WIK

Working Paper

No. 12

A cross-domain framework for auditing algorithms - From harmful content to self-preferencing

Franziska Harpenau
Faisal Abbasi
Dr. Lukas Wiewiorra

IMPRINT

WIK WORKING PAPERS

The working papers published in the series constitute work in progress circulated to stimulate discussion and critical comments. Views expressed represent exclusively the authors' own opinions and do not necessarily reflect those of the editor.

WIK Wissenschaftliches Institut für
Infrastruktur und Kommunikationsdienste GmbH
Rhöndorfer Str. 68
53604 Bad Honnef, Germany
E-Mail:info@wik.org
www.wik.org

Person authorised to sign on behalf of the organisation

Board of Directors	Dr Cara Schwarz-Schilling (Chairwoman of the Management Board, Director) Alex Kalevi Dieke (Chief Financial Officer)
Authorized representatives	Prof. Dr Bernd Sörries Dr Christian Wernick Dr Lukas Wiewiorra
Chairperson of the Supervisory Board	Dr Thomas Solbach
Registered at	Amtsgericht Siegburg, HRB 7225
Tax No.	222/5751/0722
VAT-ID	DE 123 383 795

Date: December 2025

From harmful content to self-preferencing: A cross-domain framework for auditing algorithms

Franziska Harpenau[1], Faisal Abbasi[1], Lukas Wiewiorra[1]

[1] Wissenschaftliches Institut für Infrastruktur und Kommunikationsdienste (WIK)

December 19, 2025

Abstract

As online platforms increasingly rely on algorithmic systems to curate, rank, and remove content, concerns about the societal and economic consequences of algorithmic decision-making have intensified, and with them the importance of algorithmic auditing. However, there is a lack of universal, practical guidance on the specific steps involved in designing and implementing platform audits across different problem domains and under varying conditions of data access, methodological constraints and legal obligations. This paper aims to address this gap by deriving a generalised framework with fine-grained, concrete steps required for the practical implementation of algorithmic audits using a broad focus on potential methodologies and a broad focus on potential application scenarios. The framework is based on a structural analysis of empirical audit approaches in the two domains of moderation of illegal and harmful content and of potential self-preferencing in recommendation algorithms, combined with further methodological literature. On this basis, the paper specifies the sequence of steps involved in carrying out audits in practice, identifies recurring challenges encountered across settings, and proposes a classification of audit approaches, incorporating advantages and disadvantages, and concrete examples. The proposed unified process model refines and operationalises existing high-level frameworks, while generalising more fine-grained domain-specific approaches. Furthermore, it demonstrates that meaningful audits can be conducted under restrictive access conditions, but that platform support can substantially improve the feasibility of precise, generalisable and causally informative findings. Taken together, these contributions offer regulators, platforms, and independent auditors a practical basis for systematically scrutinising complex platform algorithms and for interpreting audit results with greater rigour.

Keywords: algorithmic auditing, digital platforms, content moderation, self-preferencing

[1] E-mail addresses: F.Harpenau@wik.org, F.Abbasi@wik.org, L.Wiewiorra@wik.org
Wissenschaftliches Institut für Infrastruktur und Kommunikationsdienste (WIK), Rhöndorfer Str. 68, 53604 Bad Honnef, Germany.

1 Introduction

Algorithmic decision-making systems have become integral components of major online platforms. They determine which content is removed as illegal or harmful, which products and services are recommended most prominently, and which individuals and businesses gain or lose visibility. In response, regulators in the European Union and other jurisdictions are increasingly relying on auditing, in the sense of examining algorithmic systems regarding potential problematic behaviour and risks, as a central governance instrument for managing systemic risks on very large platforms. Despite these developments, there is still limited universal practical guidance on the concrete steps regarding the design and implementation of platform audits across different problem domains and under varying conditions of data access, methodological constraints, and legal obligations.

A growing body of empirical research has examined individual systems for various problematic behaviours, such as discrimination, the amplification of harmful or illegal content, or the prioritisation of self-interest in recommendations. However, the empirical studies are tailored to specific platforms and research questions, and available auditing frameworks are often high-level or restricted to particular domains. Consequently, a discrepancy exists between the abstract principles of responsible algorithmic governance and the detailed methodological decisions that researchers, regulators, and civil society organisations must make when planning and evaluating audits of real-world platforms.

This paper addresses this gap by providing a structural analysis of empirical auditing studies in two key application areas: the moderation of illegal and harmful content and potential self-preferencing in recommendation systems. For each domain, a mapping process is undertaken, incorporating factors such as audit objectives, methodological stance and challenges faced, as well as the kinds of data access on which the studies rely. This ranges from fully external interaction with platform interfaces to extensive platform support. Building on this, a generalised, stepwise auditing process has been developed. The process begins with the formulation of risk scenarios and system scoping, progressing to the selection of methodological approach, incorporating considerations regarding the desired types of insights and associated prerequisites, the concrete indicators, the design of data collection and labelling, the management of challenges such as ethical and technical issues, and the analysis, robustness checking and reporting of results.

The analysis yields three main results. Firstly, we derive a unified process model for auditing algorithms that refines and operationalises existing high-level frameworks and generalises domain-specific frameworks. Secondly, we develop a classification of auditing methodologies, incorporating advantages and disadvantages, and concrete examples. Thirdly, we find that under certain circumstances meaningful audits can be conducted under restrictive access conditions; however, platform support can substantially improve the feasibility of precise, generalisable and causally informative findings. Together, these contributions aim to provide regulators, platforms and independent auditors with a practical framework for systematically scrutinising complex platform algorithms themselves and understanding the results of others more thoroughly.

2 Related literature

The present paper pertains to the extensive and expanding domain of auditing algorithms. Bandy (2021) conducted the first systematic literature review of empirical studies examining public algorithm systems for potential problematic behaviour in this field. He identified four categories of problematic behaviour for which public-facing algorithms have been empirically audited: discrimination, distortion, exploitation and misjudgement, with the second two categories receiving the most attention. Auditing of various algorithms has been undertaken by researchers; however, a predominance of research focuses on search algorithms, followed by advertising and recommendation algorithms. The author categorised the used audit methods into code audits, direct scrapes, sock puppets, carrier puppets and crowdsourcing, purposefully omitting user studies from the analysis. The most prevalent method is direct scraping, followed by crowdsourcing and sock puppets. The author identified gaps for future audits to fill, including discrimination based on intersectional identities, more research in the advertising domain, more use of code audits and certain organisations being currently audited less extensively. A recent, second systematic literature review of empirical studies auditing algorithms on public-facing platforms was conducted by Urman et al. (2024a). They largely align with the previously identified categories, focus areas and gaps. Nevertheless, with regard to the audit methods, carrier puppets have been replaced with platform repurposing, which is exclusively employed in the domain of advertising algorithms, and instead of differentiating between sock puppets and direct scraping, a differentiation between personalised scraping and non-personalised scraping has been used. The most prevalent methods were identified as non-personalised and personalised scraping. In addition to the analyses conducted in Bandy (2021), further research characteristics have been analysed, leading to the identification of a strong focus on Western liberal democracies and English-language content as the research environment, with authors being concentrated in the US and a few European countries. Further summarising research synthesised quantitative empirical results from research studies for the specific problem definitions of recommendation pathways and quality (Hibert et al., 2024; Yesilda and Lewandowsky, 2022).

The present study integrates particularly into extant research regarding the establishment of frameworks for the auditing process. A number of frameworks adopt a more expansive process-oriented perspective, while others adopt a more specialised and detailed approach, focusing on the technical aspects of auditing algorithms. For instance, with regard to audits concerning the DSA, Meßmer and Degeling (2023) proposed a risk-scenario-based audit process in which, based on the developed understanding of the platform and its stakeholders, concrete testable risk scenarios related to systemic risks are defined and prioritised, measures for them analysed and prioritised, and the results evaluated and reported. Hasan et al. (2022) delineated their process for auditing algorithmic systems regarding ethical risks in a second-party manner in a similar way. The process is initiated with the identification of significant stakeholders and the most extensive range of potential harms they might encounter followed by a prioritisation of these harms and a bias assessment with concrete testing metrics. Zicari et al. (2022) proposed a three-stage process related to the EU Framework for Trustworthy AI, that applies across all stages of the AI life cycle and is carried out with operator support. The first stage is the setting up stage, which includes defining the scope and forming an adequate interdisciplinary team. The second stage is the assessment stage, which includes socio-technical scenarios based on gathered context and potential ethical issues and tensions. The third stage is the resolving stage. Raji et al. (2020) developed an internal algorithmic auditing framework for AI, encompassing the processes of scoping, mapping, artifact collection, testing, reflection and post-audit. Koshiyama et al. (2024) delineated four dimensions of auditing algorithms – development, assessment, mitigation and assurance – relating to four auditing verticals, namely explainability, robustness, fairness/bias and privacy, and described different levels of potential access for auditors.

Research focusing on frameworks with a more application-related focus includes Morales-Navarro et al. (2025), who established a generic scaffold to assist young people in auditing algorithms. This scaffold encompasses the broad, little detailed steps of developing a hypothesis, generating a set of systematic, thorough and thoughtful inputs, conducting a test, analysing the data, and reporting the results. Metaxa et al. (2021) analysed key decision points with related best practices in auditing algorithms with a focus on search engines. The identified key decision points relate to legal and ethical considerations; the selection of a research topic; the selection of an algorithm to audit; temporal considerations; data collection; measuring personalisation; which interface attributes to collect; analysing data; and communicating findings. Pattrn Analytics & Intelligence (2023) provided a detailed flow of decision forks when auditing a recommendation algorithm with regard to the dissemination of illegal and harmful content in a first-party manner. This decision-making process encompasses the elements of the party conducting the evaluation (while focusing exclusively on first-party audits); the outcome to be measured; the type of insights to be produced, such as causal or descriptive results; how to identify harmful content; how to perform causal inferences; how to survey users; what simulation to conduct; how to quantify virality; and how to quantify pathways. It should be noted that, depending on the choices made in previous decision forks, not all of the following forks may be relevant. Furthermore, a distinction is made between three overarching themes: observation, experimentation and self-reporting methodological elements. Very recently, Panigutti et al. (2025) developed structure plans with key methodological steps and best practices for analysing algorithms in relation to the DSA. The plans are for risk-uncovering studies, reverse engineering studies, interface design studies and risk-measuring studies. The most relevant plans for this paper are the reverse engineering studies and risk-measuring studies, which consist of the steps of scope determination; hypothesis formulation; experimental design definition, including the measured outcome, treatment, confounding variables and baseline comparison and noise variables; data collection; if necessary, ground truth labelling; and data analysis. The risk-measuring studies potentially fall into the categories of experimental or observational studies, which are characterised by actively interacting with the algorithm, respectively analysing historical data passively.

Further research in this area has focused more generally on potential problems and best practices to consider, more separated from also setting up frameworks (e.g., Costanza-Chock et al., 2022; Metcalf et al., 2021; Mökander and Floridi, 2023; Whittlestone et al., 2019).

The present study contributes to the extant literature by providing a comprehensive description of the fine-grained, concrete steps required for the practical implementation of algorithmic auditing using a broad focus on potential methodologies—for instance, incorporating user surveys and platform-supported approaches—and a broad focus on potential application scenarios, extending beyond specific domains or specific legislation. This analysis is based on an overview and examples of applications under varying circumstances in two different domains that have not been previously examined from this perspective, with triangulation with additional domains and methodological literature at certain points. It is therefore possible to contrast and combine our generated results with existing application-oriented frameworks, thereby creating greater generalisability across different domains and settings, and providing valuable examples and details so far left out. The identified steps can be built upon and integrated into the frameworks with a broader focus.

3 Data collection approaches

In order to establish a theoretical foundation for comprehending the following analyses and to present an overview of the generally feasible data collection approaches, a wide range of them is described and evaluated based on the existing literature in the following. Some of the approaches may be conducted in a black-box way, i.e. without platform/provider support, while others focus on how the data collection might look like with various levels of cooperation with the platform/provider for conducting the audit.

3.1 Black-box methods

A variety of methodologies can be employed to conduct an audit of an algorithmic system. The foundational framework for this research was established by Sandvig et al. (2014), with subsequent contributions extending its scope. The applicability of the various methods is contingent upon the level of access to the data, the specific case under examination, and the available resources. Black-box methods can be utilised as a preliminary step in the generation of evidence, and in the majority of cases, represent the sole available option.

Collecting data with black-box access can happen via **scraping**, in which data is automatically collected from the regarded system and subsequently analysed, thereby allowing for scale of data collection. This can be achieved by executing written scraping code, which collects specific data from the system as is. However, the available data is limited by what can be scraped, and in most cases, no active interactions are mimicked by the script. This method is suited to understanding which kind of content is displayed in the system in the situation regarded. However, no consideration is given to the interplay between users and the system. Moreover, the process of coding can necessitate a considerable investment of effort, as a distinct code is required for each system and alterations to a system can also result in the requirement for an adapted code. Additionally, certain systems explicitly prohibit the practice of scraping information in their terms of service and sometimes block the activity of scraping attempts, which may create challenges for individuals employing this auditing method. Despite the existence of some rulings in favour of violating the terms of service to conduct an audit, the situation remains uncertain (Urman et al., 2024b).

