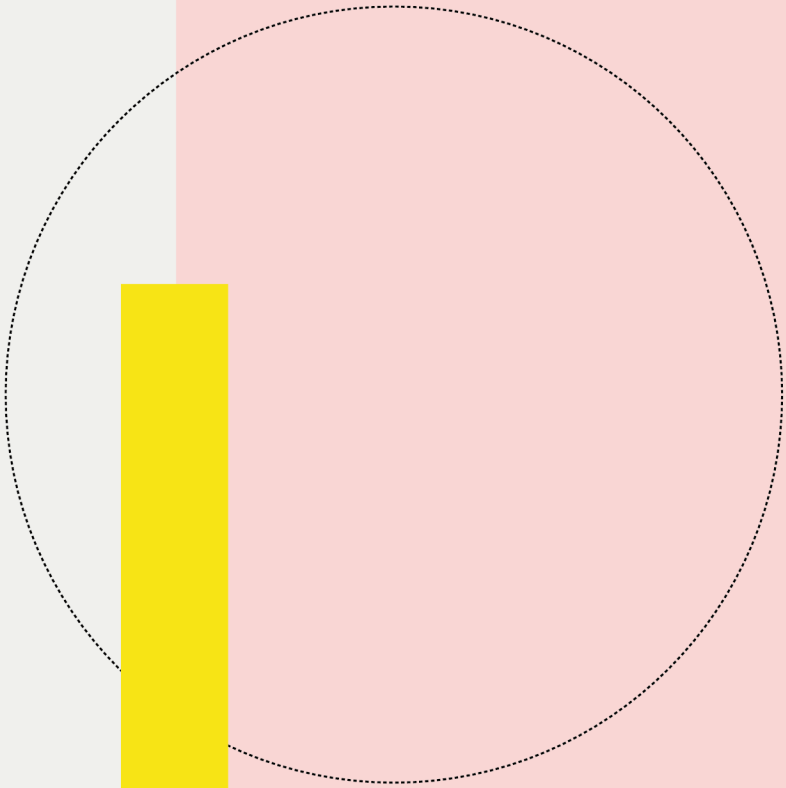
ADVERSARIAL ALGORITHMIC AUDITING GUIDE



eticas
TECH

CONTENT

EXECUTIVE SUMMARY

Executive Summary

As algorithms and AI systems proliferate worldwide, developers, regulators, communities and users need robust methods to assess their impact. This guide focuses on adversarial or third-party algorithmic auditing of systems with little transparency and oversight. The purpose of the guide is to provide an overview of the adversarial algorithmic auditing process, including:

What A process by which an independent third party or community examines the impact and, to the extent possible, the functioning of an algorithmic system in order to detect potential anomalies or practices that could be unfair or harmful towards protected groups or society as a whole.

Why Opaque algorithmic systems influence virtually every aspect of life. While internal socio-technical audits can address this to an extent, bias dynamics may not be evident before they translate into impacts. Adversarial algorithmic audits provide a way to examine this impact independently from the outside and offer a concrete toolset to quantify impacts and propose mitigation measures.

When Post-processing stage of the AI system lifecycle

Who Auditors including social science researchers, journalists, data scientists, members of civil society organizations, members of affected communities and end users.

How An adversarial audit must be a transparent and public process employing a robust socio-technical approach and specialized methods including:

- Open-source code audit
- Scraping audit
- Sock puppet audit
- Crowdsourcing
- Experimental user audit
- Comparative output audit
- Ethnographic audit

This guide outlines a set of steps and a methodology for adversarial algorithmic auditing which turn principles into a robust toolset and an effective mechanism for AI inspection and accountability. The Introduction section explains the concept of algorithmic auditing and why adversarial audits are necessary. The How-to section describes the steps for conducting adversarial audits. The Methods section covers several methodological techniques for conducting adversarial audits. The Case Studies section provides examples of Eticas' adversarial audits conducted on various systems. Finally, the guide includes an Audit Report Index, which serves as a helpful tool for organizing and structuring the results of the audit report.

INTRODUCTION

Introduction

Algorithmic auditing is an instrument for dynamic appraisal and inspection of AI systems. This guide focuses on adversarial or third-party algorithmic audits, where independent auditors or communities thoroughly examine the functioning and impact of an algorithmic system, when access to the system is restricted.

Adversarial algorithmic auditing offers a toolset for evaluating algorithms and AI systems in situations where transparent oversight is limited. Adversarial algorithmic audits are a means to enhance AI transparency and accountability, thereby bridging the gap between innovation potential and societal impact.

This guide includes actionable guidelines for conducting adversarial audits. The guide is addressed to social science researchers, journalists, data scientists, members of civil society organizations, 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 auditors to uncover the potentially negative impacts of algorithms and AI through a set of rigorous steps and methods.

This guide to adversarial algorithmic auditing consolidates the knowledge and experience of Eticas in conducting algorithmic audits. 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 adversarial (third-party) socio-technical algorithmic audits of risk assessment tools, social media, facial recognition, consumer platforms and other systems. In addition to our own experience, this guide is informed by an extensive review of previous adversarial audits, and it summarizes the best practices for adversarial algorithmic auditing.

What is algorithmic auditing?

Algorithmic auditing is a way to inspect AI systems in their specific contexts (Eticas, 2023). It is an approach and methodology that allows for a dynamic appraisal of regulations, standards, and impact. If its results are public, it is also a tool for improved transparency and accountability.

Algorithmic audits can inspect and evaluate entire AI systems.[1] An AI system can include more than one algorithm or model. Depending on their scope,

[1] 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.

algorithmic audits can inspect and evaluate one or more algorithms within an AI system.[2]

Algorithmic audits can be broadly classified into two types depending on the auditors' distance from the developers or implementers of an algorithm: internal and adversarial algorithmic audits.

An internal algorithmic audit, also known as a second-party audit, is conducted by independent auditors 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.

An adversarial algorithmic audit, also known as a third-party 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 goal of adversarial algorithmic auditing is to detect biases, inefficiencies, anomalies or other practices that could be unfair or harmful towards protected groups or society as a whole.

The key distinguishing feature between adversarial audits and internal socio-technical audits is the restricted access to the algorithm and its associated databases used for design, development, testing and validation. For this reason, adversarial 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 internal socio-technical audits that could encompass the entire end-to-end process.

Due to the limited access to internal data and information about an algorithm, third-party adversarial 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.

Why adversarial audits?

Internal socio-technical algorithmic audits are conducted in collaboration with the developers or implementers of an algorithm who are either willing or required to undergo an independent evaluation. However, organizations which develop or implement AI systems are not always willing to be audited, while the regulatory requirements for algorithmic audits remain nascent despite recent policy developments such as the Digital Services Act (DSA), which

[2] 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. 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.

requires independent audits of very large online platforms and search engines to ensure their accountability.

