performed by design, rather than reflecting their normal engagement with a system (as in crowdsourcing).

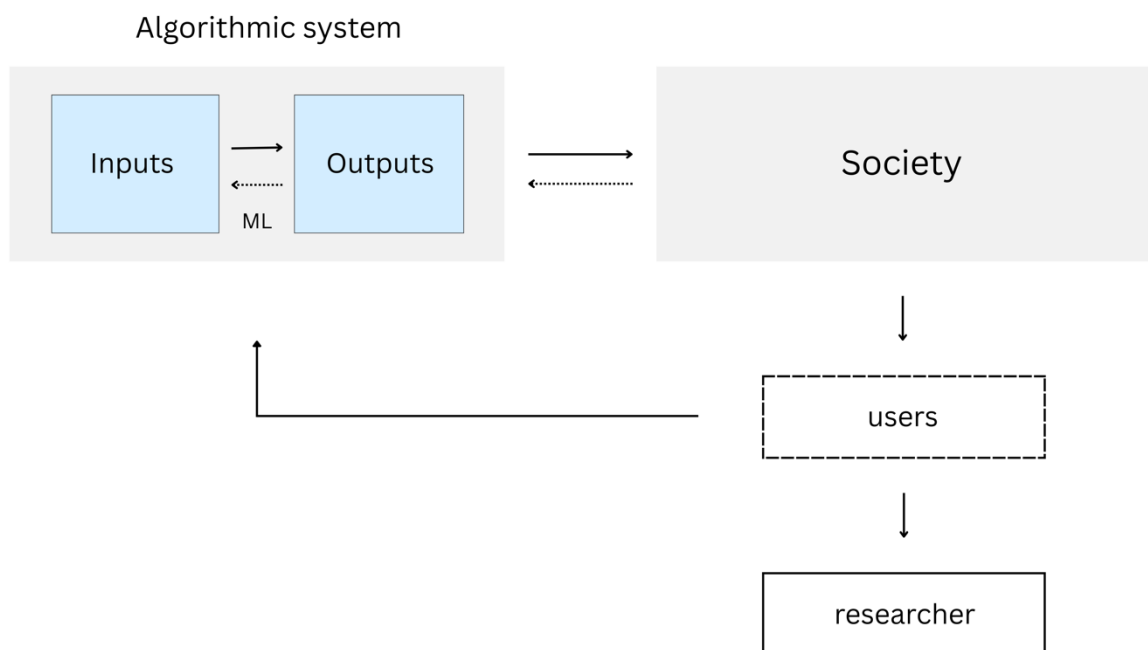


Figure 9. Visualization of the interaction between an algorithmic system and society  
Source: Eticas

When to use this method:

- When auditing systems accessible to users, including web- and app-based systems available for public use or a specific group.
- Particularly suitable for systems that do not respond well to programmatically constructed traffic, such as computer vision and risk assessment algorithms that require human participation.
- Useful for small-scale audits testing machine behaviour towards characteristics that are difficult to replicate via automated queries, such as the performance of facial recognition on people with disabilities.

Strengths:

- Like the sock puppet method, experimental user audits are an effective method to observe system outputs and identify patterns across different conditions, facilitating comparison and detection of biases.
- Results from experimental user audits provide closer approximations of real users' interactions with an algorithmic system compared to programmatically constructed traffic.
- Direct involvement of affected communities.

Limitations:

- Difficult to execute on a large scale.
- Difficulty in recruiting participants with specific characteristics.

## 5.2.6. Comparative output audit

A comparative output audit involves comparing an algorithm's predicted outcomes with the actual outcomes or comparing the performance of one system against another, a benchmark, or a statistical measure for accuracy. In the case of chatbot-like models, this can also include the use of an *LLM as a judge*<sup>19</sup>, where one model evaluates the outputs of another according to predefined criteria such as coherence, bias or toxicity.

---

<sup>19</sup> Dawei Li, et al. "From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge". 2025. <https://arxiv.org/abs/2411.16594>

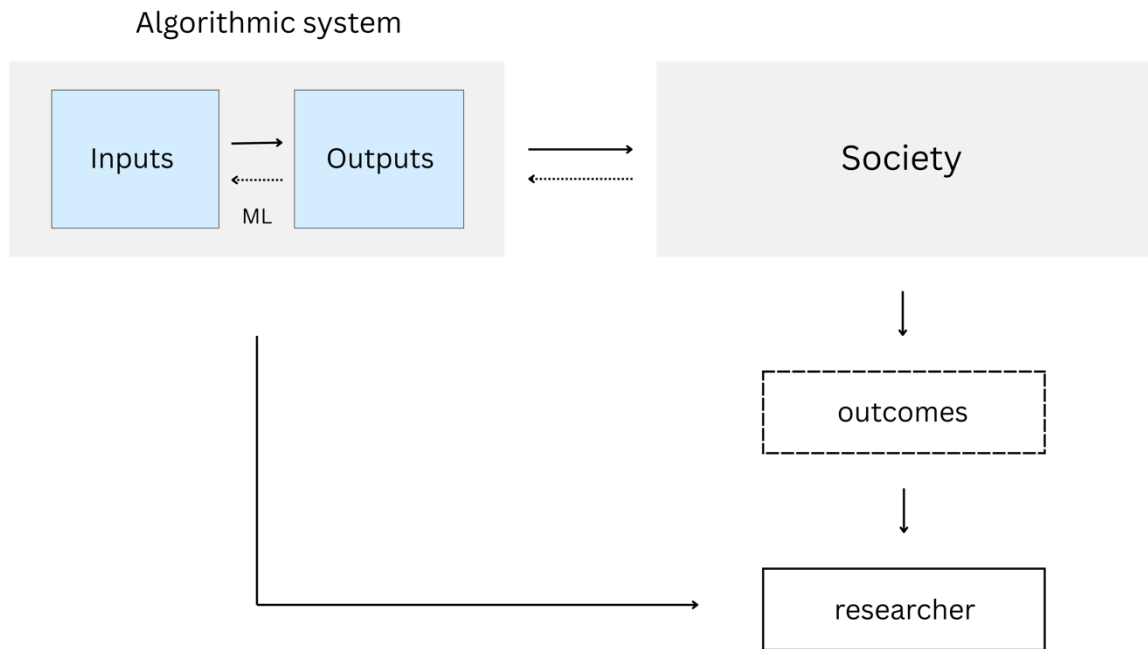


Figure 10. Visualization of the interaction between an algorithmic system and society  
Source: *Eticas*

When to use this method:

- Suitable for systems with publicly disclosed outputs and available information about actual outcomes e.g., risk assessment algorithms used in the public sector or systems that can be tested like facial recognition software.
- Appropriate for generative AI or conversational systems, where using an LLM as a judge can help assess outputs against structured rubrics.

Strengths:

- Based on accurate representation of an algorithm's outputs (e.g., predictions or risk scores) using publicly disclosed data, rather than approximations or subjective user experiences.
- Enables comparison between different systems.
- When applied to chatbot-like models, using an LLM as a judge allows for scalable and structured evaluation across large sets of outputs.

Limitations:

- Difficult to perform due to a lack of transparency in disclosing algorithm information.
- Revealing errors in the algorithm alone is not sufficient to assess efficiency or impact.

## 5.2.7. Ethnographic audit

An ethnographic audit is a qualitative method for data collection through observation, interviews and surveys to understand and analyse how end users, particularly vulnerable groups, interact with an algorithmic system.

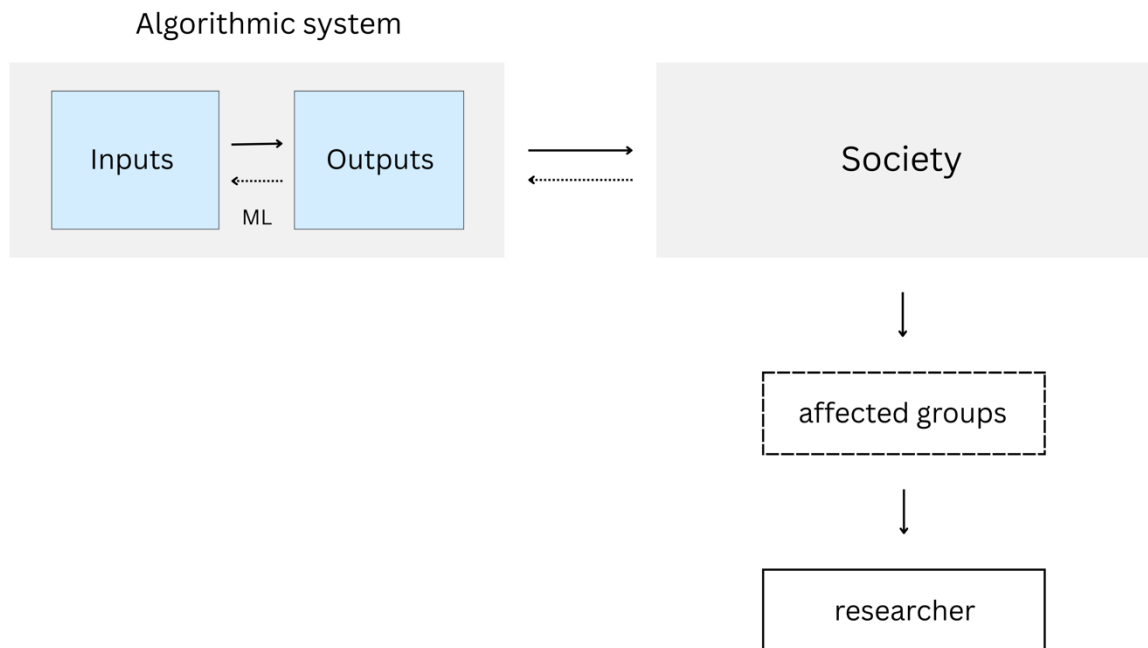


Figure 11. Visualization of the interaction between an algorithmic system and society  
Source: *Eticas*

When to use this method:

- When a vulnerable group or an affected community has been identified, and auditors can reach out to members of those groups or communities for qualitative research.

Strengths:

- Inclusive approach which considers the lived experiences of vulnerable groups or affected communities.
- Direct involvement of vulnerable groups or affected communities, and opportunities to seek redress.

Limitations:

- Difficult to execute on a large scale
- Ethnographic research is helpful for identifying harmful effects and inefficiencies in algorithmic systems as well as structural factors affecting algorithm implementation, but it may not provide definitive answers about biases in the system.
- Experiences of different user groups may be subjective, so it is important to include a representative sample of stakeholders.