In certain instances, instead of scraping the information, researchers have the option of accessing the data via an **API** of the examined system. The researchers can access the API's data through written computer programs and can thereby send data such as a search string to the system and receive the corresponding data such as search results from the system. In this case, less work is required to collect the data, but a dependence on the system regarding data breadth and quality is created. The data that is available via the API may be more extensive or more limited than that which is obtainable through scraping, and it may be accessible to all or only a select group of individuals. Moreover, the researchers must either have confidence in the coherence of the delivered data with the actual data displayed in the system, or they must implement a suitable checking procedure.

In certain instances, the platform offers functionalities that can be utilised for **repurposing** activities. This phenomenon is most prevalent within the domain of auditing advertising algorithms, where statistics provided to advertisers, for instance regarding impression distribution, can be utilised directly as a foundation for analysis (Urman et al., 2024a). Consequently, depending on the extent of data collection, the repurposed data can simply be manually collected. The specific possibilities and setting are heavily dependent on the function that is being repurposed. In certain instances, this data may be accessible via an API, thereby falling into the preceding category.

The **sock-puppet audit** is a method that combines scraping with controlled user impersonation. For this purpose, so-called sock-puppet accounts are created and programmed with the intention of impersonating users with specific characteristics and behaviours, for instance using automated browsers such as selenium or puppeteer, but a manual operation is also possible, though likely to require more effort and to be less scalable. The data created by these accounts is then automatically collected. This approach facilitates the analysis of the influence of specific user characteristics on the system's output in a controlled, large-scale setting. This approach bears a closer resemblance to an experimental setting than a conventional scraping audit. Nevertheless, it is evident that the sock puppet accounts are merely impersonating genuine user behaviour, thereby resulting in a representation that is likely to be less than perfect, particularly in instances where more than the initial few interactions need to be taken into account. In cases where the impersonation of specific characteristics is not straightforward, it is imperative to verify that the targeted characteristics are indeed being mimicked. In the context of specific systems, the verification of an account can be a laborious process. For instance, the requirement for a telephone number can incur significant expenses, thereby becoming a financially onerous undertaking. Furthermore, as with scraping, conducting this type of auditing method might be against the terms of service of certain websites and risks being blocked if unauthenticated behaviour is detected. Moreover, the workload associated with writing code is likely to be at least as onerous as that associated with scraping.

The collection of output information for different profiles can also be achieved through the implementation of a **crowdsourced audit**. This method relies on the utilisation of hired users as testers, who subsequently collect data while utilising the system. The data can be collected manually by the users or, in most cases, through automated means, such as a browser extension, achieving a controlled and more encompassing data collection. The users can either adhere to their usual behaviour or be directed to follow specific instructions so that comparable data can be collected from all users. This data pertains to the direct user experience; however, it is challenging to make any causal claims regarding specific user attributes, as there may be a multitude of unobserved variables. Moreover, as not all subjects may be willing to share their data in this manner, the achievement of a representative sample may be hindered, and costs for recruiting participants may be incurred if they do not voluntarily share their data for research purposes. Furthermore, if an automated data collection method is employed, issues pertaining to the necessary effort are similar to those associated with scraping. However, as previously discussed, the utilisation of automated or semi-automated data collection methods may result in violations of the terms of service and involves the risk of being blocked.

The administration of a **user survey** can facilitate an examination of the user experience across diverse user groups. For this purpose, a representative sample of the target population, such as the general user pool, should be recruited. However, it should be noted that user surveys are subject to issues such as social desirability and recall bias. Consequently, they may be less appropriate for applications where such biases are likely to have a significant impact. Furthermore, it is important to note that only users' experience and assumptions can be collected, and no definitive statements regarding the algorithm's functionality can be made. Moreover, expenses must be incurred with regard to the recruitment of participants.

A related approach is that of an **end-user audit**, in which the system's output is audited by the user themselves with the assistance of a supporting infrastructure (Lam et al., 2022). The users can identify discrepancies between their assessment and the algorithm's, thereby generating hypotheses concerning potential issues that can be directly quantified using the infrastructure. In the example case presented in Lam et al. (2022), recruited users would first label a few textual posts concerning their toxicity, which is used for extrapolation to a more extensive set of posts, thereby enabling analysis of a toxicity classification algorithm on a larger scale. This methodology encompasses the perspectives of actual

users, including those from marginalised communities, who may be particularly affected by erroneous behaviour, yet may not be adequately represented in conventional labels so that previously unconsidered hypotheses may emerge. Moreover, it affords a more systematic perspective than that achievable through surveys. However, users must demonstrate a certain degree of motivation in identifying issues, which may result in a non-representative group of users conducting the audit. Once more, costs for recruiting participants might need to be incurred. Furthermore, in order to employ such an approach, it is necessary to have access to the algorithm's output and to a sufficiently large set of posts with labels from different annotators. The algorithm's output is used for the comparison of labels; the set of labels is required for the extrapolation; and a sufficiently large set of data is required for the training of a personalised user model and for systemic understanding. Therefore, the utilisation of an accessible algorithm, a scraped data set of posts, and a set of labels generated by crowdsourced workers would be a viable option; however, platform support would significantly reduce the workload and expand the range of available options.

Another associated approach is that of a **sociotechnical audit**, which analyses the interplay between users and algorithms in a controlled setting (Lam et al., 2023). In addition to the analysis of the algorithm and its behaviour, the users are studied under different algorithm behaviours. This provides an opportunity to analyse the effects on the users, including any modifications in behaviour. One potential approach for this would be to first analyse the algorithm using one of the other methods for auditing, and subsequently combine this with a controlled laboratory experiment on users. Alternatively, the second part could be conducted in the real environment as well, meaning that the algorithmic output is altered directly where it is normally placed. This can be achieved through browser extensions or similar software, or, in the case of platform support, by conducting A/B testing in which the regarded system is directly altered for a subset of regular users. The modification of the algorithmic system, especially if conducted in a black-box manner, will likely require a significant investment of effort, and should participants be recruited, financial costs will incur.

3.2 Platform-supported methods

Some of the described black-box methods can also be pursued with platform support, either happening voluntarily or due to legal obligation, thereby lowering the required effort and increasing options. This was already noted for end-user or sociotechnical audits, but, for instance, also holds true for user surveys, where participants may be directly recruited through the platform. Other methods, such as scraping or crowdsourcing, might not be necessary, as the platform might already store the required data or be able to generate it oneself using the system and existing data. Of course, trust in the platform to collaborate in good faith or to adequately sanction dishonesty is necessary.

Moreover, further approaches become viable with platform support depending on the access provided. It is possible to differentiate between several levels of platform support, including white-box, grey-box and outside-the-box access (Casper et al., 2024). In the outside-the-box access, the auditor is granted access to information regarding the development and deployment of the system, such as model details, training data, deployment details, and the results of conducted internal evaluations (Casper et al., 2024). In the grey-box access model, restricted access with varying intensity to the inner workings of the system is provided, for example to intermediate computations that influence the model's decision-making process. In a highly unrestricted form of grey-box access, the de facto white-box access, auditors are able to execute arbitrary processes on a system in an indirect manner, while ensuring that the system's parameters cannot be directly accessed and therefore remain protected from duplication, for example using an API. The most unrestricted form of access is the white-box access, under which full access to

the system's inner workings is permitted, encompassing the ability to execute commands, fine-tune the model, and access information on parameters and the like. Other taxonomies differentiate further between the spectrum from black-box to white-box, for example between the possibility to sample from the model, to fine-tune the model, to inspect the model internals, to modify the model internals and access to additional system information (Bucknall and Trager, 2023) or differentiate between seven levels with increasing access from only checklist up to white-box access (Koshiyama et al., 2024).

In general, the higher the level of access to information, the more types of analysis that can be conducted. For instance, access to the training input data is necessary to detect any biases within this data; access to the training outcome data is required to precisely analyse the accuracy of the system; access to the parameter control is essential for assessing the stability of metrics such as bias. With sufficient access, a code audit, in which the code or the pseudocode of an algorithm, i.e. a written representation of the code, is examined for comprehending its functioning, could be conducted. In principle, this approach offers the most comprehensive insights, as it facilitates the comprehension of all interconnections and objectives. However, in reality, the employed codes are mostly very complex and extensive, therefore hard to understand and it is difficult to predict outcomes when just reading the code. Moreover, there is no interaction with the user at all, and the probability of misconduct being explicitly coded is low; instead, it is more probable that such behaviour is an emergent property of the algorithm in use. Consequently, a combination of analyses is valuable. In general, access to contextual outside-the-box information facilitates more efficient and impactful auditing (Zaccour et al., 2025).

A higher level of access can also facilitate enhanced quality and generalizability of the analyses, as evidenced by problems that have not been identified with black-box access due to a lack of information necessary for selecting important input data (Casper et al., 2024). Black-box methods are capable of identifying problems in certain cases; however, white-box methods can also be used to evidence the correct functioning of the algorithm (Casper et al., 2024). A code audit, but potentially all platform-supported audits, pose a challenge with regard to intellectual property, such that a safe infrastructure should be in place, and, in the absence of voluntary sharing, there must be a solid reason for requiring the provision of internal documentation and system access.

Several methods exist for preserving the privacy of users and the intellectual property of the company in cases where such safeguards are deemed to be necessary. For instance, the implementation of tailored APIs, which are designed to provide a specific level of access, could be useful (Bucknall and Trager, 2023; Casper et al., 2024). However, further research in this area could prove beneficial (Bucknall and Trager, 2023; Zaccour et al., 2025). Other approaches include conducting a white-box audit on site at the system's facilities in a safe environment, reducing any information leakages, or drawing up legal contracts regarding non-disclosure or absence of conflict of interest (Casper et al., 2024).

If concerns regarding user privacy arise, providing aggregated data in a privacy-preserving way for both the user and the platform itself, due to the application of a differential privacy mechanism, could be employed (Imana et al., 2023). One such example, as outlined in Imana et al. (2023), involves the platform querying its content-delivery algorithm with input data concerning trial content and a list of users with known demographic attributes, separately for each regarded demographic group, calculating the relevance scores, applying a differential privacy mechanism, and returning the distribution of the privacy-preserving data so that the auditor can evaluate fairness using the obtained data. The implementation of this differential privacy mechanism results in an increase in the required sample size for accurate auditing. In the specific example outlined, this would result in a sample size approximately four times larger when employing typical parameters. Using differential privacy mechanisms for aggregated data was found to be reliable, provided that the privacy budget was not set at an exceptionally low value or the sample size was small (Zaccour et al., 2025). In an ideal scenario, the platform would communicate

the selected parameter such that an evaluation of the quality can be conducted (Zaccour et al., 2025). However, access to individual-level data, as opposed to aggregated data such as is provided by the aforementioned mechanism, is invaluable for a wide range of analyses and a comprehensive audit (Zaccour et al., 2025). In contrast to the differential privacy mechanism for aggregated data, the use of synthetic individual-level data, based on a differentially private mechanism or not, does not appear to provide reliable results in at least certain cases (Zaccour et al., 2025). Conducting data minimisation by removing certain features can ensure the maintenance of auditing quality, provided that all important features are retained; however, it is not always evident which features are important (Zaccour et al., 2025). Consequently, at least in certain cases, the dissemination of actual individual-level data must be safeguarded through alternative means, such as the implementation of one of the other aforementioned privacy preservation techniques, for instance in combination with less disturbing, but thereby less secure, anonymisation methods.

4 Application domain

In the following discussion, extant scientific literature on algorithmic auditing in two different domains is analysed with regard to followed approaches and structures, and with regard to similarities and differences therein. Furthermore, potential challenges and adequate remedial actions are examined, depending on different types of circumstances. Some approaches considered are not directly related to the problem under investigation in the respective application domain but are included in the analysis because they provide valuable insights based on a close connection to the problem. Collectively, these elements offer a structured understanding of algorithmic audits in these domains, while also providing insights regarding important decision points and commonalities and differences in approaching auditing algorithms in different situations. These findings can be utilised in the following section for generating a general flow with insights into the very practical challenges.

The two regarded application domains – content moderation of illegal and harmful content, and self-preferencing in recommendations – focus on two of the four problems identified by meta-analyses for which algorithms are audited (Bandy, 2021; Urman et al., 2024a), namely misjudgement and distortion, as well as on different algorithm types in order to achieve a broader picture. In Section 5, which sets up the generalised, research regarding further problem and algorithm types is included at certain points. To facilitate a deeper comprehension of the auditing approaches, the general functionality of the two algorithm types is outlined briefly at the beginning of each respective section.

4.1 Content moderation of illegal/harmful content

4.1.1 Algorithmic content moderation

Platforms implement content moderation practices in order to ensure compliance with the relevant legal and regulatory frameworks, as well as to enforce their own platform-specific community guidelines. However, the specific moderation process and criteria itself may vary between jurisdictions, since platforms must comply with the policies and restrictions imposed by the respective jurisdictions in the markets they cover. Given the sheer volume of such processes, on large platforms, the majority of steps are automated via algorithms. In the course of this process, it is imperative for platforms to respect and appropriately balance users' freedom of expression while moderating content.