Adversarial audits are our contribution to building concrete tools to reconcile innovation potential with societal impact. They provide the means to evaluate systems typically out of reach and hence with little potential for transparent oversight. The approach towards adversarial algorithmic auditing proposed in this guide is systematic but agile. This makes the process rigorous and verifiable on the one hand, and versatile in its ability to adapt to different types of AI systems on the other. In recognition that algorithmic processes are both informed by and able to impact social dynamics, adversarial algorithmic audits follow a socio-technical approach, including:

- Qualitative contextual analysis and stakeholder mapping as first steps towards understanding the algorithm and the environment where it operates
- Evaluation of bias, inefficiencies and other anomalies through data, network interactions and impacts
- Replicating and reverse engineering system processes through research of affected parties

HOW TO AUDIT ALGORITHMS

How to audit algorithms

Steps for conducting adversarial audits

This section describes the steps for conducting adversarial algorithmic audits. 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.

Planning

Choosing a system to audit	Selecting an AI system with social impact and an initial feasibility check for identifying possible access points to the algorithm(s) for an audit
Contextual analysis	Building understanding about the AI system, the context in which it operates and the possible negative impacts it may lead to
Stakeholder mapping	Identifying relevant parties to an AI system including but not limited to the developers and implementers of the system and the communities affected directly or indirectly by it

Feasibility assessment	Data mapping to determine if the auditor can obtain sufficient information about an AI system via legal means within the relevant jurisdiction
Alliance building	Establishing relationships with communities and civil society organizations to ensure that the perspectives of affected groups are incorporated in the auditing process
Methodology design	Defining the scope of the audit, the research questions, the methods to investigate them, and the timeline of the project
	Execution
Data collection	Data gathering about the inputs, outputs and societal impact of an AI system via specialized techniques for adversarial algorithmic auditing and social science research methods
Data analysis	Translating raw data into meaningful insights via quantitative and qualitative data analysis

Mitigation and recommendations	Providing actionable mitigation measures for the developers or the implementers of an AI system and recommendations for regulators empowering them to seek accountability effectively

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 organized. This involves the following steps: choosing a system to audit, contextual analysis, stakeholder mapping, feasibility assessment, alliance building, and methodology design.

1. Choosing a system to audit

The first step in the planning for an adversarial algorithmic audit is to choose an AI system (algorithm) to audit. This entails not only identifying the algorithm, but also considering:

- Its potential impact on a community or society as a whole, including harms and inefficiencies.
- The possibilities of accessing (a part) of the AI system as an external auditor.

Previous academic research, the work of civil society organizations 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 specialized 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.

In some instances, choosing an algorithm for an adversarial audit can occur through existing relationships and partnerships e.g., when civil society contacts auditors with a specific concern. In such cases, the auditors should conduct a preliminary feasibility assessment and decide on the best way to approach the audit.

Once an algorithm of interest is identified, independent auditors should consider the following questions before proceeding to the next steps in the adversarial audit process:

- **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 adversarial 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 adversarial 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 adversarial 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.

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

2. Contextual analysis

Contextual analysis enables researchers to start building understanding about both the algorithmic system as well as the legal, social, cultural, political and economic environment in which it operates. 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 system within its broader social, legal and economic context. This includes determining what biases to check for in a system. The table below provides a non-exhaustive list of possible biases. For a more comprehensive

list of sources and moments of bias in AI systems, see Eticas' Guide to Algorithmic Auditing.

Techno-solutionist bias	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.
Population bias	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.
Omitted variable	When one or more important variables are not included in the model, resulting in biased regression coefficients and inaccurate statistical results.
Historical bias	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.
Aggregation bias	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.
Accessibility bias	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 marginalized 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 thinking outside of the box in checking for previously untested or otherwise unexpected biases or inefficiencies. This can entail identifying less common biases, as well as vulnerable groups which have been overlooked in previous research but are likely to be at risk of disparate impact. 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 color, but its performance had not been tested on individuals with disabilities with physical manifestations.

Key questions: 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?

3. Stakeholder mapping

Stakeholder mapping involves identifying the relevant parties to an algorithmic system including but 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 stakeholder groups above may be mutually exclusive: 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. Similarly, the affected communities may (partially) overlap with the users or the target population of an algorithm.

Stakeholder mapping should relate back to the contextual analysis and clarify how different stakeholders are positioned within the environment of operation. In the case of developers and implementers, 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. With regards to disparate impact and other biases, stakeholder mapping is a useful tool to identify which groups may be at risk of bias, discrimination or other harm.

Key questions: 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?

4. Feasibility assessment

One of the most challenging aspects of adversarial algorithmic auditing is the lack of access to internal data about algorithmic systems. 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.

For organizations conducting algorithmic audits, it is recommended to assess the available resources and expertise. 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, an adversarial algorithmic audit of the given AI system may not be feasible at this time.

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

5. Alliance building

As a tool for algorithmic transparency and accountability, adversarial algorithmic auditing aims to empower communities to safeguard their rights in the digital arena and beyond. For this reason, it is critical to incorporate the perspectives of affected groups in the auditing process. Alliance building with communities or civil society organizations representing them enables communication and facilitates trust between the auditors and those at risk of algorithmic harm.

Alliance building entails mapping and reaching out to members of affected communities and relevant civil society organizations in the field. In cases where access to 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.

Collaborations with civil society and affected communities can take different shapes depending on the nature of the audit and the scope of the partnership. Auditors can partner with civil society organizations and work together to define the scope and the methodology of the audit as well as the desired outcomes and strategies for policy action and advocacy campaigns upon completion of the audit. It is recommended that the obligations of each party are laid out in a contract and proper attribution is given to each depending on their role as partners, facilitators or other. Whenever possible, civil society organizations should be compensated for their time once the terms of the audit have been agreed to by both parties. In all collaborations, the robustness of the audit should guide decisions regarding engagement with partners.

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

6. Methodology design

The methodology design involves defining the scope of the audit, the research questions, the methods to investigate them, and the timeline of the project. Considerations regarding the scope of an audit may entail a focus on a particular group (for example, the portrayal of migrants in social media recommender systems), geographic location or region, and time period (for example, during elections) among other parameters. The research questions should reflect the scope and pose a clearly formulated, verifiable proposition about the presence of bias, harm or inefficiencies within an AI system to guide the audit process.

Adversarial algorithmic auditing combines traditional social science methods from a socio-technical perspective and specialized methods for algorithmic auditing including:

- Open-source code audit
- Scraping audit
- Sock puppet audit
- Crowdsourcing
- Experimental user audit
- Comparative output audit
- Ethnographic audit

The graph below serves as guidance on the selection of the most appropriate auditing method depending on the availability of data. Further considerations regarding the most appropriate method are outlined in the Methods for conducting adversarial audits section.