The seven methods for adversarial algorithmic auditing we propose consider the nature of the AI system at hand, the context in which it operates and the type of information available to the auditor. This approach allows the auditor to select the most effective means to inspect an AI system, assess its behaviour and quantify its impact. This affords flexibility in audit design according to the strengths and weaknesses of each method in different use cases, while ensuring a high level of robustness and consistency across audits of similar systems. As such, this approach to adversarial algorithmic auditing provides an effective mechanism for AI inspection and accountability.

### 5.3. Audit Report Index

It is recommended that an adversarial audit results in a comprehensive report that outlines the methodology, findings, and recommendations derived from the auditing process. The report serves as a critical document that details the assessment of the system's security and identifies areas in need of improvement. It is recommended that the adversarial audit report includes the following basic structure, however, additional information or restructuring may be necessary depending on the specifics of each audit:

*Glossary*

- I. Introduction: purpose, scope & objectives*
- II. S.o.T.A./Background/Context*
- III. Methodology*
- IV. Results*
- V. Discussion*
- VI. Recommendations and mitigation strategies*
- VII. Limitations*
- VIII. Conclusion*

*References*

*Acknowledgements*

*Annexes*

## 6. Insights from use cases

Eticas has undertaken a wide range of community-led audits over the last seven years. These are summarised in the appendix, but the table below provides an overview, split into five segments:

1. Risk assessment algorithms;
2. Social media platforms;
3. Facial recognition systems;
4. Consumer platforms; and
5. Health and Safety technologies

Audit	AI system	Domain	Methods
<i>Risk assessment algorithms</i>			
<b>VioGén audit</b>	Risk assessment algorithm	Law enforcement	Ethnographic audit; Comparative output audit
<b>RisCanvi</b>	Multi-level risk assessment algorithm	Justice: assess recidivism and violence risk	Ethnographic audit; Comparative output audit
<i>Social media platforms</i>			
<b>YouTube audit</b>	Search algorithm; Recommendation algorithm	Social media and internet platforms	Scraping audit; Sock puppet audit; Ethnographic audit
<b>TikTok audit</b>	Recommendation algorithm	Social media and internet platforms	Scraping audit; Sock puppet audit
<i>Facial recognition systems</i>			
<b>Use of facial recognition on people with disabilities</b>	Insurance and commercial facial recognition systems	Facial recognition for risk assessment; Facial attribute analysis	Experimental user audit; Open-source code audit; Ethnographic audit

## Consumer platforms

<b>Ride-hailing platform audit</b>	Pricing algorithms in consumer platforms	Sharing and gig economy	Scraping audit; Ethnographic audit
<b>Ride-hailing platforms for Roma people</b>	Supply-demand prediction; Surge/dynamic pricing; Driver-rider matching	Sharing and gig economy	Sock puppet audit

## Health & safety technologies

<b>Vape Detector Audit</b>	Audit of physical surveillance systems with algorithmic components	Surveillance technologies in education	Desk research audit; Technical/lab audit; Implementation audit; Ethnographic audit.
----------------------------	--	--	---

By examining these case studies, readers can gain a deeper understanding of the steps in the auditing process and learn how methods for community-led auditing can be combined to assess the impact of algorithms in different domains.

Looking across the case studies, the main lessons for running community-led audits are:

### 1. Limited data access is not a barrier to an audit

Most third-party audits can't access the source code or training data, but this does not mean that nothing can be done. Auditors can use proxies like public datasets, sock puppets, scraping, or comparative testing. Even partial or constrained audits can add value if their limits are documented transparently.

- *Example:* The **VioGén audit** used homicide data to test risk scores when the algorithm was not available to test
- *Example:* The **TikTok audit** relied on sock puppet accounts across cities and timeframes to reveal how political content was suppressed.

### 2. Combine methods for robust evidence

Using both statistical analysis and qualitative research makes audit findings more credible and complete. Quantitative evidence highlights measurable patterns, while qualitative insights explain their causes and consequences in people's lives.

- *Example:* The **RisCanvi audit** mixed statistical analysis of 3,600 inmate cases with interviews of former inmates and lawyers.
- *Example:* The **Facial Recognition audit** tested two Facial Recognition models, while also conducting expert interviews to situate technical errors in their social and legal context.

### 3. Use meaningful comparison groups

Comparative designs (control groups, different contexts, or timeframes), help confirm patterns and rule out coincidence.

- *Example:* The **Facial Recognition audit** compared outputs for people with Down Syndrome against a control group, showing disproportionate errors.
- *Example:* The **Ride-hailing platforms for Roma people audit** contrasted Roma and non-Roma neighbourhoods to reveal unequal ride availability and wait times.

### 4. Work to build community trust

Partnering with affected groups and civil-society organisations provides access, context, and legitimacy. Community voices ground technical findings in lived realities, making results more meaningful.

- *Example:* The **VioGén audit** included survivors of gender-based violence through the Ana Bella Foundation.
- *Example:* The **YouTube audit** drew on migrant roundtables to interpret scraped data on biased visual portrayals.

### 5. Design realistic and relevant tests

Audits should replicate real-world situations that matter for affected groups, ensuring findings have tangible social impact.

- *Example:* The **Ride-hailing platforms for Roma people** tested fares on routes to hospitals and in low-income areas, exposing unfair pricing.
- *Example:* The **Facial Recognition audit** focused on insurance-relevant variables like age and BMI to show direct consequences.

### 6. Probe closed systems with open-source mirrors

When commercial systems are opaque, combining their outputs with open-source models helps audits interrogate likely training data, biases, and design flaws.

- *Example:* The **Facial Recognition audit** tested Zurich's Azul model alongside an open-source model to reveal both practical risks and structural causes of bias.

### 7. Test across multiple contexts and timeframes

Geographic variation, user profiles, and longitudinal data collection help reveal homogenisation, invisibility, or shifts that single-snapshot audits miss.

- *Example:* The **YouTube audit** compared results for migrant vs. non-migrant profiles across London and Toronto.
- *Example:* The **TikTok audit** tracked nine profiles across three cities before, during, and after the 2022 U.S. midterm elections.

## 8. Plan for setbacks and adapt methods

Audits often face obstacles such as lack of access or limited data. Being flexible, by publishing interim results, shifting to lab simulations, or adjusting methods, helps maintain momentum and keeps community attention on the issue.

- *Example:* The **Vape Detector audit** published early planning work when real-world data access stalled.
- *Example:* The **RisCanvi audit** turned the refusal of authorities to share access into part of its political message, showing how denial itself can be evidence.

The above lessons are broad brush. We encourage you to go to the appendix to look at the details of the case studies that are the closest analogue to the audit that you are either planning or undertaking.

## 7. Conclusion

As artificial intelligence becomes ever more pervasive in daily life, its potential for adverse impacts on communities also increases. From risk assessment in justice systems to ride-hailing platforms, social media, and facial recognition, AI now influences decisions and opportunities in ways that can deepen existing inequalities or create new ones. These risks are most acute for vulnerable communities, who are often the least consulted in the design and deployment of these systems. Without meaningful accountability, the spread of AI threatens to entrench opacity, bias, and unfairness.

CLAs provide a vital counterweight to these risks. They allow those most affected to measure the real-world impacts of AI systems, challenge harmful practices, and demand accountability from the organisations that design and deploy such systems. By combining technical investigation with community knowledge and lived experience, CLAs make visible the harms that otherwise remain hidden and provide evidence for change.

This guide has outlined a clear, step-by-step process for conducting CLAs — from choosing a system to audit, to planning, execution, and making recommendations. It has also provided an overview of the main methodological tools, including ethnographic audits, scraping, sock puppets, and comparative output testing. Moreover, the guide has brought to life the steps and the methods by including insights from case studies of CLAs run by Eticas.

CLAs are not just a technical exercise. They ensure that communities have a seat at the table, that their experiences shape accountability, and that AI is held to the same standards of fairness and transparency we expect from any system that affects people's lives.

We close with a call to action. Civil society organisations, community groups, auditors, and technical experts must work together to put CLAs into practice. By doing so, they can expose hidden harms, strengthen accountability, and ensure that AI serves the public interest rather than undermining it. The tools are here. The methods are tested. What remains is for more communities to use them — and in doing so, to reclaim power and agency over the technologies shaping their lives.

## 8. Acknowledgements

**Research Director:** Dr. Gemma Galdon-Clavell, Founder of Eticas.

**Other Contributors:**

- Oliver Smith, Head of Innovation and Business Development at Eticas.
- Catalina María Bernal Murcia, Data Scientist at Eticas.
- Melissa Robles Carmona, Data Scientist at Eticas.
- Sandra Montesinos Zapata, Communication Coordinator at Eticas.

**Recommended citation:** Eticas (2025). A Guide to Community-led Algorithmic Audits. Eticas Foundation..

## 9. Glossary

**Algorithm** - A process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.

**AI System** - Software that is developed with one or more techniques and Machine Learning approaches, including supervised, unsupervised, and reinforcement learning, using a wide variety of methods including deep learning; Logic- and knowledge-based approaches (including knowledge representation), inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems; and statistical approaches, Bayesian estimation, search and optimization methods that can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with (AI Act art. 3.1). The term AI system in this guide refers to the entire technology. For a mobility service, it could be the app that integrates a Machine Learning (ML) model to predict demand and adjust pricing, including, for example, the data pipelines and protocols.

**AI Model** - The model is the trained algorithm, that is, the rules adapted to a particular domain, which constitute the foundation of the technology we audit. Models are subject to performance evaluation, and tests, and can be compared to each other via benchmark datasets. The model is the core of an AI system, but it usually relies upon other elements (e.g., data pipelines, visualization platforms) for it to work. An AI system can include more than one model.

**Algorithmic auditing** – A method for thoroughly examining AI systems within their unique contexts. It encompasses an approach and methodology that enable a comprehensive evaluation of regulations, standards, and overall impacts. Additionally, when the results of these audits are made public, they serve as valuable tools for enhancing transparency and fostering greater accountability.

**Risk assessment** - The process of evaluating the likelihood and severity of harm that may result from the processing of personal data. It helps identify potential risks and vulnerabilities and guides the development of appropriate safeguards and controls to mitigate those risks ([GDPR, Rec. 76](#); pre-deployment period, [Ada Lovelace Institute, 2020](#))

**Impact assessment** - Impact assessment, on the other hand, focuses on the potential impact of data processing activities on individuals' personal data rights. It helps identify potential risks and harms to individuals and guides the development of appropriate measures to protect those rights ([GDPR, Art. 35](#); post-deployment period, [Ada Lovelace Institute, 2020](#)).