The implementation of the moderating processes can be achieved through an algorithmic approach, manual intervention by human moderators, or a combination of both, with the majority of platforms employing a hybrid approach (Lykouris and Weng, 2024; Gosztonyi et al., 2025). A hybrid approach has the potential to facilitate the expeditious and effective management of substantial content volumes by platforms, while preserving the capacity for human judgement to exercise context-awareness and nuance, particularly in cases of ambiguity or complexity (Lykouris and Weng, 2024; Ami, 2025). Subsequent to the uploading of content, it is subjected to a process of *pre-screening and/or post-screening*. Pre-screening, also termed pre-moderation, denotes the process of content review prior to its publication, potentially resulting in a delay of the publication. Post-screening, also termed post-moderation, denotes the process of reviewing content following its publication. The problematic items may be removed later if they are found to be in violation of the guidelines. A variety of matching or classification techniques can be employed for this by algorithms. A common form of content matching relies on hashing techniques, which transform content into a "hash", meaning a string of data, that is then used to identify underlying content. Hashes are advantageous due to their reduced size and expedited processing, facilitating efficient matching with existing hashes of illegal content libraries, including those of the Global Internet Forum to Counter Terrorism (Fink and Saltman, 2025). Moreover, content classification methods are employed to evaluate newly uploaded content and allocate it to predetermined categories, such as hate speech, terrorism, sexual content, or child abuse, in most cases being based on machine learning algorithms (Gorwa et al., 2020). It is evident that each platform utilises a distinct moderation system, yet they employ analogous core technologies, such as hashing, classification, machine learning and natural language processing (NLP), for the purpose of content moderation (Gorwa et al., 2020; Boroughf, 2015; Lees et al., 2022; Steinebach, 2024). The employed algorithms might either act directly themselves, or flag it for human review, with several different combinations based on the circumstances being possible (Gorwa et al., 2020). These systems are programmed *to execute predefined remedial actions*, which may include the binary option of removal or no action, a temporary restriction on access until a subsequent review, as well as dozens of potential other remedies such as account regulation and visibility reduction (Goldman, 2021). Furthermore, a post may be *flagged* by other users or organisations, thereby instigating ex-post reactive moderation and a further review. In addition, platforms may employ *proactive ex-post moderation* techniques, such as the manual removal of certain harmful content (Klonick, 2017).

4.1.2 Analysis of empirical literature

The first analysis of an application domain relates to the potentially insufficient moderation in the form of removal of illegal or harmful content by the algorithmic content moderation system with the aim of focusing on the effects of the algorithmic system rather than those of human moderation. The following section delineates and categorises the approaches and steps of existing empirical analyses regarding this problem.

4.1.2.1 Methodological approach

The initial step in the audit of content moderation algorithms is the establishment of the objectives and the nature of the insights to be generated which exerts a significant influence on the subsequent steps of the audit. We propose that the objectives and their approaches fall into *three categories*. Some approaches can be assigned unequivocally to one group, while others have a focus regarding one group, but are related to a different group as well. The first identified group analyses the current prevalence of illegal content on the system, which can be used to draw conclusions about the false negative rate of the content moderation system. Research in this group includes the identification of posts promoting

illicit goods and services on Twitter (Wang et al., 2025), a temporal analysis of the prevalence of Russian propaganda and low-credibility content related to Russia's invasion of Ukraine on Twitter and Facebook (Pierri et al., 2023), an analysis of the "toxicity" of comments made in livestreams on Twitch (Dreier and Pirker, 2023), and an analysis of the presence of harmful content in far-right livestreams on TikTok (van Boheemen et al., 2025). Other papers in this group employ the analysis of illegal content solely as a foundational step to be combined later with the effects of the recommender system. These include an analysis of the prominence of different kinds of misinformation in various recommended content on YouTube, respectively of the prominence of misinformative health products on Amazon, and the effects of personalisation on this (Juneja et al., 2023; Hussein et al., 2020; Juneja and Mitra, 2021), of the recommendation network of ISIS content on YouTube (Murthy, 2021) and of the frequency and speed of encountering harmful videos using young persona accounts on TikTok, YouTube and Instagram (Eltaher et al., 2025). In the following analysis, the examination of this subgroup of papers will be limited as much as possible to this initial step of identifying illegal content.

The second group analyses the functioning and performance of moderation algorithms. This encompasses the examination of existing content moderation software for the identification of nudity and pornography (AlDahoul et al., 2023), of trained hate speech and CSAM classifiers (Del Vigna et al., 2017; Gutfeter et al., 2023; Vitorino et al., 2018), of Twitch's moderation system for chats in livestreams (Shukla et al., 2025), of the moderation ability and performance of LLMs regarding harmful content such as hate speech, violence, pornography and predatory chats (AlDahoul et al., 2024; Kumar et al., 2024; Nguyen et al., 2023; Mahomed et al.; 2024) and of the impact of different parameters or types of fine-tuning on the outcomes (Nguyen et al., 2023; Kumar et al., 2024). Furthermore, the analysis of alignment between an existing ML-based toxicity classifier and human judgement and decision components thereof (Muralikumar et al., 2023), the analysis of predictive multiplicity of LLMs used for detection of toxic content (Gomez et al., 2024) and the analysis of adversarial attacks on toxicity and hate speech classifiers (Gröndahl et al., 2018; Hosseini et al., 2017; Shukla et al., 2025) are included. Furthermore, a number of studies examine how the performance of content moderation algorithms varies based on user-related characteristics of the content, such as the presence of sexuality- or religion-related terms. This includes the analysis of the biases of toxicity classifiers and the application of mitigation strategies in training (Dixon et al., 2018; Sap et al., 2019; Borkan et al., 2019; Nogara et al., 2025), as well as the comparison of the algorithmic classification of content as toxic with the classification of individual users to surface potential problematic biases (Lam et al., 2022).

The third group analyses how users perceive illegal/harmful content and the moderation thereof. Approaches involving the surveying of users with regard to their experience of content moderation, as well as their online exposure to content that should be moderated, are part of this category (Haimson et al., 2021; Myers West, 2018; Lahti et al., 2024; Smahel et al., 2020), although not all of them employ a platform-specific approach. Haimson et al. (2021) highlight that the decision to utilise a survey as a research method was driven by the unavailability of access to moderation logs, owing to the black-box nature of their audit.

4.1.2.2 Causality

In conjunction with the decision on the general objective and the related group of approaches, lies the decision on whether causality in the results is necessary, meaning that certain observed factors can be clearly linked to the functioning of the algorithm. This could be a high prevalence of illegal content found on the system being clearly linked to a high false negative rate, or differences in performance being clearly linked to certain content characteristics. The first group of approaches never achieves causality, as their observational data does not allow for this. Even though Wang et al. (2025) and van Boheemen

et al. (2025) analyse evasion tactics, it can only be hypothesised that these very tactics are the reason why the content has not been moderated.

Approaches in the second group achieve causality by inputting data with known characteristics into the algorithms and analysing the results. Moreover, a number of studies analyse the effect of changes in prompts, classification threshold, training seed, fine-tuning and training data by changing only this aspect but nothing else (Kumar et al., 2024; Gomez et al., 2024; Nguyen et al., 2023; Gröndahl et al., 2018; Gutfeter et al., 2023; Vitorino et al., 2018; Shukla et al., 2025; Dixon et al., 2018). In other studies, the analysis of differences between models is conducted by utilising the same test data, and in some cases, training data, while striving to maintain as much constant as possible, except for the model (Kumar et al., 2024; Gröndahl et al., 2018; AlDahoul et al., 2024; Del Vigna et al., 2017; AlDahoul et al., 2023; Gutfeter et al., 2023; Vitorino et al., 2018). Furthermore, some studies analyse differences in performance when the same model is applied to different datasets, for instance collected from different platforms or concerning different topics or identity-terms (AlDahoul et al., 2024; AlDahoul et al., 2023; Gutfeter et al., 2023; Shukla et al., 2025; Dixon et al., 2018; Muralikumar et al., 2023; Sap et al., 2019; Borkan et al., 2019; Nogara et al., 2025). However, the capacity to ascribe variations in performance to a specific content attribute is contingent upon the extent to which other characteristics may also be altered. For instance, Muralikumar et al. (2023) try to achieve a high degree of comparability between data from different platforms by sampling only news content. In an effort to analyse the reliance of moderation on slurs, Shukla et al. (2025) replace identity terms with slurs in false negatives and, in an effort to analyse context awareness, they integrate sensitive terms of existing toxic content into a non-offensive, empowering context, all while keeping the rest constant in both cases. In order to control for other factors that might vary in a real dataset in addition to identity and increase the scope of causality, Borkan et al. (2019) and Dixon et al. (2018) utilise a synthetic dataset comprising content that varies solely in terms of identity, in conjunction with a real dataset. Nogara et al. (2025) create an altered dataset by translating text between languages to control for biasing factors such as content when analysing a bias regarding language of content, and additionally controlled for other factors such as special characters. However, they note that due to the proprietary nature of the algorithm and limited access, they cannot clearly identify the underlying cause of their results. Generally, the identified causality can be generalised to different degrees. For instance, utilising data exclusively from a single service may yield causal results, but these results may not be applicable to other contexts (Gröndahl et al., 2018), and a small sample size overall or regarding a subset might hinder valuable and definitive insights (Muralikumar et al., 2023; Mahomed et al.; 2024).

The approaches employed by the third group do not achieve causality, as their user surveys create only a subjective observational picture of stated user experiences.

It can be concluded that, in order to achieve causality, the algorithm is fed with input and an adequate strategy for isolating and identifying the causal effects of interest is in place.

4.1.2.3 Platform support

A notable commonality that is observed in all but one of the approaches that achieve causality is that they are either facilitated by platform support, utilise their own model, or employ publicly available models such as LLMs or classification APIs, through which content can be injected without others being confronted with this content. In the context of content moderation systems on online platforms, adopting these approaches without the support of the platform itself could result in the injection of harmful or illegal content into the regular system. This poses significant *legal and ethical challenges*. The approach of Shukla et al. (2025) employs the actual system for injecting harmful content, interacting with an

available API, but taking advantage of the possibility to silo the stream from other users and a legal ruling allowing this research activity. Consequently, the adoption of such an approach with a private upload environment, perhaps not the most harmful content, and the conducting of an in-depth consideration of potential legal and ethical challenges might enable the establishment of causality regarding the content moderation system of a platform without their support. However, in many cases, this might not be the case. The presence of platform support, for instance, when it can be requested by a regulator due to sufficient existing evidence, appears to be a critical factor in the auditing of content moderation algorithms.

4.1.2.4 Data collection method

Another important decision that differentiates the approaches is the chosen data collection method for conducting the audit with the previously chosen type of objective. Analyses in the first group are based on using APIs (Wang et al., 2025; van Boheemen et al., 2025; Pierri et al., 2023; Murthy, 2021; Dreier and Pirker, 2023), crowdsourcing with a browser extension (Juneja et al., 2023), sock-puppet accounts manually managed or managed by a bot (Eltaher et al., 2025; Hussein et al., 2020) and using a combination of scraping and sock-puppet accounts managed by bots (Juneja and Mitra, 2021). However, it should be noted that the opportunities afforded by sock-puppet accounts and crowdsourcing, in comparison to conventional scraping techniques, are primarily utilised for the subsequent analyses, which extend beyond the examination of the moderation system, encompassing the impact of the recommendation system. In the second group, analyses are based on model sampling alone (AlDahoul et al., 2024; AlDahoul et al., 2023; Sap et al., 2019; Borkan et al., 2019; Nogara et al., 2025; Mahomed et al.; 2024; Hosseini et al., 2017), model sampling combined with a user survey (Muralikumar et al., 2023), combined with model fine-tuning, respectively modifying model internals depending on the intensity of for the analyses conducted training (Gomez et al., 2024; Nguyen et al., 2023; Del Vigna et al., 2017; Gutfeter et al., 2023; Vitorino et al., 2018; Dixon et al., 2018; Gröndahl et al., 2018) and combined with the modification of model internals, i.e. the classification threshold (Kumar et al., 2024; Shukla et al., 2025). Approaches in the third group are based on user surveys (Myers West, 2018; Lahti et al., 2024; Smahel et al., 2020; Haimson et al., 2021) and on an end-user approach (Lam et al., 2022). Consequently, a wide range of data collection methods is represented, with the first group focusing on a variety of methods for collecting data, the second group focusing on methods listed under platform support and the third group using user-centred methods such as a survey or an end-user approach.