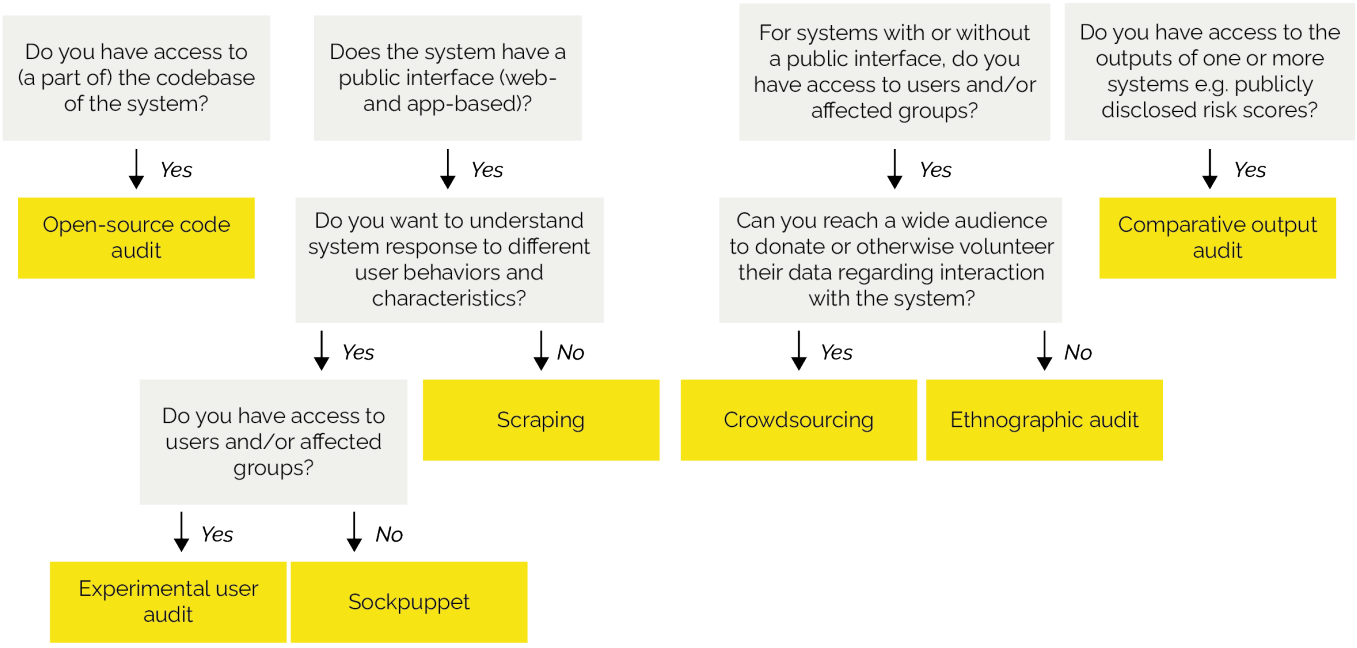


Figure 1. Selection of method for conducting adversarial audits

Key questions: 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?

Execution

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

1. Data collection

The first step in the execution phase of an adversarial algorithmic audit is data collection, which involves gathering information about the inputs, outputs, and societal impact of an algorithmic system. 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). 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.

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 anonymization where possible and secure storage. Additionally, it should communicate, risks (if any) and conditions of participation.

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

2. Data analysis

This step includes technical analysis to identify statistical bias, inaccuracies or inefficiencies in the quantitative data gathered in the previous step. It also entails qualitative analysis informed by the literature review and fieldwork to

assess impact on vulnerable groups and society as a whole. 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.

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. On the other hand, 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.

Key questions: 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?

3. Mitigation and recommendations

The final step considers the findings from the data collection and analysis within the environmental context of the AI system. It prompts the auditor to 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.

Key questions: 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?

METHODS FOR CONDUCTING ADVERSARIAL AUDITS

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.).

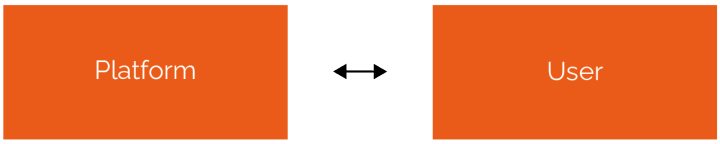


Figure 2. Visualization of the interaction between platforms and users
Source: Sandvig et al.

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 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.

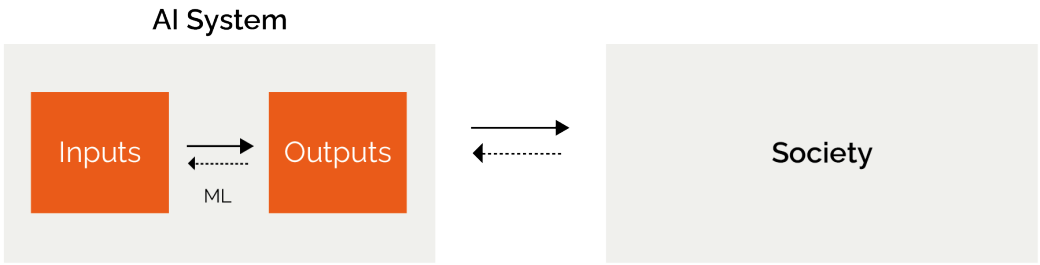


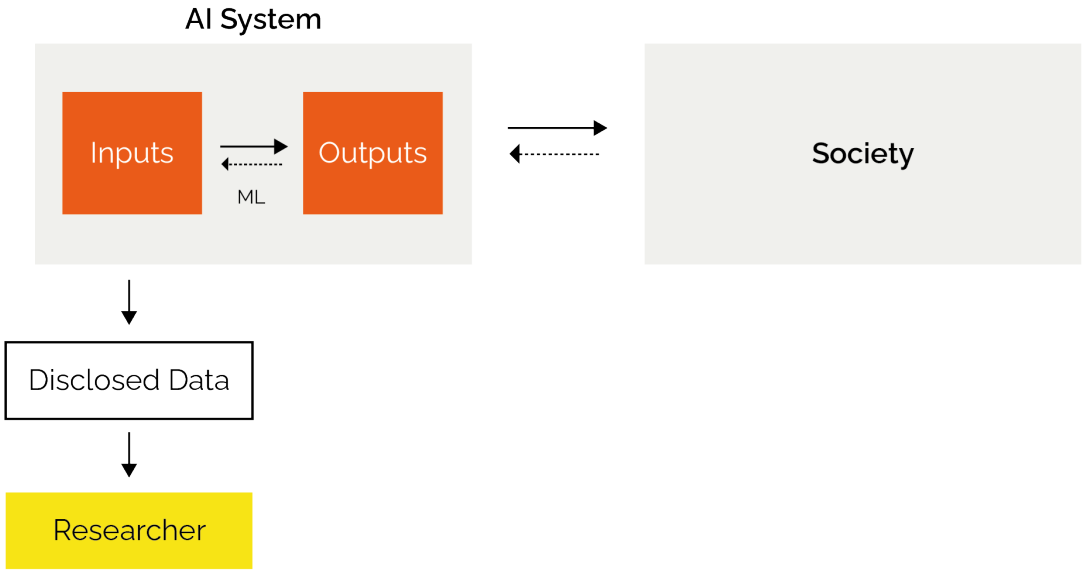
Figure 3. Visualization of the interaction between an algorithmic system and society
Source: Eticas

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.

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 Audit.



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.