**Recommender systems** - A subclass of information filtering system that provides suggestions for items that are most pertinent to a particular user.

**Large Language Model (LLM)** - A type of AI model trained on massive collections of text data to predict and generate human-like language. LLMs rely on deep learning techniques, particularly transformer architectures, to perform a wide variety of natural language tasks such as answering questions, generating text, summarising, translating, or evaluating content. They are foundational models that can be adapted for specific applications, including chatbots, content moderation, or acting as evaluators ("LLM as a judge") of other models' outputs.



## 10. Appendix: Case studies

This appendix provides an overview of several of Eticas' community-led audits. Each case study includes: the context, an overview of the planning steps (aligned to steps 1 to 6 in the guide); and overview of the execution steps (aligned to steps 7 to 10 in the guide); a summary of the findings; and key lessons learned.

### 10.1. Auditing risk assessment algorithms

Risk assessment tools are AI or algorithmic systems used for decision making and are often employed by the public sector in fields such as criminal justice, welfare, healthcare and housing. However, these tools have the potential to negatively impact protected classes and marginalised groups.

#### 10.1.1. VioGén Audit

##### 1. Context

The VioGén System ("Integral Monitoring System in Cases of Gender Violence")<sup>20</sup> is Spain's national platform for assessing and managing risks faced by women who report gender-based violence. It coordinates law enforcement, judicial, health, and social services through standardised risk assessments. Since its launch in 2007, it has generated over three million risk evaluations, making it the largest system of its kind worldwide.

Eticas selected VioGén for a community-led audit because of its profound impact on vulnerable populations and the structural risks of using algorithmic tools in sensitive social contexts. Concerns included the lack of transparency, accountability, and independent oversight, the limited involvement of end-users, and the possibility of the system evolving into a machine learning model without public debate. The audit was conducted in collaboration with the Ana Bella Foundation, a civil society organisation supporting survivors of domestic violence.

##### 2. Planning Steps

Planning began with the recognition that internal access to VioGén's datasets and algorithm was not possible. Our team set out to:

---

<sup>20</sup> Eticas. (2022). The External Audit of the VioGén System. Eticas Research and Innovation. <https://eticasfoundation.org/the-case-of-viogen-can-ai-solve-gender-violence/>

- Combine quantitative and qualitative (Comparative output audit and Ethnographic audit, respectively) methods to capture both technical performance and social impact.
- Use publicly available homicide datasets from the General Council of Judicial Power (CGPJ) to assess predictive validity across groups.
- Recruit survivors of gender violence, lawyers, and civil society actors for interviews and surveys to document lived experiences.
- Frame research questions around transparency, accountability, barriers to access, and group-specific vulnerabilities.

This design ensured that even without system access, the audit could surface biases, misrepresentations, and gaps between design and lived experience.

### 3. Execution Steps

The audit was carried out over seven months and unfolded in three main streams:

- **Quantitative analysis:** Using a dataset of 475 intimate partner homicide cases (2009–2019), our researchers assessed false negative rates and recall disparities across groups (e.g., women with/without children, migrants, older women). We found that women without children were systematically assigned lower risk scores, revealing a structural bias in the model's calibration.
- **Qualitative fieldwork:** 31 survivors, 7 lawyers, and 2 civil society representatives were interviewed. Many women described emotional, structural, and institutional barriers to accessing the system. 80% reported problems with the VPR questionnaire, ranging from poor timing (conducted in moments of shock) to lack of information, inadequate legal/psychological support, and ambiguous or rigid questions.
- **Contextual evaluation:** Our team analysed systemic issues such as police discretion, insufficient training, and the disproportionate number of "unappreciated" risk scores (around 45% of cases). They also scrutinised the governance implications of potential machine learning integration without transparency or debate.

### 4. Findings

This audit found that, despite its role in reducing recidivism and coordinating protection, the system suffers from serious transparency and fairness issues. Most victims were unaware of how their risk scores were calculated, limiting their ability to understand or contest decisions. Quantitative analysis showed that migrant women, older women, women with disabilities, and those with dependents were more likely to receive inaccurate risk assessments, with dangerous false negatives leaving some without adequate protection.

Also, some implementation problems emerged. For example, police training was inconsistent across regions, producing uneven application of the system. Victims described the questionnaire process as intrusive and poorly timed, often conducted during moments of trauma. Lawyers and advocates noted that the algorithm reduces complex

human situations to rigid numerical categories, sidelining professional judgment and victim agency.

Overall, while VioGén remains central to Spain's fight against gender violence, its opacity, rigid design, and uneven application risk reproducing inequalities and eroding trust among those it seeks to protect.

## 5. Key Lessons

- Audits are possible without direct system access: Even without entry to VioGén's internal datasets or algorithm, combining publicly available homicide records with qualitative methods revealed structural biases and transparency issues.
- Pairing quantitative and qualitative approaches strengthens evidence: Statistical analysis of homicide cases uncovered disparities in false negatives, while ethnographic interviews with survivors and legal experts exposed how the system's design and application shape lived experiences.
- Group-specific analysis is crucial: Stratifying results by age, migration status, and parental status uncovered disparities, especially for women without children, who systematically received lower risk scores.
- Civil society participation adds legitimacy: Partnering with the Ana Bella Foundation ensured survivors' voices were central, grounding technical findings in real human experience.

## 10.1.2. RisCanvi Audit

### 1. Context

The RisCanvi audit was the first community-led audit of a criminal justice AI system in Europe<sup>21</sup>. RisCanvi, implemented in Catalonia since 2009, is used to assess recidivism and violence risk, influencing parole, sentencing, and prison regime decisions. Despite its central role, the system remained largely unknown to inmates and poorly understood even by professionals. Concerns about fairness, reliability, transparency, and compliance with existing regulations motivated the audit. The project aimed not only to evaluate RisCanvi's technical validity but also to surface its social impacts, drawing attention to the asymmetries of power and information faced by those most affected.

### 2. Planning Steps

---

<sup>21</sup> Eticas (2024). Automating (In) Justice? An Adversarial Audit of RisCanvi. Eticas Research and Innovation. <https://eticasfoundation.org/automating-injustice-an-adversarial-audit-of-riscanvi/>

- Defined an **adversarial audit approach**, since no direct access to the algorithm or datasets was available.
- Partnered with **Iridia**, a human rights NGO, to build trust with affected communities and strengthen legitimacy.
- Designed a **mixed-method framework** combining quantitative analysis of public recidivism data with qualitative ethnographic research.
- Identified a dataset of **3,600 released inmates** for comparative statistical analysis of recidivism and risk factors.
- Developed **interview protocols** for former inmates, families, lawyers, psychologists, educators, and activists.
- Shaped the audit's research questions around **fairness, transparency, accountability, and social impact** of RisCanvi.

### 3. Execution Steps

- **Ethnographic audit:** Eighteen interviews with former inmates, psychologists, educators, lawyers, activists, and family support representatives explored awareness, trust, fairness, and legitimacy of RisCanvi.
- **Comparative output audit:** Public data on 3,600 released inmates was analysed statistically to test relationships between risk factors, behaviours, and outcomes. Techniques included regression analysis, intersection analysis, clustering, and factor prevalence studies. Together, these methods exposed gaps between the system's assumptions and actual recidivism patterns. Findings revealed opacity in calculations, overreliance on static factors like childhood circumstances, and inconsistent or seemingly random scoring practices.

### 4. Findings

Quantitative analysis of over 3,600 cases showed that the system heavily relied on static, biographical factors, such as childhood experiences or family background, that individuals cannot change, while undervaluing dynamic factors linked to rehabilitation. This design undermined fairness and weakened incentives for reintegration. At the same time, qualitative interviews with former inmates, lawyers, psychologists, and family support groups highlighted that risk scores were often perceived as arbitrary, inconsistent, or even random.

This audit concluded that RisCanvi not only lacks transparency but also risks producing unjust outcomes by locking people into risk categories based on unchangeable characteristics, rather than reflecting actual behaviour or progress in rehabilitation.

### 5. Key Lessons

- Building trust with former inmates was essential: without careful collaboration, many would not have shared experiences of how risk scores shaped their prison life.
- Using public recidivism datasets allowed the team to partially reconstruct the algorithm's logic but also highlighted how little can be known without official access.
- Interviewing professionals inside and outside prisons uncovered contradictions between technical expectations of the system and its real-world application.
- Combining statistical outputs with lived experiences gave the audit credibility across communities, showing that neither approach alone would have been sufficient.
- The refusal of authorities to grant access turned the audit into a political act as much as a technical one, underscoring the role of civil society in demanding accountability.

## 10.2. Auditing social media

Auditing social media is challenging due to the large, complex and dynamic algorithmic systems employed by internet platforms. At the same time, social media are vastly influential yet scarcely transparent, making community-led audits a key mechanism for inspection and accountability.

Our community-led audits of YouTube and TikTok illustrate how to inspect social media recommender systems. They exemplify how a mixed-method socio-technical approach to auditing can help identify biases and problematic behaviours in the systems used by internet platforms and provide insights into how users perceive and interact with content on these platforms. The case studies below demonstrate how the use of scraping and sock puppet audit methods, combined with qualitative ethnographic research, allows auditors to conduct comprehensive assessments of complex AI systems with multiple dynamic elements and gain a nuanced understanding of the issues they present.

### 10.2.1. YouTube Audit

#### 1. Context

YouTube is a globally popular social media platform and plays a central role in shaping public opinion, including on migration. While traditional media has long portrayed migrants in negative, stereotypical ways, social media adds a new layer of algorithmic influence through opaque recommendation and search systems. This makes it critical to understand how migrants and refugees are represented on such platforms.

Our audit<sup>22</sup> examined YouTube's portrayal of migrants and refugees because prior research suggested systemic misrepresentation, dehumanisation, and stereotyping. Our audit confirmed that migrants tend to be shown in disadvantaged frames (large groups of non-white people crossing borders with invisible faces) while refugees, particularly white Ukrainians, were depicted more humanely, in small groups with visible faces.