However, the selection of data collection method appears to be contingent on the type of content being examined, with certain combinations of content type and data collection method being either unfeasible or necessitating specific precautions. For instance, Wang et al. (2025) note the secure storage of the collected content promoting illicit goods and services as part of their approach, while the papers collecting observational data on misinformation do not mention such a precaution, likely due to a lower potential harmfulness. Should storing data in a secure manner not suffice to guarantee safety, hashes for the content could be created during data collection, such that only the hash values, not the content itself, are stored, thus allowing for the identification of exact duplicates or very similar content (Boshmaf et al., 2023). Nevertheless, this would considerably restrict other potential avenues of analysis. Utilising sock puppets or crowdsourcing involves safety considerations with regard to preventing the behaviour of the sock puppets from having a detrimental effect on other users or the user whose account is employed for the crowdsourcing (Juneja et al., 2023; Juneja and Mitra, 2021). For instance, the crowdsourcing browser extension can be set to run in the background, with the history of watched content being deleted to avoid users being confronted with raised levels of misinformation in the future (Juneja et al., 2023). Alternatively, to avoid potential effects on real user accounts being used, sock puppet accounts can be used instead, with a limited range of potential behaviours to avoid strong spillover effects on other users,

as might for instance be the case through a positive review written for a misinformation product (Juneja and Mitra, 2021). It is reasonable to assume that for specific categories of illicit or harmful content, even minor such spillover effects are considered unacceptable. Concerning the labelling of toxic content by crowd workers, Muralikumar et al. (2023) incorporate a trigger warning; nevertheless, for specific content types, the respondents may incur excessive harm even with such a warning. Even for the researchers themselves, detailed work with the content might be too disturbing; for instance, Gomez et al. (2024) minimised direct exposure to content by redacting certain parts and analysing the data in aggregate.

4.1.2.5 Metrics and indicators

With the exception of Myers West (2018), who employs exclusively a descriptive analysis of qualitative results, all other papers collect the data for analysing some chosen quantitative metrics and indicators, sometimes supplemented by a qualitative analysis. For the approaches that rely on observational data, this includes the prevalence of illegal or harmful content, and descriptive statistics related to the types of harm categories (Wang et al., 2025; van Boheemen et al., 2025; Pierri et al., 2023; Dreier and Pirker, 2023). Papers inputting data into algorithms analyse performance metrics such as the accuracy, precision, recall or F1 score (Kumar et al., 2024; Nguyen et al., 2023; AlDahoul et al., 2024; Del Vigna et al., 2017; AlDahoul et al., 2023; Gutfeter et al., 2023; Vitorino et al., 2018; Shukla et al., 2025), the effect on such metrics and the classification score resulting from adversarial attacks (Gröndahl et al., 2018; Shukla et al., 2025; Hosseini et al., 2017), pairwise disagreement and arbitrariness metrics (Gomez et al., 2024) and Krippendorff's alpha to measure human-algorithm alignment (Muralikumar et al., 2023). With regard to the input analyses regarding biased performance, metrics such as accuracy differentiated by content group (Sap et al., 2019), metrics combining performance and fairness, such as the average equality gap or pinned AUC equality difference (Dixon et al., 2018; Borkan et al., 2019), and differences in distribution of toxicity score for different content groups (Nogara et al., 2025), are examined. The user-centred approaches analyse indicators such as the frequency of encountering illegal content (Lahti et al., 2024; Smahel et al., 2020), the proportion of respondents having had content removed and regression coefficients regarding the connection with certain characteristics (Haimson et al., 2021) and the proportion of disagreement with the algorithm combined with qualitative indicators such as topic and type of problem (e.g. over- or underreporting) (Lam et al., 2022).

4.1.2.6 Implementation strategy and challenges

However, the accurate identification of these metrics and indicators poses significant challenges and necessitates an appropriate implementation strategy. Depending on whether observational data is collected from the system, the model is probed or altered, or users are consulted, different groups of challenges arise when collecting and analysing the data.

A common challenge for approaches that collect observational data from the actual system lies in determining *where to search* for this data. Limited data access and the necessity to input search terms in the chosen data collection method might result in the necessity to create a list of relevant search terms (Wang et al., 2025; Pierri et al., 2023). Moreover, a low prevalence of the regarded type of illegal content might require a refined way of searching in order to identify a relevant fraction of this content (Murthy, 2021). Analysing all content present on the system would create the most complete picture and lead to the highest representativeness, but this approach requires an infeasibly high workload in most cases. Instead, the envisioned strategy might lead to the identification of as much illegal content as possible with the given resources, or the creation of a representative picture of some part of the system. For instance, Wang et al. (2025) begin with known tags from an existing dataset concerning the promotion

of illicit goods, which are then updated with tags and accounts based on the illegal content identified in their research, thus resulting in a diverse set of keywords. Furthermore, they conduct an analysis of the proportion of illegal content identified in a random sample of general content. Van Boheemen et al. (2025) employ a similar, more manual strategy, utilising relevant tags associated with far-right accounts, algorithmic snowballing and further online sources for relevant accounts. Alternatively, the search can be based on existing keywords relating to the studied topic, in this case the Russian invasion of Ukraine, without updating (Pierri et al., 2023), a combination of Google Trends and known relevant video tags relating to electoral misinformation (Juneja et al., 2023), a combination of high-impact topics and search terms concerning vaccine- respectively misinformation-related topics identified through Google Trends and relevant search terms based on the system's auto-complete (Hussein et al., 2020; Juneja and Mitra, 2021), known seed videos of terrorist content to gather further content through recommendations (Murthy, 2021) or a diverse set of general accounts with a focus on more popular ones to analyse the toxicity in their chat messages (Dreier and Pirker, 2023). Evidently, the adequacy of a strategy depends on the overarching objective; for instance, the extent to which results should be influenced by the recommendation system, how extensive the search should be, or the extent to which current developments should be analysed. It can be observed that the analysis of misinformation is the only field in which trends data is being used, likely due to the difficulty of obtaining such data for illegal content, and because misinformation is more closely associated with topics that are not generally considered harmful than this might be the case for other harmful content. Should sock puppets be employed, their behaviour, including potential searches and interactions with content which influence which content will be shown to them, must be defined in accordance with the situation their behaviour is intended to imitate (Eltaher et al., 2025).

A second challenge that arises with the analysis of observational data is the *classification* of the collected content. A range of approaches can be identified. The paper by Pierri et al. (2023) utilises existing lists of news websites and their reliability and political leaning to completely automatically classify Tweets in as propaganda and misinformation. Dreier and Pirker (2023) employ NLP techniques to identify toxicity in chat messages. Van Boheemen et al. (2025) employ a combination of automatic and manual classification techniques for the identification of harmful video posts by, for instance, automatically screening transcripts and comments for predefined words and emojis, and manually analysing selected recordings due to the fact that harmful speech is frequently encoded and hidden in everyday language, making automatic detection more challenging. Other studies commence with human labelling in the context of posts promoting illicit goods, video posts containing electoral misinformation, and terrorist videos, either entirely conducted by researchers or in collaboration with crowd workers, leveraging this dataset to automate classification, for instance, employing a machine-learning approach (Wang et al., 2025; Murthy, 2021; Juneja et al., 2023). The remaining approaches label content regarding harmful video posts, misinformative video posts and misinformative health products completely manually (Hussein et al., 2020; Juneja and Mitra, 2021; Eltaher et al., 2025). This is for instance achieved by the utilisation of labels from a second researcher if content has been classified as harmful by an initial researcher and the inclusion of further researchers in the event of disagreement, given that certain aspects of classifying content as harmful might be subjective (Eltaher et al., 2025), or by starting labelling by a researcher and finishing it using multiple crowd workers, and the researcher deciding in case of disagreement among the crowd workers (Juneja and Mitra, 2021). The feasibility of the different methods will naturally depend on various factors, including the type of content, the accuracy with which novices might label content, the potential consequences of presenting content to crowd workers, the volume of collected content, and the presence of any consistent patterns that could inform the work of crowd workers or a classification algorithm. In general, it should be borne in mind that any illegal data observed on the platform has likely already been moderated, such that it constitutes false negatives which might be much harder to identify.

A common challenge that arises when the model is probed, sometimes after an alteration to it has been conducted, is the *obtaining of the adequate data* that is needed for this. A significant number of papers employ existing suited datasets (Gomez et al., 2024; Gröndahl et al., 2018; Vitorino et al., 2018; Shukla et al., 2025; Sap et al., 2019; Hosseini et al., 2017; Lam et al., 2022; Gutfeter et al., 2023). With the exception of Gutfeter et al. (2023) and Vitorino et al. (2018), all of them pertain to hate speech or toxicity. However, as these two papers relate to child sexual abuse material, they cannot obtain simple access to existing data related to this topic, as is the case with the general pornographic data they are using. Nevertheless, in collaboration with a public agency, they are able to gain access to it in a safe environment. However, as the illegal content is not visible to the auditors, finetuning of the classification algorithm and the training data set is more challenging (Gutfeter et al., 2023). Furthermore, a significant proportion of papers utilize newly collected or created datasets in conjunction with existing data (Kumar et al., 2024; Nguyen et al., 2023; AlDahoul et al., 2024; Dixon et al., 2018; Borkan et al., 2019; Nogara et al., 2025). For instance, AlDahoul et al. (2024) draw upon existing datasets for topics such as hate speech, adult content, and violence, while acquiring new data for topics including graphic violence and alcohol, child, and drug abuse. Nogara et al. (2025) supplement the extant hate speech datasets with a randomly collected corpus of Wikipedia data, thereby obtaining a text corpus that is highly similar between different languages and Nguyen et al. (2023) add a manually collected dataset relating to sexual predatory chats in an additional, less common language to existing ones. Kumar et al. (2024) supplement the existing hate speech data with less selective real-time data to obtain a more realistic data example and Dixon et al. (2018) create a synthetic dataset based on an existing dataset to create a more controlled environment. Other studies collect the relevant data completely themselves (Del Vigna et al., 2017; Muralikumar et al., 2023; Mahomed et al.; 2024), for instance because hate speech of a certain less commonly used language is of interest. This data is, for instance, crawled using an API and annotated by bachelor students, though this results in a low inter-annotator agreement (Del Vigna et al., 2017), or manually collected and annotated based on the reasoning that no API was available (Muralikumar et al., 2023). Moreover, studies probing the model with a focus on user-related characteristics encounter the challenge of obtaining data relating to information such as the identity referenced. This data might be collected during the initial data collection process, for instance labelled by crowd workers (Sap et al., 2019), but if an existing dataset without this information is used, it would need to be added later on, for instance by approximating identity based on dialect used or certain identity terms present (Sap et al., 2019; Borkan et al., 2019). Therefore, depending on the type of content, for instance whether it is illegal, that is analysed, the desired level of specificity, the desired size, or the desired level of control over the content, the necessary input is more or less challenging to obtain.

However, it should be acknowledged that in the case of an external auditor having access to the aggregate of the content moderation algorithms of a platform, a slightly different situation regarding input data would take place. Firstly, depending on whether access is granted to specific algorithms checking for specific violations or to a broader part checking for multiple types of violation, the utilised true negative content cannot violate any of the prohibitions. Consequently, it may be necessary for the content to be in accordance with all the terms of service, which could result in the classification of content becoming more complex, potentially necessitating expert judgement or sophisticated classification algorithms, given the multitude of nuanced rules that vary between platforms. Secondly, depending on the level of access, it might not be known on which data the algorithm has been trained. However, the performance on known content is likely to differ from the performance on new content, which more closely mirrors the real-life situation. Hence, it would need to be ensured that a sufficient fraction of the content with which the algorithm is probed is different from the training data, for instance by collecting or creating data oneself. However, the incorporation of a small proportion of content that may be known to the algorithm, for instance from a hash database, could provide insights into whether the reupload of known content is sufficiently countered.

The approaches consulting users appear to be mainly confronted with challenges related to surveys in general, such as a bias in who might answer surveys (Myers West, 2018; Haimson et al., 2021), the potential of manipulated or inaccurate answers (Myers West, 2018; Haimson et al., 2021), biases in answers (Muralikumar et al., 2023) and difficulties in obtaining exactly the information required (Muralikumar et al., 2023; Smahel et al., 2020).

A general challenge that may be encountered, depending on the level of access to the algorithm, is that the content moderation process is not solely algorithm-based, but also involves human labour. As previously outlined, the content moderation system may categorise certain content not as either permitted or prohibited, but as necessitating human evaluation, and user reports as well as human post-moderation further influence what content is available. Therefore, depending on the level of access and the part of the moderation that is of interest, strategies for removing the effects of components that are not to be analysed must be designed, or it must be acknowledged that the results relate not only to the moderation component of main interest, but to an interplay of several components.

Following the collection of data in accordance with the decisions made in the preceding steps, the data is then prepared, analysed and reported.

4.2 Self-preferencing in recommendations

4.2.1 Algorithmic recommender systems

Recommendation systems are information retrieval and filtering tools that assist users in determining items and options that are likely to be of interest. Depending on the specific circumstances, a variety of inputs can be utilised, and a range of outputs can be generated, including ranked lists of varying lengths or single-item recommendations. Stray (2022) emphasises that recommendation systems do not operate with a single objective; rather, platforms often pursue a combination of goals, such as relevance, diversity, or healthier patterns of engagement. In consideration of input from the content moderation system and the context, such as search conducted or past user behaviour, systems typically first generate a set of candidate items. Subsequently, the relevance of the item is estimated within the context provided, frequently assigning a probability or score that reflects predicted user engagement. However, broader social or civic considerations may also be incorporated. Finally, a re-ranking step can be employed to adjust the list and further emphasise the values that the platform wishes to promote, such as exposing users to a wider range of perspectives. This process can thus be viewed as a series of decision points at which different goals can be embedded.