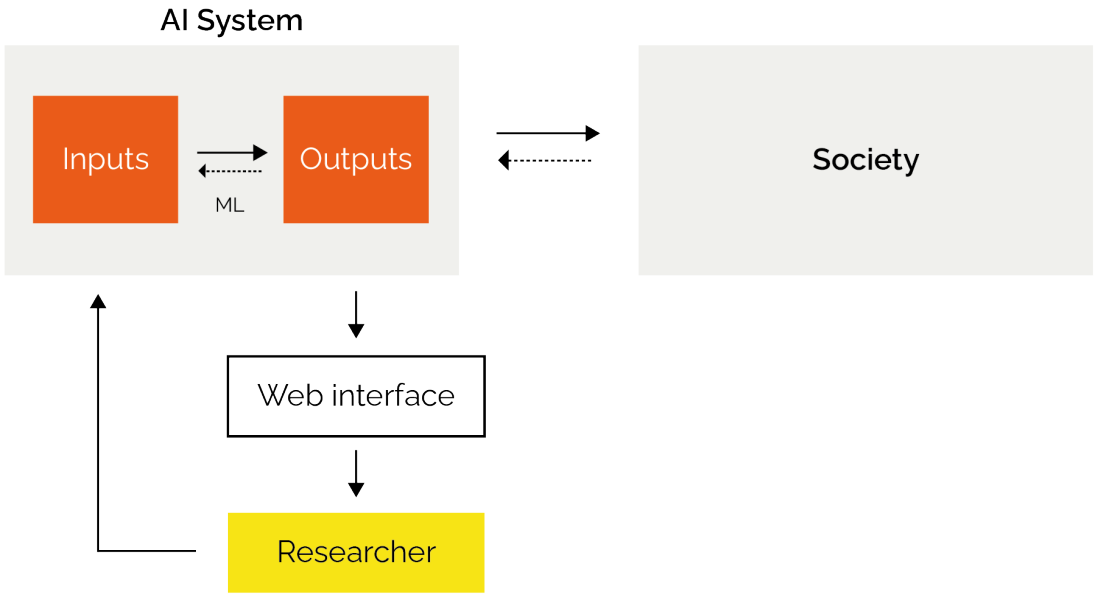
Limitations:

- Problematic machine behaviors 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.
- 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.

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.



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.

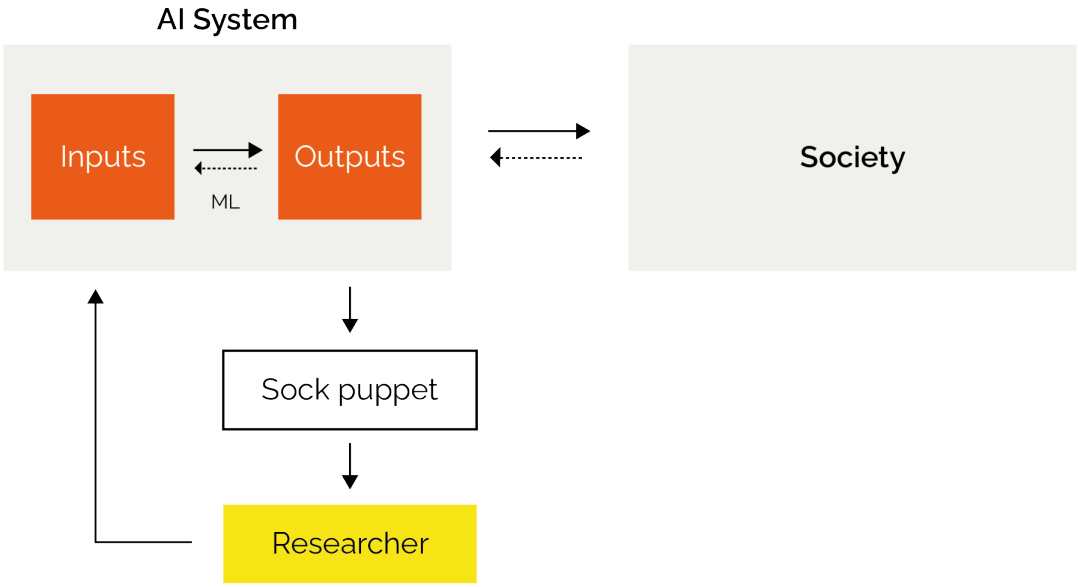
- If automated, scraping can generate a high amount of data for testing and analysis.

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 behavior 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.

Sock puppet

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



When to use this method:

- 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.
- Large-scale audits.

Strengths:

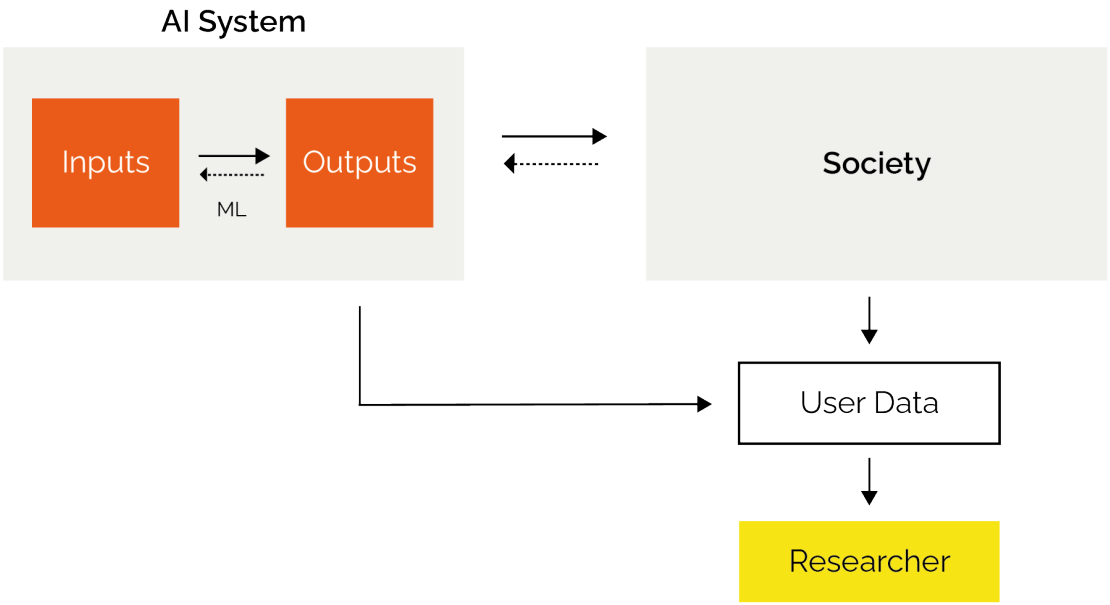
- Effective method to observe the outputs of a system and identify patterns across different conditions, enabling comparison and more effective detection of biases.
- Accessible method to all auditors and communities irrespective of their level of technical expertise (for manual scraping) and available 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 behavior 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 behavior since they lack embedded client-side information such as cookies.

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.