## 2. Planning Steps

1. **Research questions defined:** How are migrants and refugees represented in top YouTube videos? Do algorithms vary by geography (e.g., UK vs. Canada)? Do they differ for migrant vs. non-migrant accounts? How do migrants themselves perceive portrayals?
2. **Audit design:** A mixed-method socio-technical approach was chosen to combine quantitative data scraping with qualitative interviews.
3. **Data collection scope:** Focused on thumbnails, titles, and descriptions of top-watched and recommended videos for the queries "migrants" and "refugees." Thumbnails were prioritized since they are the first and most influential element shaping user impressions.
4. **Comparative contexts:** Sock puppet accounts (a migrant profile and a non-migrant profile) were created, and searches were run via VPNs in London and Toronto to test for contextual differences.
5. **Community engagement:** A roundtable with migrants was planned to ensure lived experiences were integrated into the findings.

## 3. Execution Steps

- **Quantitative analysis:**
  - Scraped 100 top-watched videos each for "migrants" and "refugees."
  - Conducted geolocation tests (London vs. Toronto) and account-based tests (migrant vs. non-migrant).
  - Coded thumbnails by variables such as group size, race, gender, visibility of faces, activities (border crossing, work, protest), and framing (victim, problem, beneficial, ambiguous).
- **Sock puppet audits:**
  - Created two blank accounts: "Kate White" (non-migrant, white British profile) and "Fatma Aydin" (migrant, Turkish background profile).
  - Both received nearly identical recommendations: migrants consistently framed as non-white, faceless groups crossing borders, mostly in victim or problem frames.
- **Qualitative work:**

---

<sup>22</sup> Eticas (2023). Auditing YouTube. Social media's treatment of migrants and migration. Eticas Research and Innovation. <https://eticasfoundation.org/auditing-youtube-social-medias-treatment-of-migrants/>

- Conducted a roundtable with four migrants to discuss perceptions of representation.
- Participants noted that positive portrayals are rare and underpromoted, while negative depictions (mass arrivals, “waves,” and border-crossing scenes) dominate and reinforce dehumanization. They also highlighted the double standards in depictions of Ukrainian refugees vs. other groups.

#### **4. Findings**

Our team found that migrants were overwhelmingly depicted as large, faceless groups of non-white people crossing borders, often in frames of crisis or threat. In contrast, refugees, particularly white Ukrainians, were more frequently shown as small groups with visible faces, framed in more sympathetic and humanizing ways.

With the two profiles, a migrant user and a non-migrant user tested in London and Toronto, we found that, despite different contexts, the recommendations were nearly identical, suggesting that YouTube’s algorithms produce homogenized representations of migration that are largely insensitive to location or user profile.

Finally, our qualitative research added depth. For example, in a roundtable discussion, migrants confirmed that positive depictions of migration are rare, while negative portrayals dominate and reinforce harmful stereotypes. They also noted double standards in portrayals of different migrant groups, which further entrenches feelings of exclusion and discrimination. The combination of scraping and testimonies highlighted how YouTube’s recommendation systems amplify biased imagery while suppressing more balanced or positive narratives.

#### **5. Key Lessons**

- Thumbnail analysis was crucial: coding visual elements gave stronger insights into bias than relying only on video titles or descriptions.
- Sock puppet testing in two countries showed limits of personalization: the near-identical results confirmed that such audits must test multiple contexts to expose homogenization.
- Paired migrant and non-migrant profiles provided contrast: this setup helped rule out user identity as a driver of recommendations.
- Community engagement shaped interpretation: migrant testimonies ensured that coded data on stereotypes was linked to real-world consequences.
- Comparing search terms uncovered hidden patterns: auditing both “migrants” and “refugees” demonstrated how wording choices in audit design can expose algorithmic bias.

## 10.2.2. TikTok Audit

Following our YouTube audit, we also conducted a community-led audit of TikTok to see how the platform's algorithms shape political discourse on migration (Eticas, 2023).

### 1. Context

TikTok has rapidly become one of the most influential social media platforms, especially among younger generations. Known primarily as an entertainment app, TikTok also plays an increasingly important role in shaping public opinion and youth politics. Its recommender system curates personalised feeds on the "For You" page, where most users consume content. This makes it a powerful force in framing global issues such as migration.

We selected TikTok for an external audit<sup>23</sup> because of its wide reach, opaque recommendation algorithms, and its relevance during politically sensitive moments like the 2022 U.S. midterm elections. Migration was a central electoral issue, providing an opportunity to test whether TikTok amplified or suppressed political discourse on the topic. The audit focused on whether TikTok's recommendations varied depending on user attitudes toward migration, political leanings of their location, or timing around the elections.

### 2. Planning Steps

- Defined the scope: examine visibility of political discourse on migration during the 2022 U.S. midterm elections.
- Selected three U.S. cities with different political leanings (San Francisco (Democrat), Oklahoma City (Republican), and Virginia Beach (ambivalent)).
- Created **nine sock puppet accounts**: one pro-migrant, one anti-migrant, and one neutral profile in each city.
- Designed training routines for each profile:
  - Pro-migrant: followed NGOs, activists, and hashtags like #immigrantrights.
  - Anti-migrant: engaged with stricter border-control accounts and hashtags like #buildthewall.
  - Neutral: interacted only with pet content using hashtags like #dogsoftiktok.
- Established collaboration agreements and research protocols to ensure methodological rigor and ethical compliance.

### 3. Execution Steps

---

<sup>23</sup> Eticas (2023). Auditing TikTok. Social media's treatment of migrants. Eticas Research and Innovation. <https://eticasfoundation.org/auditing-tiktok-social-medias-treatment-of-migrants/>

- Trained sock puppet accounts for six days by liking, sharing, and following migration-related or neutral content creators.
- Conducted an additional day of reinforcement training during the election period to simulate active user engagement.
- Scraped **20 recommended videos per account per day** from the "For You" feed before, during, and after the midterm elections (total: 1,620 videos).
- Collected metadata for each video, including author, hashtags, likes, and description, to contextualize recommendations.
- Coded videos into categories (political vs. non-political) and subcategories (e.g., migration, race, gender, economy).
- Applied frames (security, economy, validity, policy) and sentiment analysis (positive, negative, neutral) to political migration content.
- Compared recommendations across profiles (pro-, anti-, neutral), across locations, and across time periods (pre-, during, and post-election).
- Analysed variations and found that political migration content was almost absent, with recommendations dominated by entertainment videos.

#### 4. Findings

Despite creating nine sock puppet accounts with distinct political attitudes (pro-migrant, anti-migrant, and neutral) across three politically different cities, political discourse on migration was virtually absent. Fewer than 1% of recommended videos contained political themes, and only a handful directly addressed migration.

The results showed little variation across user attitudes or locations: pro- and anti-migrant profiles received almost identical feeds dominated by entertainment, lifestyle, and humour content. Over time, the share of political content dropped further, from 1.5% before the election to 0.2% after. Migration appeared only marginally in post-election recommendations. While there were shifts in language diversity and gender representation in content, these did not translate into more visibility for migration or political debate.

The audit concluded that TikTok's recommender system deprioritizes political discourse in favour of entertainment and offers weak personalization around migration-related attitudes. This invisibility is as consequential as misrepresentation, because it erases migration from the political agenda of a platform central to youth engagement.

#### 5. Key Lessons

- Training sock puppet accounts with pro- and anti-migrant hashtags showed that TikTok's algorithm resists account priming, teaching auditors that even intensive training may not trigger meaningful personalization.
- Using multiple profiles across three cities demonstrated the value of over-engineering the setup: it proved that location and political leaning had minimal impact, something a smaller audit design might have missed.

- Collecting data before, during, and after the election revealed a decline in political content over time, showing that timing choices in audits can decisively shape what patterns emerge.
- Designing a broad coding scheme for both political and non-political videos allowed the team to capture invisibility itself as a result, a reminder that sometimes the absence of expected content is the core audit finding.
- The audit showed that invisibility is harder to prove than bias, and that careful methodology (multiple profiles, longitudinal scraping, and broad coding) is required to turn “nothing to see” into robust evidence.

## 10.3. Auditing facial recognition

Our audit of the use of facial recognition (FR) in the insurance sector demonstrates how to assess the ethical and legal compliance of FR technology within a specific domain. This approach can be useful for similar audits of facial recognition technology in other sectors, providing a more in-depth understanding of the impact and implications of this technology.

### 10.3.1. Use of facial recognition on people with disabilities

#### 1. Context

The investigation concentrated on two types of systems<sup>24</sup>. The first, Azul, is a facial recognition tool developed by Zurich Insurance Group to estimate age, body mass index (BMI), and smoking status in order to calculate insurance premiums. The second, DeepFace, is an open-source framework widely used in commercial FR models for demographic and emotional analysis, including attributes such as age, gender, ethnicity, and emotion.

Given the structural challenges that people with disabilities face, including barriers to employment, higher poverty rates, and health risks, the use of FR in sensitive areas like insurance pricing raises significant ethical and legal concerns. By examining both a commercial insurance application and a widely used open-source FR system, the audit sought to reveal whether these technologies reproduce or amplify discrimination against people with disabilities

#### 2. Planning steps

---

<sup>24</sup> Eticas (2023). Invisible No More: The impact of facial recognition and Price discrimination AI on people with disabilities. Eticas Research and Innovation. <https://eticasfoundation.org/invisible-no-more-the-impact-of-facial-recognition-on-people-with-disabilities/>

The planning stage aimed to design a methodology that could expose how facial recognition systems interact with people with Down Syndrome, combining technical testing with expert perspectives.

- **Choosing the systems:** The audit focused on Azul, Zurich Insurance's FR tool for risk assessment, and DeepFace, a widely used open-source FR framework.
- **Contextual analysis:** Background research highlighted the lack of studies on disability and FR, despite evidence of bias in relation to gender and race. This justified focusing on Down Syndrome as a case where exclusion is most visible.
- **Stakeholder input:** Semi-structured interviews were conducted with technical, legal, social, and policy experts to understand the broader risks of FR for people with disabilities.
- **Methodology design:** A two-part plan was established: (1) an experimental user audit with 20 participants with Down Syndrome and 20 without, to test Azul; and (2) pilot tests with DeepFace, using curated datasets of people with and without Down Syndrome, to evaluate bias in age, gender, ethnicity, and emotion recognition.

### 3. Execution steps

The audit was carried out through a combination of experimental testing and expert input to capture both technical and social dimensions of bias.