The majority of extant literature categorises recommendation systems into three main types: content-based filtering, collaborative filtering and hybrid systems. Content-based filtering can be defined as a recommendation system that functions by suggesting items of a similar nature to those previously selected by the user. This constitutes the basic model of recommendation. Collaborative filtering analyses data from multiple users to identify patterns of co-preferences, and uses these patterns to make recommendations. Collaborative filtering is known to encounter issues such as scalability and sparsity. Consequently, recent research has focused on hybrid filtering systems that address these challenges, resulting in significant advancements and heightened research interest.

4.2.2 Analysis of empirical literature

The second application domain relates to the gathering of indications as to whether a platform with incentives and possibilities to self-preference does so in their recommendation algorithms, for instance regarding listed recommendations or in single-item recommendations. In this context, self-preferencing involves the recommendation of specific items or offers, such as the platform's own products or the delivery of third-party products by the platform itself, with greater prominence without a valid explanation, but due to anti-competitiveness, with the objective of generating profit. The identification of self-preferencing is, in general, a challenging undertaking, for instance due to the definition of a neutral alternative, legitimate business interests, or the possible ambiguity regarding used variables being objective or subjective, often requiring a case-by-case analysis (Dash et al., 2024; Carugati, 2022). The following section delineates and categorises the approaches and steps of existing empirical analyses with regard to this problem.

4.2.2.1 Methodological approach

As in the other application domain, a significant difference between the approaches lies in the chosen objective and the nature of the insights to be generated. In this regard, we propose a categorisation of the approaches into two groups that are directly related to the problem, and one group that is indirectly related. The first group uses observational data from the platform and analyses it for indications of self-preferencing based on observable characteristics. It encompasses several analyses of different recommendation structures on Amazon regarding the visibility, either the assigned rank or being included at all, of Amazon's own private-label products in comparison to similar third-party products, such as of the sponsored related item recommendations (Dash et al., 2021), of the similar item to consider recommendations (Dendorfer and Seibel, 2025), of search recommendations in general (Farronato et al., 2023), and of the search recommendations regarding changes in the relative ranking of these private-label products after the designation as a gatekeeper under the DMA (Waldfogel, 2024). Jürgensmeier and Skiera (2023) analyse the search results in the aforementioned general regard, as well as regarding the ranking of products concerning whether Amazon or a third party holds the buy box, meaning being the primary seller offer shown, for this product. Further analyses concern Amazon's buy box selection regarding products sold and/or shipped by Amazon compared to the identical products sold and shipped by third-party sellers (Hartzell and Haupt, 2025; Hartzell, 2025; Raval, 2022; Lee and Musolff, 2025; Mouton and Rottembourg, 2024), Amazon's frequently bought together recommendations regarding the presence of products sold and shipped by Amazon compared to identical products sold and shipped by third-party sellers (Chen and Tsai, 2024), and Amazon's Kindle Daily Deal page regarding the ranking of titles published by Amazon Publishing compared to similar titles (Reimers and Waldfogel, 2023). Analyses of other platforms examine Apple's App Store ranking in search results regarding Apple's own apps compared to similar apps (Teng, 2022), Netflix's top 10 recommendations regarding the presence of its original content compared to similar content (Leung et al., 2025) and Kayak's search results ranking regarding whether hotels for which its through vertical integration affiliated online travel agent, Booking.com, is price leader are ranked higher than similar hotels, as well as Kayak's sales channels recommendation regarding whether Booking.com is more likely to be displayed than similar offers (Cure et al., 2022). Not directly related to self-preferencing, but similar in structure and therefore being included in the first group, is the examination of the ranking in search results on Booking.com respectively Expedia regarding hotels which set lower prices on a different website than Booking.com respectively Expedia, compared to those who do not, closely being related to the in some countries forbidden price parity clause (Hunold et al., 2020).

The second group leverages interactions with users to gather more information regarding their actual preferences, which is then used to control for this information when examining for self-preferencing. Farronato et al. (2025) analyse self-preferencing of Amazon's own products in rankings using elicited consumer preferences and observed consumer behaviour under different controlled circumstances. Dash et al. (2024) analyse self-preferencing by Amazon regarding the buy box assignment and Amazon's policy of striking through some reviews for products being fulfilled by them using elicited consumer preferences and effects on consumers.

Furthermore, although not analysing self-preferencing, but related topics, some papers directly probe recommendation algorithms and alter them, analysing the fairness of the distribution of results concerning certain content characteristics. As this could be employed in more or less similar ways regarding self-preferencing with sufficient access, and as it constitutes a noticeably different approach than the first two groups, these papers are regarded as the third group with the aim of generating insights that are relevant for the domain of self-preferencing. Given the unfeasibility of third parties inputting content that mimics first-party content on the real system, and the low incentives for a first party to publish results that might point to self-preferencing, it appears reasonable that very little literature employing this type of approach directly relating to self-preferencing can be found. The third group encompasses analyses of recommendations on an online marketplace platform concerning seller-side fairness, which pertains to all sellers receiving a minimum baseline level of visibility (Ye et al., 2023), of recommendations of books and education courses concerning preference distribution-aware provider fairness, which relates to providers receiving at least some level of visibility while taking into account user preferences when distributing this visibility (Gómez et al., 2025), of recommendations of a large-scale, production recommender system concerning the under-recommendation of items from a certain group conditional on anticipated engagement (Beutel et al., 2019), and of LinkedIn's recruiter recommendations of potential job candidates concerning fair representation regarding sensitive attributes (Geyik et al., 2019). Three of these papers are written with the assistance of platform affiliates, and the remaining paper employs a by them newly developed algorithm in addition to existing algorithmic approaches, a distinction from the other two groups, whose analyses are all conducted in a black-box manner.

4.2.2.2 Model setup

A further discernible foundational difference between the approaches lies in whether models based on certain assumptions are set up for analysing the setting, thereby influencing the further procedure. Several papers in the first and second group set up demand and supply models (Hartzell and Haupt, 2025; Hartzell, 2025; Lee and Musolff, 2025; Reimers and Waldfogel, 2023; Teng, 2022; Hunold et al., 2020; Farronato et al., 2025). The supply models analyse, for instance, the price setting of products, the entry of a supplier into a market or the update frequency of apps. Other papers in the first group do not set up a complex model but analyse regressions, for example of rank or of presence in certain recommendation structures on product characteristics and further controls (Dendorfer and Seibel, 2025; Farronato et al., 2023; Waldfogel, 2024; Jürgensmeier and Skiera, 2023; Raval, 2022; Chen and Tsai, 2024; Leung et al., 2025; Cure et al., 2022), one being machine-learning based (Mouton and Rottembourg, 2024). In relation to these approaches, it should be noted that some critique has been raised regarding the inadequacy of employing a linear regression of rank on relevant independent variables, abstracting away from the intermediate step of generating scores used for creating the ranking, because the coefficients and standard errors may be biased (Ford, 2023; McKelvey and Zavoina, 1975). The only paper in the first group that does not fall into these two categories is Dash et al. (2021), which analyses characteristics of a related-item network; in the second group, Dash et al. (2024) focus rather on descriptive statistics regarding their collected data and user survey results, even though the paper also used a structured machine-learning-based model approach to analyse important factors in consumer choices. Approaches

in the third group focus on evaluating their chosen fairness metrics in a non-model-based manner, though formally formulating the recommendation problem prior.

4.2.2.3 Metrics and indicators

Connected to these differences is the selection of metrics and indicators to be analysed. Frequently, this involves an analysis of regression coefficients and their size and significance (e.g., Farronato et al., 2023; Jürgensmeier and Skiera, 2023; Reimers and Waldfogel, 2023). For the papers setting up a welfare model, this also includes components such as consumer surplus or overall welfare (e.g., Hartzell, 2025; Teng, 2022; Farronato et al., 2025). In the context of user-centred approaches, this encompasses additional components, including estimates of willingness to pay, reported consumer satisfaction levels, stated most influential choice factors, and the impact of experimental conditions on behaviour (Dash et al., 2024; Farronato et al., 2025). Further metrics include measures of representation or exposure in a recommendation network (Dash et al., 2021), or in the third group, fairness measures, recommendation performance and business outcomes when different algorithms are directly compares (Gómez et al., 2025; Beutel et al., 2019; Geyik et al., 2019; Ye et al., 2023).

4.2.2.4 Data collection method

In order to conduct the chosen analysis, a number of approaches are taken to obtain the necessary data. Some papers in the first group collect their observational data completely from a third party, such as from an API regarding search result rankings on Amazon (Waldfogel, 2024), a price-tracker API for products on Amazon (Raval, 2022; Mouton and Rottembourg, 2024), a high-volume third-party seller on Amazon in combination with data from a commercial third-party analytics provider (Hartzell and Haupt, 2025; Hartzell, 2025) or app store optimization service providers (Teng, 2022). Other papers combine data obtained from third parties with data from other sources, such as data from a third party offering repricing services for Amazon, in combination with estimates from existing research (Lee and Musolff, 2025), data from a third-party collecting device-level viewing behaviour data and from third-party APIs, in combination with own collected data regarding Netflix (Leung et al., 2025), data from a commercial third-party data provider regarding book sales, combined with own collected data on Amazon (Reimers and Waldfogel, 2023), data from a third-party API price tracker for Amazon products, supplemented with wider ranging scraped data (Chen and Tsai, 2024), or supplemented with data on visibility in search results from a search engine optimization firm and a user survey (Jürgensmeier and Skiera, 2023). The remaining papers collect the data completely by themselves, scraping it without an account or using a single sock-puppet account (Dash et al., 2021; Dendorfer and Seibel, 2025; Cure et al., 2022; Hunold et al., 2020) or obtaining the data using crowdsourcing with a browser extension (Farronato et al., 2023).

With regard to the second group, one paper collects data through scraping using a single sock-puppet account, conducting a sociotechnical experiment on the influence of Amazon's policy of striking through some reviews for products that are fulfilled by them on consumer choices, and conducting user surveys, in which participants indicate whether they would choose Amazon-related products that have been assigned the buy box even though they are not the lowest priced and which product characteristics are important influencers of choice (Dash et al., 2024). The second paper employs a combination of crowdsourcing and a sociotechnical field experiment, wherein Amazon's private label products are hidden, alongside an incentivised shopping task and a user survey that investigates factors such as willingness to pay and significant choice influencers (Farronato et al., 2025).

The third group's data is derived from model sampling and advanced access to model modification, utilising proprietary data in conjunction with existing public data (Ye et al., 2023), exclusively existing datasets (Gómez et al., 2025), synthetic data in conjunction with employment in the actual environment with actual user interactions and data (Geyik et al., 2019), and exclusively actual data from the real environment (Beutel et al., 2019).

Therefore, in the first group, a broad reliance on third-party data is evident. The acquisition of data pertaining to marketplaces serves the commercial interests of third-party sellers, thus explaining the wide availability of such data from third-party providers. However, it should be noted that such APIs are not universally available across all online platforms (Mouton and Rottembourg, 2024). The second group places a greater emphasis on the execution of user surveys and experiments, while the third group employs its access to the model primarily in conjunction with existing datasets, sometimes of a proprietary nature.

When choosing to scrape data or query an API, the decision which content to regard needs to be made. Approaches to this issue include the examination of the most popular products, categories and/or keywords (Dash et al., 2024; Waldfogel, 2024; Mouton and Rottembourg, 2024; Chen and Tsai, 2024; Teng, 2022), of a combination of highly popular and less popular products (Jürgensmeier and Skiera, 2023; Raval, 2022; Hunold et al., 2020), of categories with a higher presence of the party potentially self-preferencing (Dash et al., 2021; Dendorfer and Seibel, 2025), of as much of the inventory as possible (Leung et al., 2025), and of a specific, arbitrary situation, in this case a single city and a chosen request (Cure et al., 2022).

4.2.2.5 Implementation and challenges

A common issue identified in studies of the first and second group regarding implementation is the *unavailability and inaccuracy of data*, which hinders the envisaged analysis due to the necessity of a substantial depth of data. Generally, the objective of the studies is to analyse whether products that differ only in a relevant attribute, such as seller identity or brand, and are equally attractive to consumers, are treated differently, thereby constituting self-preferencing. In order to achieve this objective, it is necessary to collect information either regarding consumer preferences directly or regarding all factors that may have a significant influence on these preferences. However, factors relating to the platform's knowledge of information that is not accessible to researchers hinder this latter analysis. These include the unavailability of certain potentially influential factors, such as a seller rating of Amazon, inaccuracy when scraping data, such as the failure to clearly identify all sponsored results, and the necessity to use approximations for certain data, such as approximating sales using sales rank. Some papers partially alleviate this by cooperating with a third party. To this end, three studies gather insights into actual sales data by obtaining data from a past seller, respectively by obtaining data from a third-party having access to their customers' sales data (Hartzell and Haupt, 2025; Hartzell, 2025; Lee and Musolff, 2025). With regard to the equivalent data for hotel bookings, information on past bookings of hotels was stated directly on one platform (Hunold et al., 2020). Consequently, in the absence of external support from a third party or the platform, or exceptional circumstances, acquiring outcome data, such as sales or bookings, is challenging in the majority of cases. However, even with access to a large variety of relevant platform data, there might be still additional characteristics that could influence consumer preferences, which could be overcome by obtaining data on this from users. However, data concerning actual consumer preferences or behaviour is scarce in the regarded approaches. In order to gather information on user behaviour and thereby ascertain their preferences, Lee and Musolff (2025) estimate arrivals at product pages using the sales data obtained from a third party, and Hartzell and Haupt (2025), Hartzell (2025), Teng (2022) and Leung et al. (2025) obtain third-party data on product views, consumer

conversion and watch data. Eliciting consumer preferences in a more direct manner, Dash et al. (2024) conduct a user survey to assess the alignment of algorithmic decisions with user preferences. As Farronato et al. (2025) employ crowdsourcing alongside an experiment and incentivised choice tasks, they are able to analyse consumer behaviour and preferences in detail. Furthermore, the utilisation of random variation in ranks, stemming from the experimental setup, facilitates the identification of unobserved characteristics that influence preferences, which are reflected in the assigned rank, independent of the effect of higher ranks being more likely to be selected due to increased visibility. However, it should be borne in mind that the choice architecture created by the platform may influence consumer preferences, thereby altering the "true" preferences, and thus rendering reliable collection more complex (Dash et al., 2024). Furthermore, in general, consumer choices in an artificial environment might not accurately reflect real preferences, especially without incentivisation. Additionally, the pool of respondents might not be representative for the actual user base, for instance because only a subgroup is willing to have a browser extension track their behaviour.