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:

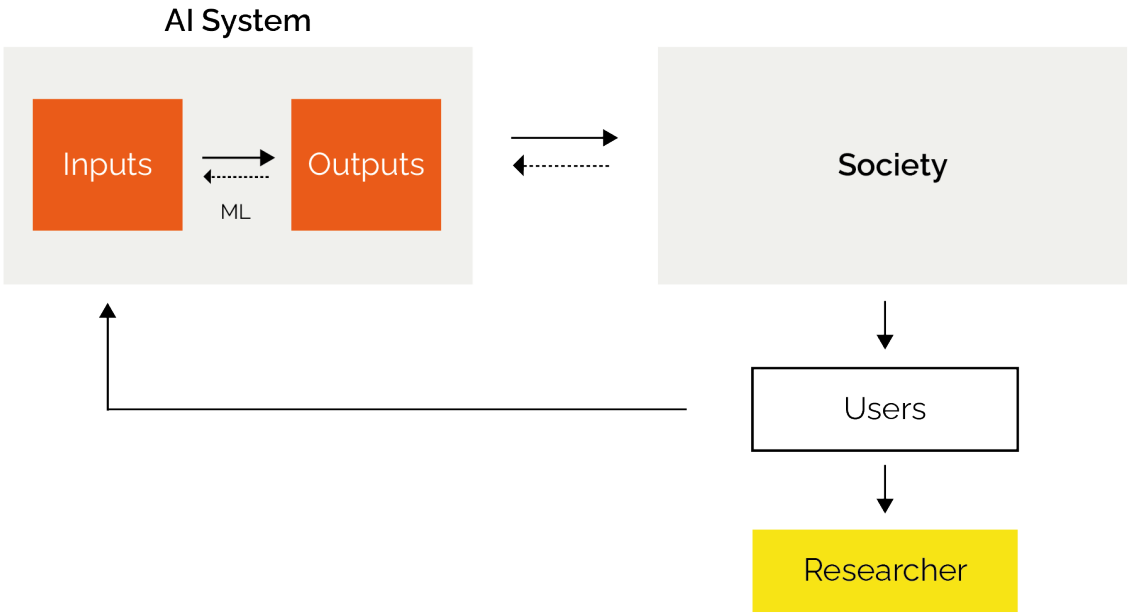
- Reflective of real users' experience and the most accurate approximation between the interaction between an algorithmic system and society (via users).
- Direct involvement of the user community.

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.

Experimental user audit

The experimental user audit is a systematic method for observing and recording system responses to real user behaviors 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).



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 behavior towards characteristics that are difficult to replicate via automated queries, such as the performance of facial recognition on people with disabilities.

Strengths:

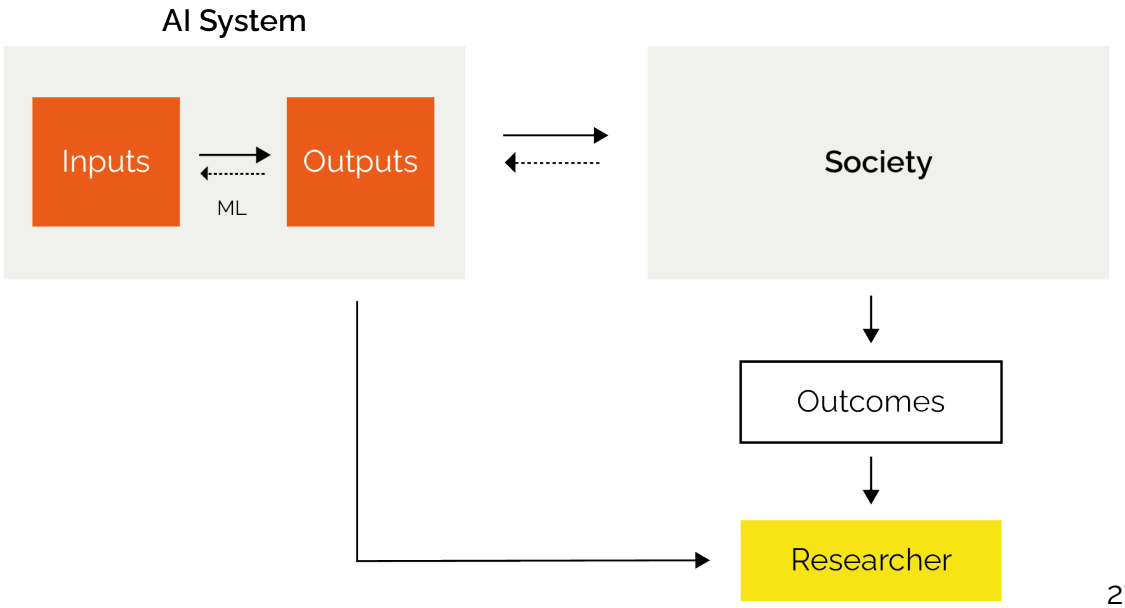
- Like the sock puppet method, experimental user audits are an effective method to observe system outputs and identify patterns across different conditions, facilitating comparison and detection of biases.
- Results from experimental user audits provide closer approximations of real users' interactions 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.

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.



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.

Strengths:

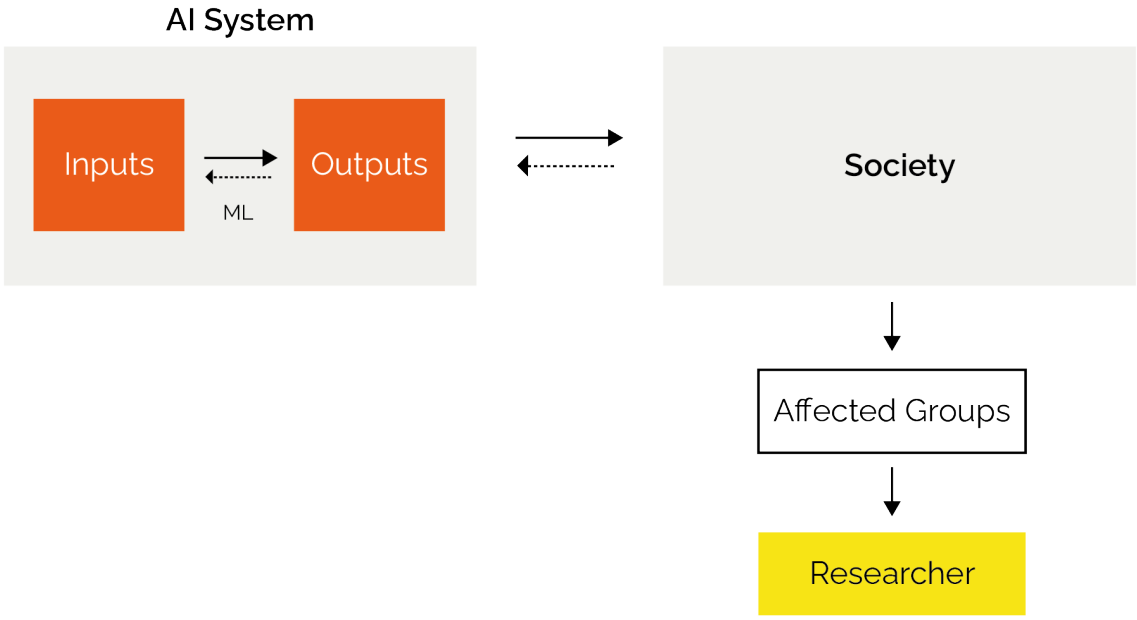
- Based on accurate representation of an algorithm's outputs (e.g., predictions or risk scores) using publicly disclosed data, rather than approximations or subjective user experiences.
- Enables comparison between different systems.