- **Expert interviews:** Four semi-structured interviews were held with specialists in technology, law, policy, and social issues to frame risks and ethical concerns around FR for people with disabilities.
- **Azul testing:** Forty participants (20 with Down Syndrome, 20 without) used Zurich's Azul system. The tool's outputs for age, BMI, and smoking status were recorded and compared with actual data to assess accuracy and fairness.
- **DeepFace testing:** Two datasets were compiled: 60 images of people with Down Syndrome and 60 of well-known individuals without. DeepFace was then tested for age, gender, ethnicity, and emotion classification to identify disparities.
- **Analysis:** Statistical measures (e.g., error rates, mean absolute error, recall) were applied to compare results across groups and highlight patterns of bias

### 4. Main findings

The audit found that facial recognition systems performed poorly and unevenly when tested with participants with Down Syndrome. Age prediction was highly inaccurate in both Zurich's Azul system and commercial FR models, with errors larger for disabled participants. In Azul, strong gender disparities emerged: women's ages were drastically underestimated, sometimes to the point of classifying adults as children, while men's ages were consistently overestimated, raising both fairness and legal concerns. Commercial models also showed weaknesses in gender classification, with much lower accuracy for women with Down Syndrome compared to controls. Emotion classification was weak across all groups, but predictions were made with less confidence for disabled

participants. Finally, ethnicity was often misclassified, particularly for Asian and white participants with Down Syndrome, revealing structural gaps in how these models represent human diversity.

## 5. Key lessons

- **Expert interviews add depth:** Conversations with specialists in technology, law, policy, and social issues grounded the technical results in their social and legal consequences, showing how misclassifications directly affect rights and protections.
- **Combine commercial and open-source models:** Testing Zurich's Azul revealed practical risks in insurance pricing, while using DeepFace made it possible to explore hidden aspects like training data, representativity, and structural bias that commercial systems keep opaque.
- **Representativity is crucial:** Including a control group of participants without Down Syndrome allowed for statistically meaningful comparisons, showing that errors disproportionately affected disabled users.
- **Design realistic tests:** By focusing on variables tied to real-world outcomes—age, BMI, smoking status in insurance, and demographic/emotional classification in FR—the audit ensured findings reflected tangible impacts on people's lives.
- **Community-centered framing:** Concentrating on Down Syndrome highlighted a major gap in existing research and showed how community-led audits can bring neglected groups into broader debates on fairness in AI.

## 10.4. Auditing consumer platforms

Auditing consumer platforms providing services in the sharing and gig economy include ride-hailing apps, food and product delivery apps, and marketplaces for homestay such as Airbnb as examples. Like social media, they can employ large, complex and dynamic systems which may be challenging to audit.

Our audit of ride-hailing platforms in Spain is an example of how to conduct a community-led audit of consumer platforms and identify instances of bias and discrimination in their algorithms. This case study demonstrates how scraping and ethnographic audits, and the combination of quantitative and qualitative methods can uncover the harmful impacts of the pricing algorithms used by ride-hailing platforms.

### 10.4.1. Audit of ride-hailing platforms

We examined how the pricing algorithms of Uber, Bolt and Cabify impact competition, workers and consumers in Spain.

#### 1. Context

This audit<sup>25</sup> began with concerns that ride-hailing apps in Spain may not be fully complying with competition, labour, and consumer law. Platforms such as Uber, Bolt, and Cabify dominate Spain's ride-hailing market but operate in a regulatory grey area under the private hire vehicle (PHV) framework. These platforms affect multiple stakeholder groups, including drivers, passengers, taxi companies, and regulators. In particular, PHV drivers and passengers in low-income areas were identified as groups at risk of algorithmic discrimination. Given these conditions, Eticas partnered with Taxi Project, an organization advocating for taxi workers, and Observatorio TAS, an organization defending workers in the platform economy, to conduct an adversarial audit of ride-hailing algorithms. The aim was to test whether these platforms' algorithmic practices complied with competition, labour, and consumer protections.

## 2. Planning Steps

- The audit began with a **contextual analysis**, reviewing relevant Spanish and EU legislation in competition, labour, and consumer law. Expert interviews complemented this desk research, clarifying how ride-hailing platforms operate within ambiguous regulatory frameworks.
- Next, a **stakeholder mapping** exercise identified groups directly and indirectly affected by algorithmic decisions, ranging from PHV drivers and license-holding companies to passengers, regulators, and policymakers. Among these, drivers and passengers in disadvantaged areas were singled out as particularly vulnerable to bias.
- In the **alliance-building** stage, the project formalized roles, responsibilities, and agreements with *Taxi Project* and *Observatorio TAS* to ensure that both worker perspectives and sector expertise informed the audit.
- The **methodology design** combined two approaches: scraping audits of platform pricing to examine issues of competition and consumer law, and ethnographic audits (interviews with drivers) to assess labour-related impacts.

## 3. Execution Steps

- Scraped fare data on selected routes in Madrid and Andalusia to test for price collusion between Uber, Bolt, and Cabify.
  - Found moderate to strong correlations between Uber–Cabify and Uber–Bolt fares, suggesting possible algorithmic price fixing.
- Scraped pricing data for routes in high-, medium-, and low-income neighbourhoods in Madrid and Málaga to test for consumer discrimination.
  - Linear regression revealed a weak to moderate negative correlation between neighbourhood income and fares, meaning trips in low-income areas were more expensive.
- Conducted ethnographic interviews with PHV drivers to assess labour protections.

---

<sup>25</sup> Eticas, the Taxi Project, & Observatorio TAS. (2023). Adversarial audit of ride-hailing platforms: Algorithmic compliance with competition, labor and consumer law in Spain. Eticas Research and Innovation. Taxi Project 2.0. Observatorio TAS

- Drivers reported opaque sanctions for lawful absences and a lack of transparency in platform payment systems, indicating algorithms undermined labour rights.

#### **4. Findings**

The audit of Uber, Bolt, and Cabify in Spain revealed important risks across competition, consumer, and labor law. Scraping of fares in Madrid and Andalusia showed moderate to strong correlations between Uber–Cabify and Uber–Bolt prices, raising concerns of algorithmic price coordination. Further analysis of routes across high-, medium-, and low-income neighborhoods in Madrid and Málaga identified a weak to moderate negative correlation between income and fares, meaning users in lower-income areas were systematically charged more.

Qualitative fieldwork with private hire vehicle (PHV) drivers revealed how algorithmic management practices undermine labor protections. Drivers reported being penalized for lawful absences and facing opaque, non-transparent payment systems. Together, the evidence suggested that algorithmic decision-making not only distorts competition and consumer fairness but also exacerbates precarity for workers, reinforcing vulnerabilities within the platform economy.

#### **5. Key lessons**

- Scraping fare data across multiple cities and routes showed that selecting the right geographic comparisons is crucial to detect both collusion and geographic price discrimination.
- Combining quantitative scraping with ethnographic interviews proved essential: scraping exposed market-level risks, while interviews revealed individual labour harms invisible in data.
- Including low-, medium-, and high-income neighbourhoods in the audit design demonstrated how socioeconomic context shapes algorithmic outcomes.
- The process showed that auditing platforms requires multi-dimensional methods: competition, consumer, and labour issues can only be uncovered when technical data and lived experiences are examined together.

### **10.4.2. Ride-hailing platforms for Roma people**

#### **1. Context**

This audit examined how Uber's algorithmic systems impact access to mobility in Roma neighbourhoods in Madrid. The Roma are the largest ethnic minority in Europe and face persistent structural barriers, including poverty, segregation, and limited access to public transport. Given their high dependence on affordable and reliable mobility, the audit set out to test whether Uber's systems, particularly supply–demand prediction, surge pricing,

and driver–rider matching, provide services of equal reliability in Roma and non-Roma neighbourhoods.

## 2. Planning steps

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.

- **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.
- **Contextual analysis:** Desk research and prior Eticas audits highlighted concerns about fairness in ride-hailing platforms. Input from *Fundación Secretariado Gitano* 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 and 10 nearby control neighborhoods. For each, three destination types were chosen (city center, nearest commercial center, nearest hospital), creating 60 routes for comparative testing.

## 3. Execution steps

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.

- **Data collection setup:** Researchers created sock puppet accounts to simulate rider behaviour and submitted 240 trip requests across 20 neighbourhoods, 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.

## 4. Main findings

The audit revealed clear disparities in Uber's service between Roma and non-Roma neighbourhoods in Madrid. In Roma areas, Uber was unavailable for 27% 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 neighbourhoods. Although preliminary, these findings point to the role of Uber's algorithmic systems in reinforcing unequal access to mobility resources. To address these issues, the audit highlights the need for greater algorithmic transparency and the establishment of legal frameworks that enable independent audits, ensuring fair access to mobility for marginalized communities.

## 5. Key lessons

- **Simulations can replace direct access:** Even without platform data or community participation, the use of sockpuppet accounts made it possible to reproduce rider experiences in a controlled way and collect reliable evidence on service disparities.
- **Representativity strengthens findings:** Including a control group of non-Roma neighbourhoods allowed for meaningful statistical comparisons and demonstrated that service gaps were not random but systematic.
- **Designing for real-life relevance:** Routes were chosen to essential destinations—hospitals, commercial centers, and the city center—so that results reflected not only technical differences in availability and wait times but also the social impact of unequal access to mobility.

## 10.5. Auditing Health & Safety Technologies

### 10.5.1. Vape Detector

#### 1. Context

The project focuses on auditing vaping detection (VD) technologies installed in school settings in the Twin Cities, Minnesota. These devices, marketed as anti-vaping solutions for spaces where cameras are not feasible (like bathrooms), use particulate matter and chemical sensors to detect vaping and send alerts to school staff. Increasingly, VD systems also include additional surveillance capabilities such as audio recording, aggression detection, and integration with video management systems (VMS). Vendors frame them as tools for student health and safety, but these claims are unverified and raise concerns about privacy, surveillance overreach, and disproportionate impacts on vulnerable student groups.

This audit seeks to examine both the technical effectiveness of VD devices and their broader consequences on school environments and student well-being. It also situates VD technologies within larger sociopolitical trends of surveillance in education, where monitoring is normalised in ways that may exacerbate racial and social inequities.

#### 2. Planning Steps

- Defined two overarching research questions:
  - Do VD technologies fulfill the health and safety purposes claimed by vendors?
  - What are their impacts on school environments and society at large?
- Designed three interconnected research streams:
  - **Stream 1:** Technical capabilities and performance, verifying vendor claims and testing detection accuracy.
  - **Stream 2:** Surveillance and impacts, examining how VD implementation affects school dynamics, students, parents, and staff.