The challenge of obtaining all necessary data is also influenced by the type of self-preferencing that is regarded as this influences the scope of characteristics for which one needs to control. To illustrate, when comparing identical products sold by different sellers, there is a reduced need for controls when compared to the sale of different products by different sellers. Furthermore, three papers utilise an outcome-based approach, in which sales data is regressed on the treatment of the platform, for instance the rank given to the product, and on further control characteristics, in order to analyse whether outcomes differ for products that are ranked similarly and therefore should be equally successful (Dendorfer and Seibel, 2025; Jürgensmeier and Skiera, 2023; Reimers and Waldfogel, 2023). This is undertaken to minimise reliance on information regarding characteristics that are relevant to consumer preferences, as these are implicitly accounted for through the outcome data. However, in return, information on attributes that correlate with a product's success is required.

The third group faces noticeably smaller difficulties in obtaining the data due to their ability to access proprietary data, as the analyses are conducted by the platforms themselves, or they can use general, existing data to make their point.

4.2.2.6 Causality

This challenge of not observing all the required data hinders the ability to make causal claims in many papers, especially in the first group. Nevertheless, this is fundamentally the ultimate objective when it comes to analysing self-preferencing. However, certain studies have been able to establish causal claims regarding self-preferencing and related practices, albeit partly based on assumptions that may potentially limit their accuracy. Within the first group, one paper notably approaches true causal results more closely than the others, while two additional papers distinguish themselves as well, though with a greater distance from causality. Chen and Tsai (2024) employ exogenously judged within-product variation, generated by Amazon stockouts, to analyse whether the identical product sold by a third-party seller receives fewer recommendations. Certain assumptions must still be met, such as the stockout being exogenous, though this is more likely to be fulfilled than the assumptions that would need to hold in regard to the other studies. Moreover, indications for the assumption holding are collected, such as by conducting a placebo test with a third-party seller stockout, controlling for indicators which might be related to potential changes in consumer preferences and analysing the development of prices and sales regarding smoothness prior to the stockout. Teng (2022) leverages a change in Apple's recommendation algorithm that downshifts the ranking of its own apps for a difference-in-difference design to analyse the effects on independent apps competing against Apple versus the independent apps in categories not competing with Apple, for instance regarding changes in downloads. However, for the

purpose of analysing self-preferencing, a model with several, more complex assumptions is designed. Waldfogel (2024) examines a change in Amazon's recommendation algorithm subsequent to its designation as a gatekeeper under the DMA, analysing its impact on the ranking of Amazon's own products. In the absence of a change in the products' market appeal, this could be interpreted as a change in the extent of self-preferencing. In order to analyse the specificity to Amazon, a comparison is drawn between the effect on Amazon brands and the effect on other major brands. However, for both of these studies, it is challenging to make a causal assessment of the absolute extent of self-preferencing present before and after the change, and not only in terms of the relative change.

In the second group, Farronato et al. (2025) establish a causal relationship with their approach. One building block for this is the removal of Amazon's brands in a field experiment with a control group for which no such change occurs and a second comparison group for which the same amount of products, but a random selection, are hidden to control for the effect that might be induced by the hiding in general, not being specific to Amazon's brands. One building block for this is the removal of Amazon's brands in a field experiment with a control group for which no such change occurs and a second comparison group for which the same amount of products, but a random selection, are hidden to control for the effect that might be induced by the hiding in general, not being specific to Amazon's brands. The other building block is an incentivised shopping task within the three experimental groups to achieve greater statistical power regarding the estimation of consumer preferences, as much heterogeneity therein is present. This enables a causal analysis of changes in behaviour and satisfaction when Amazon's brands are hidden, as well as a precise analysis of consumer preferences for the categories and respondents they are regarding. Consequently, their estimates of current self-preferencing are more likely to withstand the causality test. However, their subsequent welfare model is, by its very nature, an abstraction, for instance, by focusing exclusively on short-term effects. Dash et al. (2024) seek to advance towards a causal understanding by examining the impact of Amazon's strike-through policy on consumer choices, and whether user preferences indicated in user surveys correspond with algorithm choices. However, the survey choices may not adequately reflect real choices and are not incentivised, the respondent pool is relatively small and statistical significance regarding differences is not regarded. Consequently, the capacity to formulate definitive causal statements is to a certain extent constrained in this study.

All papers in the third group achieve causality by directly interacting with the algorithm and having access to data with all the required and correct data. All of them compare the performance of different algorithms while maintaining a constant testing environment, including the test dataset. In addition, some approaches apply the same algorithms to different datasets, thereby analysing the effect of dataset (Gómez et al., 2025; Geyik et al., 2019; Ye et al., 2023). Geyik et al. (2019) utilise a simulated dataset for the evaluation, enabling the analysis of a broader range of potential ranking scenarios and eliminating position bias. Additionally, an alteration of the algorithm is implemented in the actual environment, applied randomly to 50% of users in an A/B test, in order to assess the impact in a real-world setting.

4.2.2.7 Robustness

Nevertheless, it is evident that papers that do not attain causal claims are attempting to enhance their results' robustness to a certain extent by conducting additional analysis and tests. Approaches for this include testing the quality of the estimated model by comparing its predictions to the truth, in the best case for data on which the model has not been trained on (Hartzell and Haupt, 2025; Raval, 2022; Lee and Musolff, 2025; Mouton and Rottembourg, 2024), by analysing the likelihood of other potential explanations such as reverse causality (Hunold et al., 2020), by conducting a placebo test (Cure et al., 2022) and by conducting a plausibility test (Dendorfer and Seibel, 2025). Moreover, certain studies have analysed the stability of the results by removing specific data points, which could potentially introduce

bias, in an additional analysis (e.g., Farronato et al., 2023; Jürgensmeier and Skiera, 2023; Leung et al., 2025) and by examining different categories, countries, or time periods separately to identify whether results are influenced by outliers (e.g., Dash et al., 2024; Jürgensmeier and Skiera, 2023; Leung et al., 2025). Furthermore, the sensitivity to the assumptions made is examined by analysing different specifications (e.g., Dash et al., 2024; Dendorfer and Seibel, 2025; Farronato et al., 2023) and conducting a specification curve analysis, which examines the way in which varying assumptions influence the results (Jürgensmeier and Skiera, 2023).

Finally, the results obtained from the implementation of the chosen strategy and analysis are reported.

5 Generalized auditing process

Utilising the insights derived from the two application domains concerning important decision points, considerations, and challenges, as well as the commonalities and differences between different approaches, a generic flow for the auditing process has been formulated. This flow can be followed when applying algorithmic auditing under different circumstances in practice. It is enriched with existing research in this topic and triangulated with applications from other application domains at certain points.

5.1 Setting up

As indicated in the literature review, several pre-considerations should be made prior to commencing the auditing process as outlined in the application domains. This process constitutes step zero of our framework: the setting up. This encompasses the fundamental task of selecting an algorithm to be audited and a research topic, respectively a risk, for which the algorithm should be audited (Metaxa et al., 2021; Panigutti et al., 2025). As has been identified in the literature review, several frameworks with a broader focus develop risk-based scenarios based on an in-depth understanding of the platform, its stakeholders and the general circumstances (Hasan et al., 2022; Meßmer and Degeling, 2023; Zicari et al., 2022). These scenarios are subsequently prioritised, if necessary due to limited resources. In this context, it may be advantageous to specify concrete, testable scenarios that define the affected party with its characteristics, the harm, the involved elements of the platform, for instance the specific algorithms of the algorithmic system, and the further impacts (Meßmer and Degeling, 2023). When analysing the effect of the algorithm on society or certain individuals, research is needed to ensure a proper understanding of any effects beyond the direct interplay between system and user (Meßmer and Degeling, 2023). In general, the socio-technical interplay must be given due consideration (Hasan et al., 2022; Meßmer and Degeling, 2023). It is frequently emphasised that a diverse set of stakeholders should be involved in the auditing process and in exchange with each other (Metaxa et al., 2021; Hasan et al., 2022; Meßmer and Degeling, 2023; Zicari et al., 2022; Ojewale et al., 2024; Whittlestone et al., 2019; Costanza-Chock et al., 2022). For instance, in order to identify which technological constraints are currently in place and which are not, it may be beneficial to consult several technological experts, whose expertise is generally highly valuable for conducting the audit (Meßmer and Degeling, 2023; Zicari et al., 2022; Whittlestone et al., 2019). Researchers with a background in risk, ethics, psychology, or a related field are valuable contributors to the identification and comprehension of risk scenarios (Meßmer and Degeling, 2023; Zicari et al., 2022), the consultation of end users or their representatives can facilitate the identification of potential harms to users (Hasan et al., 2022; Zicari et al., 2022; Whittlestone et al., 2019; Shen et al., 2021), and the involvement of legal experts may be essential in cases where the audit relates to regulatory matters (Meßmer and Degeling, 2023; Zicari et al., 2022). Moreover, when feasible, conducting interviews with expert platform officials can yield valuable insights (Hasan et al., 2022; Meßmer and Degeling, 2023). However, when individuals from different disciplines collaborate,

there is often a divergence in the utilisation of terminology. Therefore, it is imperative to establish precise definitions for all significant words and concepts prior to the audit (Whittlestone et al., 2019).

5.2 Causality

As has been identified in the application domains, a crucial first step in deciding on the approach to auditing an algorithm is determining whether causality regarding a selected part of the algorithm's functionality or its effects should be established, given that certain types of approaches are unable to achieve causality in most cases. The utilisation of observational data can be employed to formulate causal claims through the employment of an econometric method, provided that the circumstances are deemed suitable for its implementation. However, in the majority of cases, appropriate circumstances are not met. All of the remaining approaches achieving causality either feed input into the algorithm or employ a controlled experiment with users in which the effect of a chosen treatment compared to a control group is analysed. In order to achieve a high level of specificity in the causality, strategies aimed at varying only one specific relevant factor are employed, with the resulting differences then analysed. Pattrn Analytics & Intelligence (2023) similarly differentiate between causal methods, which attribute outcomes to the algorithm, speculative methods, which analyse what might happen under certain assumptions, and descriptive methods, which provide insights into current outcomes but not their causes, in their framework.

5.3 Platform support

The second step in this process is to address the question of whether the platform itself is involved in the audit, and if not and if this power is given, whether its support needs to be requested for the insights desired, potentially in conjunction with the considerations made in the following step. Pattrn Analytics & Intelligence (2023) differentiate further between first-party, second-party and third-party audits. As has been demonstrated in the two application domains, the capacity to input data into the algorithm of a platform is, in the majority of cases, a prerequisite for achieving certain causality. One regarded approach takes the advantage of utilising a siloed environment on the real platform to upload potentially harmful content without platform support. However, in many cases, such an environment is not available and thereby platform support is required for providing inputs. Nevertheless, it should be noted that for other problems analysed, support for the platform might be less essential for the input of data and the achievement of causality. To illustrate this point, even in the absence of platform support, user characteristics could be inputted into the system in a controlled manner and the output systematically analysed in relation to the different inputs. Moreover, if the focus is not on the products of a platform, but rather on content that can be generated and uploaded in a non-harmful manner by the user, then recommendation algorithms could theoretically be probed with input using the real open system. Nevertheless, the creation of a wide variety of content can be demanding, depending on the type of content, and the utilisation of existing content may result in legal or ethical challenges, for instance with regard to copyright. Triangulating with the realm of advertising algorithms, several papers have created and uploaded advertisements without platform support to analyse how the algorithm delivers the ads to users (e.g., Imana et al., 2024; Ali et al., 2019; Kaplan et al., 2022; Sapiezynski et al., 2022). Fewer uploaded items might be sufficient to generate significant results, as significance can stem from biased distributions of only a few advertisements and from increasing the number of people who see an advertisement, not only from the number of advertisements published, although generalisability still depends on the scope of the advertisements. Nonetheless, placing advertisements on the actual platform may entail substantial expenses, given that advertisers are required to pay a fee (Ali et al., 2019).