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.

Ethnographic audit

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



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 behavior 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.

CASE STUDIES

Case studies

The Case studies section provides insights on how to approach adversarial audits for different AI systems. It showcases examples of adversarial audits conducted by Eticas, focusing on auditing risk assessment algorithms, social media platforms, facial recognition systems, and consumer platforms. By examining these case studies, readers can gain a deeper understanding of the steps in the auditing process and learn how methods for adversarial auditing can be combined to assess the impact of algorithms in different domains.

Audit	AI system	Domain	Method
VioGén audit	Risk assessment algorithm	Law enforcement	Ethnographic audit; Comparative output audit
YouTube audit	Search algorithm; Recommendation algorithm	Social media and internet platforms	Scraping audit; Sock puppet audit; Ethnographic audit
TikTok audit	Recommendation algorithm	Social media and internet platforms	Scraping audit; Sock puppet audit

Use of facial recognition in insurance audit	Facial recognition; Risk assessment algorithm	Insurance	Experimental user audit
Ride-hailing platform audit	Pricing algorithms in consumer platforms	Sharing and gig economy	Scraping audit; Ethnographic audit

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 marginalized groups.

VioGén is automated risk assessment algorithm used by the public administration in Spain that determines the level of risk faced by a victim of gender-based violence. The system establishes her protection measures. Our adversarial audit of the system demonstrates the effectiveness of adopting a multi-methods approach that combines quantitative analysis of publicly available data (comparative output audit) with qualitative research involving interviews with affected stakeholders and civil society organizations (ethnographic audit). This allows for a comprehensive understanding of the strengths and weaknesses of a risk assessment algorithmic system.

VioGén Audit

1. **Choosing a system to audit**

Eticas selected the VioGén System (The "Integral Monitoring System in Cases of Gender Violence") because it affects vulnerable populations, and automated systems like VioGén often raise concerns about transparency, accountability, and social impact, including the lack of independent oversight and user participation. Gender violence is a very complex social issue, and any automated approach used to address it must be held accountable.

2. **Contextual analysis**

The VioGén System is a web application integrated in the Spanish SARA Network designed to coordinate the actions of public professionals in charge of monitoring, assisting, and protecting women who report gender violence and their children. During our contextual analysis, we identified concerns around transparency, independent oversight, accountability, end-user engagement and the possible transition of the algorithmic system to machine learning. With these concerns in mind, Eticas conducted an adversarial audit.

3. **Stakeholder mapping**

The auditors team identified the following key stakeholder groups:

- Spanish Ministry of the Interior and the Gender Violence Unit
- Ana Bella Foundation
- Researchers and developers who contributed to the development of VioGén
- Software company SAS
- Police officers who use VioGén to assign risk scores
- Women who have survived domestic violence and have had a risk score produced by the VioGén system
- The public, as VioGén is a publicly funded decision-making system of enormous social impact

4. **Feasibility assessment**

Eticas requested information and meetings from the Spanish Ministry of Interior regarding VioGén's functioning and impact. However, the Government did not take action and therefore Eticas could not conduct its proposed internal audit. This situation prompted Eticas to explore the feasibility of conducting an adversarial audit, that is an audit without the cooperation of the Ministry of the Interior. Despite the lack of access to the full original database, Eticas identified a public record of homicide victims, including victims of intimate partner violence, i.e., a subset of the targeted database. Eticas then conducted a partial comparative output audit. Additionally, Eticas partnered with the Ana Bella Foundation to enable an ethnographic audit that included interviews with women victims of gender-based violence.

5. **Alliance building**

For this adversarial audit, Eticas partnered with the Ana Bella Foundation, a leading civil society organization (CSO) working with women who have survived domestic violence and have had a risk score produced by the VioGén system. We established a formal partnership and hired a representative from the Foundation to conduct the interviews due to the sensitivity of the issue.

6. **Methodology design and data collection**

The audit employed a multi-methods research approach. The methodology combined the comparative output audit method with the ethnographic audit method. To compare the risk scores assigned by the algorithm and the actual outcomes for victims of domestic violence across groups of women, we used a public record of Intimate Partner Homicide (IPH) victims, a subset of women victims of domestic violence. For the ethnographic audit, we conducted qualitative fieldwork with 31 women who had gone through the VioGén system. to explore affected communities' perceptions and experiences with the system:

7. **Data analysis**

The statistical analysis of the IPH identified false negative rates and disparities in recall across different strata to contextualize the algorithm's predictive accuracy and potential biases. From a qualitative perspective, we conducted thematic analysis of the survey and interview data. Our analysis showed that VioGén adapts the clustering of risk assessments to limited police resources, with only 1 out of 7 women who reached out to the police for protection receiving it. We found that not having children has a significant negative impact on how extreme risk cases are perceived, and that police officers only increase VioGén's observed risk score in 5% of cases, highlighting a potential bias in the system. We also confirmed our concerns that VioGén lacks transparency, accountability and engagement with end-users.

For recommendations and mitigation measures and more details on our auditing process, see The External Audit of the VioGén System report.

Auditing social media

Our adversarial 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 behaviors 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.

YouTube Audit

Eticas conducted an adversarial audit of YouTube's search and recommendation algorithms to examine how migrants are represented in YouTube content (Eticas, 2023b).

1. **Choosing the system**

For our first audit of social media, we selected YouTube because it is the second most popular social media platform in the world, and it plays a major role in informing people about global issues, including vulnerable social groups, such as migrants. Its visual component is also influential in shaping people's emotional attitudes toward these groups.

2. **Contextual analysis**

Our literature review of harms associated with social media helped us identify concerns that YouTube's algorithms may be exposing users to divisive and potentially harmful content, including misinformation and conspiracy theories. We identified a gap in the existing literature on the topic and decided to focus on YouTube's search and recommendation algorithms with a focus on the portrayal of migration.

3. **Stakeholder mapping**

As social media is a major source of news and entertainment for a large part of the population, the biases and anomalies present in the algorithms employed by internet platforms affect society with far-ranging impacts, concerning stakeholders including:

- YouTube and other social media platforms
- migrant and refugee communities
- migrant led CSOs or other organizations working with individuals from migrant background
- public institutions and policy makers
- researchers and experts
- users of YouTube and other social media platforms

4. **Feasibility assessment**

Our feasibility assessment for the YouTube audit included an assessment of resource availability and in-house technical expertise for data collection, and a legal feasibility assessment. Since YouTube is a public platform available for anyone to use with or without an account, it provides many access points for auditors to observe the outputs of the system. In this case, our feasibility assessment also included outlining and evaluating the strengths and weaknesses of different possible access points.

5. **Alliance building**

The YouTube audit was conducted as a part of the Re:framing Migrants in the European Media pilot project, which is co-funded by the European Union. Eticas has collaborated with members of the consortium and migrant communities to help achieve its goal of changing the media narrative around migrant and refugee communities in Europe.