- **Stream 3:** Surveillance in education, analysing how surveillance technologies in schools intersect with racial inequities and policing practices.
- Identified sub-streams on transparency, awareness, data privacy, and governance.
- Built a methodology that combined desk research, stakeholder interviews, and quantitative and lab-based testing.
- Defined possible collaborations with schools for real-world data collection, while also preparing fallback lab-based tests with procured devices.

### 3. Executing Steps

Completed

- **Desk research:** Collected vendor documentation, procurement records, and technical specifications.

Planned

- **Stakeholder interviews:** Designed to capture testimonies from students, teachers, parents, administrators, and security staff to understand the lived impacts of VD deployment.
- **Quantitative analysis:** Planned collection of real-world alert data from schools (or lab simulations) to test reliability, accuracy, and false positives in detecting vaping, aggression, or gunshots.
- **Comparative setups proposed:**
  - Schools with VD vs. schools without VD.
  - Schools with add-on surveillance features vs. basic VD.
  - Socio-demographically diverse schools (income, race/ethnicity).
- **Lab testing (backup plan):** Controlled experiments with particulate, vapour, and sound samples, with varying levels of community participation (none, low, high).
- **Coding & evaluation:** Measurement of whether alerts are valid, how staff respond, and what outcomes students face (e.g., counselling vs. punishment).
- **Community engagement:** Town halls and workshops with students and parents to gather perspectives and share findings.

### 4. Findings

None as yet as this project is on pause due to less data being available than was initially anticipated. The lead community organisation is currently considering the best way forward, but in the meantime is publishing some of the planning work as part of maintaining a local focus on this issue.

### 5. Key Lessons

- This audit is a prime example of step 10 in this audit process – managing limitations and set backs.

- In this instance, the community organisation has reacted to an unanticipated change in data availability by publishing some early work. The intention being to attempt to resolve the data availability issue by maintaining attention to the issue.

## 11. Bibliography

- Abid, A., Farooqi, M., & Zou, J. (2021). Large language models associate Muslims with violence. *Nature Machine Intelligence*, 3(6), 461–463. <https://doi.org/10.1038/s42256-021-00359-2>
- Ada Lovelace Institute (2021). Technical methods for the regulatory inspection of algorithmic systems in social media platforms. <https://www.adalovelaceinstitute.org/report/>
- Ali, M., Sapiezynski, P., Korolova, A., Mislove, A., & Rieke, A. (2019). Ad Delivery Algorithms: The Hidden Arbiters of Political Messaging (arXiv:1912.04255). arXiv. <http://arxiv.org/abs/1912.04255>
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. ProPublica. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- ASA and CAP News. (2019). Harnessing new technology to tackle irresponsible gambling ads targeted at children. <https://www.asa.org.uk/news/harnessing-new-technology-gambling-ads-children.html>
- Asplund, J., Eslami, M., Sundaram, H., Sandvig, C., & Karahalios, K. (2020). Auditing Race and Gender Discrimination in Online Housing Markets. *Proceedings of the International AAAI Conference on Web and Social Media*, 14, 24–35. <https://doi.org/10.1609/icwsm.v14i1.7276>
- Auret, D., & Barrientos, S. (2004). Participatory social auditing: A practical guide to developing a gender-sensitive approach. IDS Working Papers, 237. Institute of Development Studies. [https://opendocs.ids.ac.uk/articles/report/Participatory\\_social\\_auditing\\_a\\_practical\\_guide\\_to\\_developing\\_a\\_gender-sensitive\\_approach/26480614?file=48231346](https://opendocs.ids.ac.uk/articles/report/Participatory_social_auditing_a_practical_guide_to_developing_a_gender-sensitive_approach/26480614?file=48231346)
- Auret, D., & Barrientos, S. (2006). Participatory social auditing: Developing a worker-focused approach. In S. Barrientos & C. Dolan (Eds.), *Ethical sourcing in the global food system*. Routledge.
- Automated discrimination: Facebook uses gross stereotypes to optimize ad delivery. (n.d.). AlgorithmWatch. <https://algorithmwatch.org/en/automated-discrimination-facebook-google/>
- Avin S. et al. Filling Gaps in Trustworthy Development of AI. *Science* 374, no. 6573 (December 10, 2021): 1327–29, <https://doi.org/10.1126/science.abi7176>
- Bandy, J. (2021). Problematic Machine Behaviour: A Systematic Literature Review of Algorithm Audits. <https://doi.org/10.48550/ARXIV.2102.04256>

- Bandy, J., & Diakopoulos, N. (2020). Auditing News Curation Systems: A Case Study Examining Algorithmic and Editorial Logic in Apple News. *Proceedings of the International AAAI Conference on Web and Social Media*, 14, 36–47. <https://ojs.aaai.org/index.php/ICWSM/article/view/7277>
- Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., Horowitz, A., Lum, K., & Venkatasubramanian, S. (2021). It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. <https://doi.org/10.48550/ARXIV.2106.05498>
- Barlas, P., Kyriakou, K., Kleanthous, S., & Otterbacher, J. (2019). Social B(eye)as: Human and Machine Descriptions of People Images. *Proceedings of the International AAAI Conference on Web and Social Media*, 13, 583–591. <https://doi.org/10.1609/icwsm.v13i01.3255>
- Barocas, S., Hood, S., & Ziewitz, M. (2013). *Governing Algorithms: A Provocation Piece*. SSRN Electronic Journal. <https://doi.org/10.2139/ssrn.2245322>
- Bashir, M. A., & Wilson, C. (2018). Diffusion of User Tracking Data in the Online Advertising Ecosystem. *Proceedings on Privacy Enhancing Technologies*, 2018(4), 85–103. <https://doi.org/10.1515/popets-2018-0033>
- Bashir, M. A., Arshad, S., & Wilson, C. (2016). "Recommended For You": A First Look at Content Recommendation Networks. *Proceedings of the 2016 Internet Measurement Conference*, 17–24. <https://doi.org/10.1145/2987443.2987469>
- Bashir, M. A., Arshad, S., Robertson, W., & Wilson, C. (n.d.). *Tracing Information Flows Between Ad Exchanges Using Retargeted Ads*. <https://personalization.ccs.neu.edu/Projects/Retargeting/>
- Bashir, M. A., Farooq, U., Shahid, M., Zaffar, M. F., & Wilson, C. (2019). Quantity vs. Quality: Evaluating User Interest Profiles Using Ad Preference Managers. *Proceedings 2019 Network and Distributed System Security Symposium*. Network and Distributed System Security Symposium, San Diego, CA. <https://doi.org/10.14722/ndss.2019.23392>
- Bechmann, A., & Nielbo, K. L. (2018). Are We Exposed to the Same "News" in the News Feed? *Digital Journalism*, 6(8), 990–1002. <https://doi.org/10.1080/21670811.2018.1510741>
- Brown, M. A., Bisbee, J., Lai, A., Bonneau, R., Nagler, J., & Tucker, J. A. (2022). Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users (SSRN Scholarly Paper No. 4114905). <https://doi.org/10.2139/ssrn.4114905>
- Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, 81, 1–15.

- Cabañas, J.G., Cuevas, Á., & Rumín, R.C. (2018). Unveiling and Quantifying Facebook Exploitation of Sensitive Personal Data for Advertising Purposes. USENIX Security Symposium.
- Cano-Orón, L. (2019). Dr. Google, what can you tell me about homeopathy? Comparative study of the top10 websites in the United States, United Kingdom, France, Mexico and Spain. *El Profesional de La Información*, 28(2). <https://doi.org/10.3145/epi.2019.mar.13>
- Chakraborty, A., & Ganguly, N. (2018). Analyzing the News Coverage of Personalized Newspapers. 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 540–543. <https://doi.org/10.1109/ASONAM.2018.8508812>
- Chen, L., Ma, R., Hannák, A., & Wilson, C. (2018). Investigating the Impact of Gender on Rank in Resume Search Engines. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–14. <https://doi.org/10.1145/3173574.3174225>
- Chen, L., Mislove, A., & Wilson, C. (2015). Peeking Beneath the Hood of Uber. Proceedings of the 2015 Internet Measurement Conference, 495–508. <https://doi.org/10.1145/2815675.2815681>
- Chen, L., Mislove, A., & Wilson, C. (2016). An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace. Proceedings of the 25th International Conference on World Wide Web, 1339–1349. <https://doi.org/10.1145/2872427.2883089>
- Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. Conference on Fairness, Accountability and Transparency, 134–148.
- Christl, W. (2022). Digital Profiling in the Online Gambling Industry. A report on marketing and risk surveillance by the UK gambling firm Sky Betting and Gaming, TransUnion, Adobe, Google, Facebook, Microsoft and other data companies. <https://crackedlabs.org/en/gambling-data>
- Competition and Markets Authority. (n.d.). CMA Digital Comparison Tools (DCT) Mystery Shopping Research. Technical Report. <https://assets.publishing.service.gov.uk/media/59c9380e40f0b6440a8b5310/gf-k-mystery-shopping-research-technical-report.pdf>
- Courtois, C., Slechten, L., & Coenen, L. (2018). Challenging Google Search filter bubbles in social and political information: Disconforming evidence from a digital methods case study. *Telematics and Informatics*, 35(7), 2006–2015. <https://doi.org/10.1016/j.tele.2018.07.004>
- Cucchiatti, F., Moll, J., Esteban, M., Reyes, P., & García Calatrava, C. (n.d.). carbolytics, an analysis of the carbon costs of online tracking. <https://carbolytics.org/report.html>