In general, platform support reduces the burden of certain ethical or legal risks, such as the risk of influencing users' experiences negatively or violating the platforms' terms of service by scraping their content, thereby expanding the scope of feasible analyses. Furthermore, the potential increase in the availability of data may facilitate additional analysis, reduce the workload of certain analyses, and enhance the depth and accuracy of results, for instance, due to the elimination of a reliance on a workaround or proxy. For instance, with platform support and a sufficiently high level of access, the algorithm can be examined at different stages, thereby enabling the precise identification of the origin of any issues (Casper et al., 2024). However, when collaborating with the system operator, it is crucial to establish clear boundaries regarding their respective responsibilities and areas of work (Mökander, 2023).

5.4 Methodological approach

The third step in this framework concerns the selection of a general methodological approach. The application domains have demonstrated that the approaches can be categorised into three groups, each exhibiting distinct types of insights and objectives, with some applications falling more clearly into one category than others.

The first type of approach is that of an *observational analysis*. This approach is characterised by providing an overview of certain aspects, focusing on the current state of the system. This process entails the more or less passive collection of data from the system. This may also encompass the input of specific search terms or user characteristics of a sock-puppet account into the algorithm, though not for the purpose of directly analysing the effects of these inputs, but rather to facilitate the collection of data from several parts of the system. The results obtained can be utilised to generate descriptive statistics and conduct approximate analyses of certain connections based on observable factors. This approach can be used in many realms for obtaining a general overview and first insights into what is present on the system. However, in many cases, it might only be possible to achieve a partial picture, as the process of collecting data for the whole system part would be too cumbersome. In addition, as previously stated, the generated results are exclusively descriptive in nature in most cases, precluding the ability to make any causal assertions. Triangulating with other realms, such as the auditing of advertising algorithms, observational approaches are utilised as well (e.g., Kingsley et al., 2020). Nevertheless, in this context as well, a multitude of factors influence the outcomes of the algorithms, thereby complicating the clear attribution of causes using solely observational data (Imana et al., 2021; Lambrecht and Tucker, 2019). Should platform support be given for an observational analysis, it could be conducted with less effort and in a more encompassing way, as access to more data is possible and, in many cases, the underlying data might already be stored. Additionally, data that is not visible to the user but only to the platform can also be utilised. This approach can be mapped to one of the overarching themes regarding first-party audits concerning the identification of harmful and illegal content in Pattrn Analytics & Intelligence (2023) and one of the approach types of risk-measuring studies defined in Panigutti et al. (2025). Both of them are defined by the analysis of passive real-world data.

The second type of approach is an *input-output analysis*. This approach is characterised by presenting the algorithm with inputs and analysing the resulting outputs in relation to these inputs, thereby evaluating its functioning and performance. This can facilitate a more profound comprehension of the algorithm and provide insights into potential outcomes, not only current outcomes. This form of analysis is often essential for deriving causal results; however, depending on the specific causal assertions being made, merely adopting this approach is insufficient. For instance, even if an audit causally demonstrates that an algorithm performs worse for a particular group or type of input, causally attributing this disparity to a specific factor requires further analysis in which only that suspected factor is systematically varied.

Potential factors to vary include the data set or profile characteristics used as input, or the algorithm to which input is provided. The input may also be constructed to most likely display malfunctions, analyse potential adversarial attacks, respectively conduct red teaming. Red teaming has become a widely adopted approach to address concerns regarding the security, safety and reliability of AI algorithms (Majumdar, 2025). It is a team-based exercise that simulates adversarial attacks or subjects systems to stress tests in order to identify vulnerabilities, failure points, and unintended behaviour. The objective of this approach is to design inputs or scenarios that are intended to elicit failure, which may take the form of operational deficiencies or behavioural issues such as bias, unfairness, and lack of transparency. Red teaming is a sophisticated form of scrutiny that extends beyond the conventional evaluation of whether an algorithm performs as intended. It can be conducted in various forms, including black-box, white-box, or grey-box testing. As previously mentioned, in other domains such as the auditing of advertising algorithms, input-output analyses are conducted as well, including also alterations of the algorithms in a platform-supported way (e.g., Imana et al., 2024; Timmaraju et al., 2023). As has been previously indicated, a variety of analyses necessitate platform support for conducting an input-output analysis. Furthermore, it aids in controlling the input in the desired manner Furthermore, depending on the scope of access granted under platform support, additional types of analysis can be conducted, such as conducting model fine-tuning or employing potential mitigation strategies in development and analysing the effects of this. Moreover, when access to the output is granted at various computational steps, the reasons for different outputs given different inputs can be understood in a more fine-grained manner and attributed to the different systems.

This type of approach is also related to the classification of overarching themes in Pattrn Analytics & Intelligence (2023) and the approach types in Panigutti et al. (2025). Pattrn Analytics & Intelligence (2023) define their "experimentation methods" as the alteration of components of an algorithmic system, with the subsequent analysis of the consequences, a process frequently essential for the establishment of causal conclusions. This largely fits within our broader category of input–output analyses, which also includes studies that take the algorithm as given, although some user-focused experiments in Pattrn Analytics & Intelligence's (2023) group fall instead into our next category. These differences reflect our wider analytical scope, which extends beyond the specific topic of first-party audits of harmful content dissemination, and also encompasses the broader sociotechnical system. Panigutti et al. (2025) define their group of “experimental studies” in the realm of risk-measuring studies as interactions with the algorithmic system to evaluate its effects on systemic risks by varying independent scenario variables such as sock-puppet account behaviour. This closely overlaps with the input–output group resulting from our focus.

The third type of approach is a *user-centred analysis*. This approach places users at the centre, examining the algorithm’s impact on them, their experience when using the system, and whether it aligns with their interests. Such an approach can facilitate a more nuanced understanding of how the algorithm and users interact, helping to uncover potential problems that might remain hidden if the user component is ignored, and to assess whether suspected issues are in fact problematic once user factors are taken into account. The analysis can be conducted in qualitative or quantitative terms, depending on the chosen method. Nevertheless, a qualitative assessment may yield less concrete results, thereby hindering the ability to make causal attributions and form clear statements. In principle, causal statements may be possible with this approach if the actual system is included sufficiently and in a controlled way. This could be achieved, for instance, by altering its functioning in an experiment or by obtaining controlled user data to analyse the alignment of user and algorithm. Such approaches are also present, for instance, in the domain of advertising algorithms (e.g., Lam et al., 2023). If platform support were given, the workload of analyses could be reduced and the quality increased. For instance, an A/B test could be conducted on the platform in the most realistic way by directly altering the algorithm, thereby

analysing effects for a random sample of users and not a self-selected one. Moreover, survey participants could be directly surveyed or recruited through the platform, or the necessary data for an end-user audit could already be at hand. The third group of methodological elements of Pattrn Analytics & Intelligence (2023) relates to self-reported user experiences, thereby constituting a subpart of this type of approach, which is, once again, related to our broader focus. In consideration of the more restrictive focus in Panigutti et al. (2025), and given that only two groups are constructed, it is not possible to match this third group to one of theirs.

Furthermore, a fourth type of approach, namely a *documentation analysis*, which is not based on concrete examinations in any of the two application domains, is conceivable. This may comprise checklist approaches regarding the steps taken in the development of an algorithm, as well as the review of documents or associated materials, such as training data, for instance, based on considerations as in Raji et al. (2020), Koshiyama et al. (2024) and National Institute of Standards and Technology (2024). For instance, Tagharobi and Simbeck (2022) conduct a code audit in such a way, utilising 35,000 lines of the open-source code of an algorithm employed on a learning analytics platform. Nevertheless, in most cases, this will require the platform's support due to the proprietary nature of the code. A documentation analysis has the capacity to provide insights into the underlying infrastructure and employment of the algorithm, thereby facilitating a more comprehensive understanding of the system in its entirety. It can help with identifying potential issues that have not yet manifested but may become active in the future. However, it does not necessarily provide insights into how the system behaves when being employed. The derivability of causal statements is contingent upon the type of documentation provided.

Overall, it can be summarised that the most suitable approach depends on the depth of insights that is to be generated, the resources available, including potential platform support, and the concrete type of potential problem. A combination of approaches might also prove to be beneficial to complement findings (Meßmer and Degeling, 2023; Mökander et al., 2021).

5.5 Metrics and indicators

The fourth step concerns the selection of metrics or indicators for analysis, an aspect that is reflected in multiple other frameworks (Panigutti et al., 2025; Hasan et al., 2022; Meßmer and Degeling, 2023; Pattrn Analytics & Intelligence, 2023). This step is also related to whether a structural model is established, which describes certain relations between variables, and which consequently leads to potential metrics and indicators being analysed. Furthermore, certain metrics may necessitate a model with specific assumptions, which may or may not be fulfilled, to be analysable. The choice of metrics should be made with close consideration of the problematic behaviour in question, with the aim of both quantifying the problem and avoiding the risk of p-hacking (Bandy, 2021). For instance, if real-world harms are to be analysed, it is essential that the chosen measurements are effective proxies thereof (Metcalf et al., 2021). A prioritisation of potential measures might be conducted on the basis of their capacity to accurately measure the intended phenomenon and the complexity of implementation (Meßmer and Degeling, 2023). The subsequent two subchapters illustrate two relevant challenges relating to this step, namely trade-offs and fairness definition, in greater depth.

5.5.1 Challenges related to trade-offs

The first challenge relates to potential trade-offs between certain values and objectives that an algorithm should fulfil. It may be the case that not all objectives can be fulfilled perfectly, but rather that an acceptable combination of values for the regarded use case must be found. This complicates the

interpretation of metrics and indicators, since it is unlikely that a single metric can fully capture all sides of a trade-off. As a result, the selection of metrics must be particularly well justified, and a subjective judgement about which combinations of trade-offs are acceptable will often be required. In certain exceptional cases, and with sufficiently high access, Pareto frontiers regarding the examined trade-offs, based on chosen metrics, might be derivable, which can facilitate the identification of adequate combinations.

In order to facilitate a better understanding of the potential trade-offs that may need to be considered, some exemplary cases are outlined below. One such trade-off is that between fairness and accuracy. Should there be a difference in outcomes between groups that should be treated equally, not allowing the algorithm to act on this difference may reduce the model's accuracy. A further trade-off that may be considered is that between explainability and accuracy. In cases where the underlying relationship is complex, the employment of a complex model is likely to yield the highest levels of accuracy. However, with such a model, it may prove challenging to explain the construction of the resulting outcomes and the role of the various variables. In certain contexts, more abstract and less apparent trade-offs may be relevant, such as those between personalisation and solidarity (Whittlestone et al., 2019). Further examples of trade-offs can be found, for instance, in Whittlestone et al. (2019) and Koshiyama et al. (2024).

In order to identify which tensions and trade-offs might arise between important values, it can be helpful to consider possible tensions arising from winner versus losers, short term versus long term or local versus global tensions (Whittlestone et al., 2019). It may be the case that certain individuals benefit due to the algorithm, whilst others experience disadvantage, that different effects are present in the short term as opposed to when employed over a longer period, and that the effect on individuals may differ from the effect on a global level.

In addressing the identified tensions, it may be advantageous to differentiate between true dilemmas, dilemmas in practice, and false dilemmas (Whittlestone et al., 2019). This can facilitate the differentiation between substantial argumentations concerning trade-offs and those that are less substantiated. A true dilemma arises when the achievement of all significant values is unfeasible, as they are inherently conflicting with each other. Nevertheless, an identified tension may also be a dilemma in practice, in the sense that it is the present technological capabilities and other present constraints, such as those relating to time and resources, that give rise to the tension, rather than any inherent values. Consequently, a decision must be made between either directly implementing the system and accepting the associated trade-offs or investing more resources to avoid them. Furthermore, false dilemmas may be present, in which case a previously unconsidered option could potentially resolve the tension.

However, conducting an audit of a system from an external perspective can present challenges in accurately comprehending the tension between multiple values, as well as the combinations of values that are feasible within the specified use case. Consequently, it may prove challenging to evaluate the precise quality of the solution chosen to a trade-off. Nevertheless, these considerations facilitate a more profound comprehension of the system, thereby enabling the generation of more sophisticated and deliberate analyses and conclusions. For instance, in a classification situation involving a clear trade-off between false negatives and false positives, consideration of this trade-off can facilitate the selection of more suitable, encompassing metrics such as the F1-score, which takes into account both components, as opposed to merely analysing false negatives. Furthermore, in specific circumstances, a combination of values that is strictly dominated may be identifiable, for example, if one value is significantly more important than the other, if an adequate comparison regarding possible combinations is available, or if the law prohibits such a trade-off.