6. Methodology design

Based on the contextual analysis informed by previous research and the feasibility assessment regarding possible access points to audit the platform, we decided to focus on four specific questions about the representation of migrants on YouTube:

- How are migrants and refugees represented in the top-watched YouTube video search results?
- Does YouTube's search and recommendation algorithms suggest differently framed migration videos in different national settings?
- Does YouTube's search and recommendation algorithms suggest differently framed migration videos to migrant and non-migrant accounts?
- How do individuals with a migrant background perceive the portrayal of migrants on YouTube videos?

In this step, we also determined the most appropriate methods for data collection and analysis to examine those questions. For the first three research questions, we chose to employ scraping and sock puppet audit methods to study migrant and refugee representation in the thumbnails of YouTube videos. Our decision to focus on thumbnails rather than full videos was informed by previous research on the evocative effect of images in the formation of public opinion and it was influenced by considerations regarding the time resources involved in scraping and analyzing full-length videos. For the final research question, we decided to collect qualitative insights and employ the ethnographic audit method to capture the experiences of individuals with migrant backgrounds.

7. Data collection

For RQ1, we scraped the top-watched videos worldwide for "migrants" and "refugees" on YouTube. RQ2 involved scraping the top-100 recommended videos for "migrants" using a VPN to change the location to Canada and the UK. For RQ3, we created two sock puppet (migrant and non-migrant) accounts and scraped the top-100 recommended videos separately. Finally, for RQ4, data was collected through a roundtable discussion with individuals having a migrant background at the "Decolonizing the Newsroom" event in Madrid, where they were shown preliminary results of a content analysis conducted on the top 100 most watched videos worldwide, and their perceptions and reactions were recorded.)

8. Data analysis

We used content analysis to identify the key visual features of each thumbnail, including categories such as predominant gender, group size, activity and facial visibility. We then calculated the frequency distributions of each category across the three datasets we collected for our first three research questions. Our findings revealed that YouTube's portrayal of migrants and refugees in popular YouTube videos is biased and dehumanizing, perpetuating negative stereotypes across national settings and user backgrounds. Our qualitative data reaffirmed the negative impacts of this portrayal as individuals with migrant backgrounds found the narratives promoted by YouTube to be victimizing and limiting their agency.

For recommendations and mitigation measures and more details on our auditing process, see the Auditing Social Media: Portrayal of Migrants on YouTube report.

TikTok Audit

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

1. Choosing the system

We chose to audit TikTok because it is one of the fastest growing and most popular social media platforms, especially among young users. TikTok is especially popular among young users, with increasing impact in the social and political realms. Another factor in our decision was the unique challenges posed by the app, including concerns about data privacy, national security or external influence. Unlike most other large social media platforms, TikTok is not a U.S.-based company.

2. Contextual analysis

Our contextual analysis was primarily informed by a literature review of previous audits and other studies of TikTok which documented the negative impacts of the app's highly compelling recommendation algorithm. These included TikTok leading users to "rabbit holes" of increasingly extreme content on sensitive topics such as depression and suicide and promoting disinformation on political issues within minutes or hours of using the platform. We sought to examine in greater detail how TikTok's recommendation algorithm works across different settings with regards to political discourse on migration.

3. Stakeholder mapping

Similar to our YouTube audit, we identified broad groups of stakeholders including:

- TikTok and other social media platforms
- migrant and refugee communities
- migrant-led CSOs or other organizations working with people from migrant background
- public institutions and policy makers
- researchers and experts
- users of TikTok and other social media platforms

4. Feasibility assessment

Unlike YouTube which is often used in a browser on a computer, TikTok primarily interacts with users via its mobile app. During our feasibility assessment, we noted the limitations with collecting data from the mobile phone app and instead, we identified access points via the browser version of the platform. We also conducted a legal feasibility assessment, noting that TikTok' Terms of Services are more prohibitive than other social media platforms.

5. Alliance building

The TikTok audit report was the second adversarial audit conducted as a part of the Re:framing Migrants in the European Media pilot project, co-funded by the European Union, with the aim of ensuring appropriate media representation of migrant and refugee communities.

6. Methodology design

In our methodology design, we decided to limit the geographic and temporal scope of our audit to the U.S. in the period between October 8, 2022, until December 1, 2022 during the midterm elections. This afforded us the opportunity to track differences in political content recommendations on TikTok before, during and after the election, and it provided an opportunity to study migration as one of the most important voting issues in the election. Within this scope, we formulated three research questions:

- Does the recommended content vary depending on the users' attitude towards migration?
- Does the recommended content vary depending on the location's political leaning?
- Does the recommended content vary over time during the U.S. midterm election?

After careful consideration, we concluded that the most suitable method to study the differences in outputs across the above conditions in a platform with public interface such as TikTok was a combination of sock puppet and scraping audit methods. Our approach with sock puppets involved training 9

accounts to reflect different attitudes towards migration (positive, negative and neutral), based in U.S. cities with different political leaning (Democrat, Republican and 'ambivalent'). Additionally, we employed scraping of the outputs i.e., recommended videos in TikTok's ForYou page at different points before, during and after the U.S. midterm election.

For data analysis, we chose to conduct content analysis of the full videos due to their shorter format, focusing on categories including type of content, subject of the video, sentiment and language among others. The categories for the content analysis we developed were informed by previous literature on migrant representation and our own insights about the goal of the audit.

7. Data collection

We used a custom virtual agent (bot) that simulates human interaction within the browser version of TikTok to train the sock puppet accounts by watching, liking and sharing videos with political messages on migration, and scraped the first 20 recommended videos on the "For You" feeds of each profile from October 8, 2022, until December 1, 2022.

8. Data analysis

Using content analysis, we analyzed a total of 1620 videos recommended to the sock puppet accounts in the selected time period. Our findings revealed that, despite little variation in recommended content based on users' attitude towards migration and their location's political leaning, political discourse on migration was virtually absent from the platform, indicating weak personalization for political content and a focus on entertainment rather than politics.

For recommendations and mitigation measures and more details on our auditing process, see the Auditing Social Media: (In)visibility of Political Content on Migration report.

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.

Use of facial recognition in the insurance sector

This audit evaluates the implementation of facial recognition in the insurance sector through a case study of the virtual assistant "Azul" by Zurich Seguros.

Specifically, we examine the experiences of individuals with disabilities with facial manifestations and in particular, those with Down Syndrome.

1. **Choosing the system**

We selected the insurance sector because the utilization of facial recognition is a relatively new application of FR technology that carries significant potential for negative impact and social harm. Previous audits of facial recognition software have highlighted deficiencies in performance, particularly on women and people of color. We identified individuals with disabilities and facial manifestation as a group at risk of discrimination, underrepresented in studies assessing the performance of FR technology.

2. **Contextual analysis**

During our contextual analysis, we reviewed existing audits of facial recognition software and investigated the extent to which insurance companies are using FR technology to process insurance applications and claims. We identified Azul by Zurich Seguros as an example of a virtual assistant that employs a set of algorithms for facial analysis including age, smoking status, and body mass index. The system then uses a risk assessment tool to generate a life insurance quote based on the attributes identified by FR. Given the outcomes of previous FR audits, we hypothesized that facial recognition in the insurance sector is likely to result in discrimination against people with disabilities.