- Dash, A., Chakraborty, A., Ghosh, S., Mukherjee, A., & Gummadi, K. P. (2021). When the Umpire is also a Player: Bias in Private Label Product Recommendations on E-commerce Marketplaces (arXiv:2102.00141). arXiv. <http://arxiv.org/abs/2102.00141>
- Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies*, 2015(1), 92–112. <https://doi.org/10.1515/popets-2015-0007>
- Deng, W. H., Lam, M., Lee, M. K., & Zhu, T. (2023). Understanding practices, challenges, and opportunities for user-engaged algorithm auditing in industry practice. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, 1–18. Association for Computing Machinery. <https://doi.org/10.1145/3544548.3581026>
- Desmarais, S., Johnson, K., & Singh, J. (2016). Performance of Recidivism Risk Assessment Instruments in U.S. Correctional Settings. *Psychological Services*, 13. <https://doi.org/10.1037/ser0000075>
- DeVries, T., Misra, I., Wang, C., & van der Maaten, L. (2019). Does Object Recognition Work for Everyone? (arXiv:1906.02659). arXiv. <http://arxiv.org/abs/1906.02659>
- Diakopoulos, N. (2014). Algorithmic Accountability Reporting: On the Investigation of Black Boxes. <https://doi.org/10.7916/D8ZK5TW2>
- Duwe, G. (2019). Better Practices in the Development and Validation of Recidivism Risk Assessments: The Minnesota Sex Offender Screening Tool–4. *Criminal Justice Policy Review*, 30(4), 538–564. <https://doi.org/10.1177/0887403417718608>
- Duwe, G., & Kim, K. (2017). Out With the Old and in With the New? An Empirical Comparison of Supervised Learning Algorithms to Predict Recidivism. *Criminal Justice Policy Review*, 28(6), 570–600. <https://doi.org/10.1177/0887403415604899>
- Edelman, B. G., & Luca, M. (2014). Digital Discrimination: The Case of Airbnb.com. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2377353>
- Epstein, R., Robertson, R. E., Lazer, D., & Wilson, C. (2017). Suppressing the Search Engine Manipulation Effect (SEME). *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW), 1–22. <https://doi.org/10.1145/3134677>
- Eriksson, M. C., & Johansson, A. (2017). Tracking Gendered Streams. *Culture Unbound*, 9(2), 163–183. <https://doi.org/10.3384/cu.2000.1525.1792163>
- Eslami, M., Aleyasen, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). FeedVis: A Path for Exploring News Feed Curation Algorithms. *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, 65–68. <https://doi.org/10.1145/2685553.2702690>

- Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). "I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in the news feed.
- Eslami, M., Vaccaro, K., Karahalios, K., & Hamilton, K. (2017). "Be Careful; Things Can Be Worse than They Appear": Understanding Biased Algorithms and Users' Behaviour Around Them in Rating Platforms. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), 62–71. <https://doi.org/10.1609/icwsm.v11i1.14898>
- Eticas (2021). *Guide to Algorithmic Auditing*. Eticas Research and Innovation.
- Eticas (2024). *Automating (In) Justice? An Adversarial Audit of RisCanvi*. Eticas Research and Innovation. <https://eticasfoundation.org/automating-injustice-an-adversarial-audit-of-riscanvi/>
- Eticas (2023). *Invisible No More: The impact of facial recognition and Price discrimination AI on people with disabilities*. Eticas Research and Innovation. <https://eticasfoundation.org/invisible-no-more-the-impact-of-facial-recognition-on-people-with-disabilities/>
- Eticas (2023). *Adversarial Algorithmic Auditing Guide*. Eticas Research and Innovation.
- Eticas (2023). *Auditing Social Media: (In)visibility of Political Content on Migration*. Eticas Research and Innovation.
- Eticas (2023). *Auditing TikTok. Social media's treatment of migrants*. Eticas Research and Innovation. <https://eticasfoundation.org/auditing-tiktok-social-medias-treatment-of-migrants/>
- Eticas (2023). *Auditing YouTube. Social media's treatment of migrants and migration*. Eticas Research and Innovation. <https://eticasfoundation.org/auditing-youtube-social-medias-treatment-of-migrants/>
- Eticas, the Taxi Project, & Observatorio TAS. (2023). *Adversarial audit of ride-hailing platforms: Algorithmic compliance with competition, labor and consumer law in Spain*. Eticas Research and Innovation. Taxi Project 2.0. Observatorio TAS.
- Eticas. (2022). *The External Audit of the VioGén System*. Eticas Research and Innovation. <https://eticasfoundation.org/the-case-of-viogen-can-ai-solve-gender-violence/>
- Eticas (2021). *Guide to Algorithmic Auditing*. Eticas Research and Innovation.
- Fabris, A., Mishler, A., Gottardi, S., Carletti, M., Daicampi, M., Susto, G. A., & Silvello, G. (2021). *Algorithmic Audit of Italian Car Insurance: Evidence of Unfairness in Access and Pricing*. ArXiv:2105.10174 [Cs]. <http://arxiv.org/abs/2105.10174>
- Gelauff, L., Goel, A., Munagala, K., & Yandamuri, S. (2020). *Advertising for Demographically Fair Outcomes*. ArXiv:2006.03983 [Cs]. <http://arxiv.org/abs/2006.03983>

- Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019). Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2221–2231. <https://doi.org/10.1145/3292500.3330691>
- Hannak, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., & Wilson, C. (2013). Measuring personalization of web search. Proceedings of the 22nd International Conference on World Wide Web, 527–538. <https://doi.org/10.1145/2488388.2488435>
- Hannak, A., Soeller, G., Lazer, D., Mislove, A., & Wilson, C. (2014). Measuring Price Discrimination and Steering on E-commerce Web Sites. Proceedings of the 2014 Conference on Internet Measurement Conference, 305–318. <https://doi.org/10.1145/2663716.2663744>
- Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., & Wilson, C. (2017). Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 1914–1933. <https://doi.org/10.1145/2998181.2998327>
- How Facebook's ad targeting may be in breach of UK equality and data protection laws. (n.d.). Global Witness. <https://en/campaigns/digital-threats/how-facebooks-ad-targeting-may-be-in-breach-of-uk-equality-and-data-protection-laws/>
- Hu, D., Jiang, S., E. Robertson, R., & Wilson, C. (2019). Auditing the Partisanship of Google Search Snippets. The World Wide Web Conference on - WWW '19, 693–704. <https://doi.org/10.1145/3308558.3313654>
- Hupperich, T., Tatang, D., Wilkop, N., & Holz, T. (2018). An Empirical Study on Online Price Differentiation. Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy, 76–83. <https://doi.org/10.1145/3176258.3176338>
- Hussein, E., Juneja, P., & Mitra, T. (2020). Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW1), 048:1-048:27. <https://doi.org/10.1145/3392854>
- Imana, B., Korolova, A., & Heidemann, J. (2021). Auditing for Discrimination in Algorithms Delivering Job Ads. Proceedings of the Web Conference 2021, 3767–3778. <https://doi.org/10.1145/3442381.3450077>
- Iqbal, U., Bahrami, P. N., Trimananda, R., Cui, H., Gamero-Garrido, A., Dubois, D., Choffnes, D., Markopoulou, A., Roesner, F., & Shafiq, Z. (2022). Your Echos are Heard: Tracking, Profiling, and Ad Targeting in the Amazon Smart Speaker Ecosystem. ArXiv:2204.10920 [Cs]. <http://arxiv.org/abs/2204.10920>
- Jeffries, A., & Yin, L. (n.d.). Amazon Puts Its Own "Brands" First Above Better-Rated Products – The Markup. Retrieved November 30, 2021, from <https://themarkup.org/amazons-advantage/2021/10/14/amazon-puts-its-own-brands-first-above-better-rated-products>

- Jiang, S., Chen, L., Mislove, A., & Wilson, C. (2018). On Ridesharing Competition and Accessibility: Evidence from Uber, Lyft, and Taxi. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 863–872. <https://doi.org/10.1145/3178876.3186134>
- Jiang, S., Robertson, R. E., & Wilson, C. (2019). Bias Misperceived: The Role of Partisanship and Misinformation in YouTube Comment Moderation. *Proceedings of the International AAAI Conference on Web and Social Media*, 13, 278–289. <https://ojs.aaai.org/index.php/ICWSM/article/view/3229>
- Juneja, P., & Mitra, T. (2021). Auditing E-Commerce Platforms for Algorithmically Curated Vaccine Misinformation. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–27. <https://doi.org/10.1145/3411764.3445250>
- Kawakami, A., Umarova, K., & Mustafaraj, E. (2020). The Media Coverage of the 2020 US Presidential Election Candidates through the Lens of Google's Top Stories. *Proceedings of the International AAAI Conference on Web and Social Media*, 14, 868–877. <https://doi.org/10.1609/icwsml.v14i1.7352>
- Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3819–3828. <https://doi.org/10.1145/2702123.2702520>
- Kenway, J., François, C., Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Bug Bounties for Algorithmic Harms?
- Kliman-Silver, C., Hannak, A., Lazer, D., Wilson, C., & Mislove, A. (2015). Location, Location, Location: The Impact of Geolocation on Web Search Personalization. *Proceedings of the 2015 Internet Measurement Conference*, 121–127. <https://doi.org/10.1145/2815675.2815714>
- Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, 13(3), 795–848. <https://doi.org/10.1007/s40685-020-00134-w>
- Koshiyama A. et al. Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms. SSRN Scholarly Paper (Rochester, NY, January 1, 2021), <https://doi.org/10.2139/ssrn.3778998>.
- Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., & Karahalios, K. (2017). Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 417–432. <https://doi.org/10.1145/2998181.2998321>
- Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., & Karahalios, K. (2019). Search bias quantification: investigating political bias in social media and

- web search. *Information Retrieval Journal*, 22(1–2), 188–227. <https://doi.org/10.1007/s10791-018-9341-2>
- Kulynych, B. (2021). bogdan-kulynych/saliency\_bias. [https://github.com/bogdan-kulynych/saliency\\_bias](https://github.com/bogdan-kulynych/saliency_bias)
- Kyriakou, K., Barlas, P., Kleanthous, S., & Otterbacher, J. (2019). Fairness in Proprietary Image Tagging Algorithms: A Cross-Platform Audit on People Images. *Proceedings of the International AAAI Conference on Web and Social Media*, 13, 313–322. <https://doi.org/10.1609/icwsm.v13i01.3232>
- Lai, C., & Luczak-Roesch, M. (2019). You can't see what you can't see: Experimental evidence for how much relevant information may be missed due to Google's Web search personalisation (arXiv:1904.13022). arXiv. <http://arxiv.org/abs/1904.13022>
- Lambrecht, A., & Tucker, C. (2019). Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science*, 65(7), 2966–2981. <https://doi.org/10.1287/mnsc.2018.3093>
- Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (n.d.). How We Analyzed the COMPAS Recidivism Algorithm. ProPublica. Retrieved August 17, 2021, from <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm?token=UAmqdwJgjl-rDCbXUOl5ZnMFbqkg5b6w>
- Lecuyer, M., Ducoffe, G., Lan, F., Papancea, A., Petsios, T., Spahn, R., Chaintreau, A., & Geambasu, R. (2014). XRay: Enhancing the Web's Transparency with Differential Correlation. <https://doi.org/10.48550/ARXIV.1407.2323>
- Ledford, H. (2019). Millions of black people affected by racial bias in health-care algorithms. *Nature*, 574(7780), 608–609. <https://doi.org/10.1038/d41586-019-03228-6>
- Li, D., Liu, X., Zhao, Z., Li, J., & Sun, M. (2025). From generation to judgment: Opportunities and challenges of LLM-as-a-judge (arXiv:2411.16594). arXiv. <https://arxiv.org/abs/2411.16594>
- Lurie, E., & Mustafaraj, E. (2019). Opening Up the Black Box: Auditing Google's Top Stories Algorithm. The Florida AI Research Society.
- Mähler, R., & Vonderau, P. (2017). Studying Ad Targeting with Digital Methods: The Case of Spotify. *Culture Unbound*, 9(2), 212–221. <https://doi.org/10.3384/cu.2000.1525.1792212>
- Matias, J. N., Hounsel, A., & Feamster, N. (2021). Software-Supported Audits of Decision-Making Systems: Testing Google and Facebook's Political Advertising Policies. ArXiv:2103.00064 [Cs]. <http://arxiv.org/abs/2103.00064>
- Matthews, J., Babaeianjelodar, M., Lorenz, S., Matthews, A., Njie, M., Adams, N., Krane, D., Goldthwaite, J., & Hughes, C. (2019). The Right To Confront Your Accusers: Opening the Black Box of Forensic DNA Software. *Proceedings of the 2019 AAAI/ACM*