3. **Feasibility assessment**

In this case, the AI system we intended to inspect did not provide a clear set of access points for an audit. Zurich Seguros does not disclose any information about the models and the virtual assistant Azul requires individuals to be present in front of a camera for the assessment – making open-source code audits, scraping, sock puppet and comparative output audit methods unfeasible. At the same time, reaching people with Down Syndrome for conducting experimental user or ethnographic audit without the cooperation of a civil society organization was challenging.

To illustrate the problem which we tried to tackle in this audit and secure the collaboration of civil society and affected communities, we conducted a feasibility assessment in the form of a pilot study to evaluate the performance of facial recognition software on people with Down Syndrome and people without Down Syndrome using publicly available images from the internet. The findings of our preliminary analysis clearly demonstrated that FR performed significantly worse for people with Down Syndrome compared to people without Down Syndrome across all metrics including gender, race, age and emotion classification. The pilot study confirmed the necessity of our audit and allowed us to proceed to the next steps.

4. **Stakeholder mapping and alliance building**

We proceeded to map civil society organizations working with people with Down Syndrome which could help facilitate access to the community for our audit. We contacted multiple organizations and presented the findings of our pilot study to secure a collaboration agreement with a suitable civil society organization.

5. **Methodology design**

Given the constraints of the system, which made it impossible to simulate user interaction, we opted for the experimental user audit method. Our primary goal was to assess the accuracy of the system's facial analysis for individuals with Down Syndrome compared to those without, and determine if any biases against people with Down Syndrome exist in the generation of life insurance quotes. To this end, we designed an experimental setup involving 20 participants with Down Syndrome and 20 participants without Down Syndrome carefully matching their characteristics such as age to compare the results across the two groups.

This audit is currently in its data collection phase, and we anticipate proceeding to data analysis soon.

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 an adversarial 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.

Audit of ride-hailing platforms

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

1. **Choosing the system**

This audit started with a concern that ride-hailing apps may not fully comply with competition, labor and consumer law in Spain. To examine these issues,

we partnered with Taxi Project, an organization that aims to improve conditions for taxi workers, and Observatorio TAS, an organization that defends the interests of workers in the platform economy, to conduct an adversarial audit of the algorithms of the three largest ride-hailing platforms in the country: Uber, Bolt and Cabify.

2. **Contextual analysis**

Our contextual analysis focused on understanding the legal and social environment in which ride-hailing apps operate in Spain. We did this through desk research of applicable legislation in the areas of competition, labor and consumer law, and expert interviews who provided insights into ride-hailing platforms' operation in the gray area of the private hire vehicles regulatory framework. This exercise served to focus our audit and shaped our initial research questions.

3. **Stakeholder mapping**

During our stakeholder mapping, we identified the following stakeholder groups:

- ride-hailing platforms such as Uber, Bolt and Cabify
- passengers who use ride-hailing services
- private hire vehicle (PHV) license holding companies
- PHV drivers
- traditional taxi companies and drivers
- customers of taxis and ride-hailing apps
- people in remote or low-income areas
- regulatory agencies for mobility services
- regulatory and enforcement bodies for competition, labor and consumer protection
- policy-makers in Spain and the EU

We singled out PHV drivers and users of ride-hailing apps in low-income areas as groups at risk of algorithmic bias and discrimination.

4. **Alliance building**

This step included establishing the roles, responsibilities and contractual agreements of all partners and collaborators in this audit.

5. **Methodology design and data collection**

Ride-hailing apps allow for the use of the scraping method for data collection, so we opted for this approach with regards to competition and consumer law. Additionally, we utilized the ethnographic audit method to incorporate the perspectives of workers in the sector and assess the implications for labor law.

Our research setup included:

- Selecting key routes in Madrid and Andalusia and scraping pricing data for the selected routes from Uber, Bolt and Cabify to examine whether ride-hailing platforms are indirectly fixing prices via algorithmic means.
- Conducting interviews with PHV drivers to assess the extent to which algorithmic processes incorporate existing labor legal protections, specifically in relation to leave of absence and payment transparency.
- Selecting routes in high, medium and low-income areas in the cities of Madrid and Málaga and scraping price data for the selected routes from Uber, Bolt and Cabify to determine the presence of geographic discrimination in consumer prices.

6. **Data analysis**

We used linear regression analysis to interrogate the correlations between the prices of Uber, Bolt and Cabify in relation to competition law. We found moderate to positive correlations in the prices of Uber and Cabify, and Uber and Bolt, indicating possible price collusion on selected routes.

We used the same approach in investigating the correlation between the median income of a neighborhood and trip fares in the area in relation to consumer law. We found a weak to moderate negative correlation between median income and trip fares i.e., fares in low-income neighborhoods tend to be higher, indicating possible geographic price discrimination.

Finally, the thematic analysis of our qualitative data from interviews with PHV drivers highlighted that workers are not adequately protected from algorithmic sanctions in cases of lawfully protected reasons for absence from work and that their payments through platforms are not transparent.

For recommendations and mitigation measures and more details on our auditing process, see the Adversarial audit of ride-hailing platforms report.

AUDIT REPORT INDEX

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:*

- *Quantitative findings*
- *Qualitative findings*

V. *Discussion*

VI. *Recommendations and mitigation strategies*

VII. *Limitations*

VIII. *Conclusion*

References

Acknowledgements

Annexes

ACKNOWLEDGEMENTS

Project team: Adversarial Audits
Research Director: Dr. Gemma Galdon-Clavell, Founder of Eticas Tech
Research Lead: Iliyana Nalbantova, Junior Ethics and Technology Researcher at Eticas Tech
Contributing Researchers:

- Asylai Akisheva, Ethics and Technology Researcher at Eticas Tech
- Luís Rodrigo Gonzalez, Ethics and Technology Researcher at Eticas Tech

Other Contributors:

- Matteo Mastracci, Ethics and Technology Researcher at Eticas Tech
- Isabela Miranda, Project Manager at Eticas Tech
- Patricia Vázquez, Marketing and Communications Manager at Eticas Tech

Recommended citation: Eticas (2023). Adversarial Algorithmic Auditing Guide. Association Eticas Research and Innovation.

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.

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

Automated discrimination: Facebook uses gross stereotypes to optimize ad delivery. (n.d.). AlgorithmWatch. https://algorithmwatch.org/en/automated-discrimination-facebook-google/

Bandy, J. (2021). Problematic Machine Behavior: 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/gfk-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

Cucchietti, 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

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' Behavior 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. Association Eticas Research and Innovation.

Eticas (2023). Auditing Social Media: (In)visibility of Political Content on Migration. Association Eticas Research and Innovation.

Eticas (2023). Auditing Social Media: Portrayal of Migrants on YouTube. Association Eticas Research and Innovation.

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

Eticas. (2022). The External Audit of the VioGén System. Association 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/icwsm.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

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

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=UAmqdwJ9jl-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

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

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

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

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

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

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/

eticas
TECH

CONTACT

info@eticas.tech
www.eticas.tech

ADVERSARIAL ALGORITHMIC AUDITING GUIDE