Conference on AI, Ethics, and Society, 321–327.  
<https://doi.org/10.1145/3306618.3314279>

McMahon, C., Johnson, I., & Hecht, B. (2017). The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), 142–151.  
<https://doi.org/10.1609/icwsm.v11i1.14883>

Mikians, J., Gyarmati, L., Erramilli, V., & Laoutaris, N. (2012). Detecting price and search discrimination on the internet. *Proceedings of the 11th ACM Workshop on Hot Topics in Networks - HotNets-XI*, 79–84. <https://doi.org/10.1145/2390231.2390245>

Mikians, J., Gyarmati, L., Erramilli, V., & Laoutaris, N. (2013). Crowd-assisted search for price discrimination in e-commerce: first results. *Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies*, 1–6.  
<https://doi.org/10.1145/2535372.2535415>

Miles Brundage et al., *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*. <https://doi.org/10.48550/arXiv.2004.07213>

Minderoo Centre For Technology And Democracy. (2022). *A Sociotechnical Audit: Assessing Police Use of Facial Recognition*. Apollo - University of Cambridge Repository. <https://doi.org/10.17863/CAM.89953>

Moe, H. (2019). Comparing Platform "Ranking Cultures" Across Languages: The Case of Islam on YouTube in Scandinavia. *Social Media + Society*, 5(1), 205630511881703.  
<https://doi.org/10.1177/2056305118817038>

Mökander, J and Floridi L. Operationalising AI Governance through Ethics-Based Auditing: An Industry Case Study, *AI and Ethics* 3, no. 2 (May 1, 2023): 451–68, <https://doi.org/10.1007/s43681-022-00171-7>

Mökander, J and Floridi L. Ethics-Based Auditing to Develop Trustworthy AI, *Minds and Machines* 31, no. 2 (June 1, 2021): 323–27, <https://doi.org/10.1007/s11023-021-09557-8>

Jakob Mökander et al., "Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations," *Science and Engineering Ethics* 27, no. 4 (July 6, 2021): 44, <https://doi.org/10.1007/s11948-021-00319-4>

Mökander, J et al. Auditing Large Language Models: A Three-Layered Approach, SSRN Scholarly Paper (Rochester, NY, February 16, 2023), <https://doi.org/10.2139/ssrn.4361607>

Noble, S. U. (2013). *Google Search: Hyper-visibility as a Means of Rendering Black Women and Girls Invisible*.

- Northcutt, C. G., Athalye, A., & Mueller, J. (2021). Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. ArXiv:2103.14749 [Cs, Stat]. <http://arxiv.org/abs/2103.14749>
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>
- Pandey, A., & Caliskan, A. (2021). Disparate Impact of Artificial Intelligence Bias in Ridehailing Economy's Price Discrimination Algorithms. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 822–833. <https://doi.org/10.1145/3461702.3462561>
- Raji, I. D., & Buolamwini, J. (2019). Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435. <https://doi.org/10.1145/3306618.3314244>
- Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2020). Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 131–141. <https://doi.org/10.1145/3351095.3372879>
- Rieder, B., Matamoros-Fernández, A., & Coromina, Ò. (2018). From ranking algorithms to 'ranking cultures': Investigating the modulation of visibility in YouTube search results. *Convergence*, 24(1), 50–68. <https://doi.org/10.1177/1354856517736982>
- Robertson, R. E., Jiang, S., Joseph, K., Friedland, L., Lazer, D., & Wilson, C. (2018). Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 1–22. <https://doi.org/10.1145/3274417>
- Robertson, R. E., Jiang, S., Lazer, D., & Wilson, C. (2019). Auditing Autocomplete: Suggestion Networks and Recursive Algorithm Interrogation. *Proceedings of the 10th ACM Conference on Web Science*, 235–244. <https://doi.org/10.1145/3292522.3326047>
- Robertson, R. E., Lazer, D., & Wilson, C. (2018). Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 955–965. <https://doi.org/10.1145/3178876.3186143>
- Robertson, R. E., Olteanu, A., Diaz, F., Shokouhi, M., & Bailey, P. (2021). "I Can't Reply with That": Characterizing Problematic Email Reply Suggestions. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–18. <https://doi.org/10.1145/3411764.3445557>
- Ryan C. LaBrie and G. Steinke, Towards a Framework for Ethical Audits of AI Algorithms, 2019, <https://www.semanticscholar.org/paper/Towards-a-Framework-for-Ethical-Audits-of-AI-LaBrie-Steinke/c103601dbf79c05c7f72b865ce05e6f82048c1ca>

- Sánchez-Monedero, J., Dencik, L., & Edwards, L. (2020). What does it mean to "solve" the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 458–468. <https://doi.org/10.1145/3351095.3372849>
- Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms*.
- Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The Risk of Racial Bias in Hate Speech Detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678. <https://doi.org/10.18653/v1/P19-1163>
- Sapiezynski, P., Kassarnig, V., & Wilson, C. (2017). *Academic performance prediction in a gender-imbalanced environment*. Boise State University. <https://doi.org/10.18122/B20Q5R>
- Silva, M., de Oliveira, L. S., Andreou, A., de Melo, P. O. V., Goga, O., & Benevenuto, F. (2020). Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook (arXiv:2001.10581). arXiv. <http://arxiv.org/abs/2001.10581>
- Snickars, P. (2017). More of the Same – On Spotify Radio. *Culture Unbound*, 9(2), 184–211. <https://doi.org/10.3384/cu.2000.1525.1792184>
- Soeller, G., Karahalios, K., Sandvig, C., & Wilson, C. (2016). MapWatch: Detecting and Monitoring International Border Personalization on Online Maps. *Proceedings of the 25th International Conference on World Wide Web*, 867–878. <https://doi.org/10.1145/2872427.2883016>
- Sultan, T. (2020, September 30). The A-levels Exam Fiasco: Ofqual's Discriminatory Algorithm. *Gair Rhydd*. <https://cardiffstudentmedia.co.uk/gairrhydd/the-a-levels-exam-fiasco-ofquals-discriminatory-algorithm/>
- Sweeney, L. (2013). *Discrimination in Online Ad Delivery*. <https://doi.org/10.48550/ARXIV.1301.6822>
- Tolan, S., Miron, M., Gómez, E., & Castillo, C. (2019). Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, 83–92. <https://doi.org/10.1145/3322640.3326705>
- Trielli, D., & Diakopoulos, N. (2019). Search as News Curator: The Role of Google in Shaping Attention to News Information. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15. <https://doi.org/10.1145/3290605.3300683>
- Tschantz, M. C., Egelman, S., Choi, J., Weaver, N., & Friedland, G. (2018). The Accuracy of the Demographic Inferences Shown on Google's Ad Settings. *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, 33–41. <https://doi.org/10.1145/3267323.3268962>

- Turner, E., Medina, J., & Brown, G. (2019). Dashing Hopes? The Predictive Accuracy of Domestic Abuse Risk Assessment by Police. *The British Journal of Criminology*, 59(5), 1013–1034. <https://doi.org/10.1093/bjc/azy074>
- Urman, A., Makhortykh, M., & Ulloa, R. (2021). Auditing Source Diversity Bias in Video Search Results Using Virtual Agents. *Companion Proceedings of the Web Conference 2021*, 232–236. <https://doi.org/10.1145/3442442.3452306>
- Vecchione, B., Levy, K., & Barocas, S. (2021). Algorithmic auditing and social justice: Lessons from the history of audit studies. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21)* (pp. 1–9). Association for Computing Machinery. <https://doi.org/10.1145/3465416.3483294>
- Venkatadri, G., Sapiezynski, P., Redmiles, E. M., Mislove, A., Goga, O., Mazurek, M., & Gummadi, K. P. (2019). Auditing Offline Data Brokers via Facebook's Advertising Platform. *The World Wide Web Conference on - WWW '19*, 1920–1930. <https://doi.org/10.1145/3308558.3313666>
- Vincent, N., Johnson, I., Sheehan, P., & Hecht, B. (2019). Measuring the Importance of User-Generated Content to Search Engines. *Proceedings of the International AAAI Conference on Web and Social Media*, 13, 505–516. <https://doi.org/10.1609/icwsm.v13i01.3248>
- WarTok: TikTok is feeding war disinformation to new users within minutes — even if they don't search for Ukraine-related content. (n.d.). NewsGuard. Retrieved March 25, 2022, from <https://www.newsguardtech.com/misinformation-monitor/march-2022>
- Weber, M. S., & Kosterich, A. (2018). Coding the News: The role of computer code in filtering and distributing news. *Digital Journalism*, 6(3), 310–329. <https://doi.org/10.1080/21670811.2017.1366865>
- Whyte, W. F. (1989). Advancing scientific knowledge through participatory action research. *Sociological Forum*, 4(3), 367–385. <https://doi.org/10.1007/BF01115015>
- Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., Trindel, K., & Polli, F. (2021). Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 666–677. <https://doi.org/10.1145/3442188.3445928>
- YouTube Regrets. (2021). Mozilla Foundation. <https://foundation.mozilla.org/en/campaigns/regrets-reporter/findings/>
- Zeng, A., et al. (2024). Can large language models judge research papers? arXiv preprint arXiv:2411.16594. <https://arxiv.org/abs/2411.16594>