5.5.2 Challenges related to the definition of fairness

A further challenge in selecting an appropriate metric arises when analysing the concept of fairness and attempting to conceptualise it. There exists a multitude of criteria for fairness, with each one conceptualising a distinct notion, and it is not feasible to fulfil all of them simultaneously in every situation (Kleinberg et al., 2016; Chouldechova, 2017; Castelnovo et al., 2022; Friedler et al., 2021; Berk et al., 2021). Furthermore, it has been demonstrated that, for certain criteria, if they can be satisfied concurrently yet only in a highly specific, artificial situation, fulfilling them approximately requires a situation that is very close to the specific, artificial one (Kleinberg et al., 2016). Consequently, it proves essential to identify which version of fairness might be adequate in the case under consideration and to consider the consequences of violations of any fairness criteria. Similarly, with regard to the EU law for discrimination, there is not one fairness metric and a given threshold that is suitable for all situations, as the evaluation is highly context-dependent (Wachter et al., 2021).

Popular fairness metrics include demographic/statistical parity (equal probability of prediction of the positive outcome across groups), conditional demographic parity (like the unconditional case but holding conditional on certain characteristics), equality of odds (equal false-positive-rate and false-negative-rate across groups), equal opportunity (equal probability of a predicted positive outcome when the true outcome is positive), predictive parity (equal probability of an actual positive outcome when predicted across groups), calibration of scores (equal probability of a true positive outcome given any predicted score value across groups). The majority of metrics employed for binary outcomes can also be utilised for multi-class outcomes; however, when examining ranking outcomes or continuous outcomes, such as related to continuous regressions, these criteria must be modified or different ones must be employed, such as those proposed by Zehlike et al. (2021), Zehlike et al. (2022), Steinberg et al. (2020), Perera et al. (2022), Gursoy and Kakadiaris (2022) and Agarwal et al. (2019). Further methods include the utilisation of individuals' preferences in defining fair situations, for example, by allowing deviations from certain fairness criteria if in line with user preferences (Kim et al., 2019; Ustun et al., 2019; Zafar et al., 2017; Do et al., 2022). However, the process of identifying user preferences may present significant challenges.

The adequacy of these criteria is dependent on the specific circumstances of each application. For instance, when it is reasoned that the predicted outcome should be entirely unrelated to group affiliation, the use of a criterion such as demographic parity might be adequate (Castelnovo et al., 2022). In circumstances where the true outcome is reliably measuring what it should, and when discrimination can be argued for as long as it is based on reliable data, using criteria such as equality of odds or equal opportunity might be adequate (Castelnovo et al., 2022). Under conditions where the true outcome is reliable and the predicted outcome should be equally predictive of the true outcome across groups, criteria such as calibration and predictive parity can be employed. In instances where the distribution of the relevant characteristic, such as risk, varies across groups, which is a common occurrence, error metrics may not be optimal for measuring individual equity or social well-being, such that the calibration of estimated scores should be examined prior to concluding that any modifications regarding a bias are required if error metrics indicate this (Corbett-Davies et al., 2023). Intuitively, if, for instance, the conditional risk distributions of people with negative true outcomes differ and the score is well-calibrated, there are necessarily score values for which more people of one group than the other are above these scores, thereby leading to a higher false positive rate in the first group. In the context of EU discrimination legislation, investigating the conditional demographic parity may represent a valuable initial approach for the collection of evidence (Wachter et al., 2021).

However, the argument has been made for a shift in focus away from any particular fairness metrics and towards a more direct examination of the inherent trade-offs involved, including the weighting of costs and benefits of different policies and their effect on real-world quantities (Corbett-Davies et al., 2023). For instance, it has been demonstrated that deviating from a utility-maximising approach to satisfying a fairness metric can result in an outcome that is detrimental to both individuals and groups, as evidenced in the context of college student selection, where this approach is not on the Pareto curve regarding the student body's skill and diversity, but Pareto-dominated, given any utility function that increases in both variables (Corbett-Davies et al., 2023). Nevertheless, these critics have noted that the utilisation of fairness metrics can still carry value, but that a less uncritically approach to their application with a broader view should be adopted (Corbett-Davies et al., 2023).

A further point to consider in the context of fairness is whether intersectional fairness is a more adequate concept than one-dimensional fairness, given the use of algorithms increases the likelihood of such fine-grained unfairness (Gerards and Borgesius, 2022). For instance, in the context of image classification, it has been demonstrated that the error rates for darker-skinned women are significantly higher than would be expected based solely on error rates for women with any skin colour or for dark-skinned people in general (Buolamwini and Gebru, 2018). However, certain computational and practical issues emerge in the context of intersectional fairness, primarily due to the exponential growth in the number of subgroups when adding sensitive attributes, and that with real samples, a significant proportion of these subgroups may possess either zero or a limited number of observations (Castelnovo et al., 2022). Furthermore, the relevance regarding EU law remains ambiguous (Wachter et al., 2021; Gerards and Xenidis, 2020).

5.6 Data collection method

The fifth step concerns the selection of the most appropriate data collection method, a matter which is also addressed in many regarded frameworks as a standalone or sub-step (Metaxa et al., 2021; Panigutti et al., 2025; Hasan et al., 2022; Meßmer and Degeling, 2023; Pattrn Analytics & Intelligence, 2023; Morales-Navarro et al., 2025). This step is closely related to the previous one, as the choice of metric can dictate the breadth and depth of data that needs to be collected. Certain data collection methods cannot be employed for certain metrics, and certain metrics might not be analysable due to a lack of a suitable data collection method. Therefore, an inventory of the possibilities of data collection on the system itself and from other sources must be set up and analysed in order to assess which metrics would be analysable and between which data collection can be chosen from if multiple ones are possible. In certain instances, a particular data collection method may serve as the foundation for another primary method, such as the scraping of information for utilisation in a user survey, or multiple methods may be employed at equivalent levels. This step further encompasses considerations regarding which components exactly should be collected with the chosen data collection method. In the event that data is obtained by an external party using platform support, this step also includes consideration of the manner in which access to the data is structured. For instance, whether data is freely accessible or whether access to it is only granted in a secure environment, as described in Section 3.2. As discrepancies regarding suitable and chosen data collection methods depending on the methodological approach were identified in the application domains, an analysis of the possible methods and arguments for and against them depending on the circumstances is conducted separately for the different types of approaches.

Should an observational analysis be selected, a number of different data collection methods may be employed, such as scraping, leveraging a first- or third-party API if available, crowdsourcing, a sock-puppet approach, or obtaining other data available by commercial third parties. In general, a broad range of results with control over the collection process can be achieved with scraping. Should the desired data be obtainable only upon login, or should a specific user-characteristic-related scenario be analysed, scraping can be conducted using a sock-puppet account. If a trustworthy and free or affordable API provides equivalent or even more information than available via scraping, it might be preferable as the amount of effort required is substantially lower and legal risks regarding scraping are avoided. Other commercial third-party data may facilitate the acquisition of otherwise inaccessible data. Should personalisation or real user behaviour be influencing the observed data, crowdsourcing could be used, which enables the provision of guidance and the exercise of control over data collection, but also the absence of such control to analyse the data of real users when typically engaging with the system. If an observational analysis were to be conducted with platform support, the platform might provide the auditor with the necessary data or access to it, or grant sufficient grey-box access such that the sought-after statistics can be generated.

In the event of an input-output approach being selected, and if this is conducted with platform support, certain levels of grey-box access or de facto white box access could be employed, depending on the specific analysis, such that, for instance, the model can be sampled, fine-tuned or modified. Moreover, access to proprietary datasets may be advantageous in instances where no suitable datasets are available. With regard to input-output analyses conducted without platform support, the sole paper in the application domains that employs such an analysis utilises a platform API. Furthermore, several regarded studies within the domain of content moderation, for which the observational analysis constitutes the foundation for subsequent analyses, undertake input-output analysis concerning user characteristics. This is achieved through the utilisation of both automated and manual sock-puppet accounts, in addition to crowdsourcing. In the domain of advertising algorithms, numerous papers employ platform repurposing, whereby the statistics provided to advertisers by the platform are utilised for analysis, with explicit mention of this process occurring through the platform's API in some cases (Imana et al., 2024; Kaplan et al., 2022; Sapiezynski et al., 2022; Imana et al., 2021; Ali et al., 2019). In most instances, this data is supplemented with additional necessary data concerning missing characteristics and a strategy for combining these data. Furthermore, within the domain of advertising, yet employing user characteristics as inputs to analyse exploitation of and discrimination based on user data, Datta et al. (2014) utilise automated sock-puppet accounts. A similar pattern regarding the methods and their suitability as identified for the observational analyses can be observed. In instances where personalisation should be considered or input and output can only be provided and collected upon login, respectively, using a sock-puppet approach, entailing the use of a controlled yet somewhat artificial environment capable of high scalability could be beneficial. Alternatively, crowdsourcing with real users and real user history, albeit in a less controlled environment, with reduced scalability could be selected. The less controlled environment may hinder causal claims, as many attributes may be different between user profiles, acting as confounding factors when trying to isolate the effects of certain user characteristics. Conversely, in the context of sock-puppet audits, profiles must be meticulously designed and scrutinised to ascertain their authenticity if the analysis is about more than simply changing profile settings such as gender, thereby ensuring the attribution of causal effects relating to user behaviour to the intended simulation. In instances where personalisation is not necessary, the utilisation of the API may be a beneficial approach, provided it offers the desired output and input options and its data is reliable, as it demands less effort than the alternative methods and avoids legal risks. Moreover, certain platform functionalities might be repurposed for the analysis, reducing the workload and providing further data in some cases. Alternatively, simple scraping provides a broad array of options, with the exception of analysing personalisation.

If a user-centred audit is selected, potential options comprise user surveys, end-user audits, crowdsourcing and sociotechnical audits. In circumstances where comprehensive context pertaining to the user experience is sought or first insights are to be gathered, the implementation of a user survey might prove advantageous. It is possible to collect a broad array of information, some of which is not visible when only looking at the system itself. This allows for a more complete picture of the user experience to be created. However, it should be noted that surveys are susceptible to common biases. Conducting an end-user audit enables the analysis of the direct user experience and opinions based on specific system output, as well as elements that users do not notice in their everyday use. Nevertheless, this necessitates substantial resources, including a multi-labelled dataset and a diverse range of motivated users. The sociotechnical audit can provide insights into the effect of changes in the algorithm on the users, potentially generating causal results. Depending on the circumstances, this may necessitate a substantial amount of work, as an experiment must be designed. Moreover, a dissimilarity of the setting to the real world may potentially hinder the generalisability of the results. The examination of interaction from real accounts with the system in a real, uncontrolled way can be facilitated through crowdsourcing. If platform support were given, sociotechnical audits, end-user audits or user surveys might be conducted as well, though with more possibilities and potentially less resources required. Moreover, provisions of data concerning user behaviour and outputs could be considered.

In the context of a documentation audit, the platform may provide a range of outside-the-box information for analysis.

5.7 Implementation and handling of challenges

The sixth step involves the preparation for the handling of implementation challenges, i.e. the hurdles that arise when attempting to utilise the methodological approach and the data collection method to analyse the chosen metrics, followed by the implementation itself. Drawing upon the more detailed insights generated in the application domains, the following section delineates the common challenges faced.

A common challenge encountered concerns *ethical and legal risks*. This encompasses potential violations of the terms of service, risks associated with the possession of certain content, and potential adverse effects through the actions undertaken on users, researchers, or the platform itself. In addition, this encompasses the potential adverse effects on the platform resulting from the increased computational resources required, for example, if rate limits are not adhered to (Metaxa et al., 2021). Potential solutions include the limitation of actions and analyses conducted on the platform, security precautions relating to the storage of data, data anonymisation, obtaining permission by the platform, resorting to the necessity of platform support, or, in the most extreme case, to not conduct such an analysis at all if no adequate resolving approach can be identified.

A further challenge relates to the *obtaining of adequate data*, which is often scarce and difficult to obtain accurately or at all. This may be due to factors such as information asymmetry between the platform and researcher or between the researcher and users, a certain specificity or depth of data being required for a valuable analysis, or required scalability, which excludes certain, maybe more accurate, approaches due to a high workload. Associated workloads entail the evaluation of existing datasets to ascertain their adequacy for the analysis, and in instances where these datasets prove to be inadequate, the formulation of a strategy for the collection of the missing data. This may involve the collection of observational data from the system, such that it fulfils the desired criteria by establishing a search strategy, for instance, by identifying relevant keywords related to a particular topic or by focusing on the most popular keywords. Furthermore, the necessity for data classification may arise. Potential approaches

for this include the utilisation of existing proxies, manual classification by researchers or crowd workers, automated classification, or a combination of manual and automated classification, for instance, automated classification based on initial manual classification or employing one or the other based on the type of content regarded. Moreover, certain data may be obtained only with inaccuracies, for instance, if the actual data is unavailable and proxies are used instead such that an adequate strategy for handling this must be developed. In relation to this subject, and with regard to the auditing of advertising algorithms, it is frequently the case that voter data is utilised for the purpose of inferring characteristics of recipients that are not reported by the platform; however, this can result in reduced statistical power if not adequately accounted for (Imana et al., 2025). Furthermore, potential biases inherent to surveys must be identified and mitigated. In general, if data for probing the algorithm is to be used, it should be ensured that the data is of good quality, meaning that it is close to the conditions under which the algorithm is normally used, has the relevant characteristics such as race or gender correctly labelled and includes enough data points in regions of interest such that statistical power can be achieved (Hasan et al., 2022). The issue of ground truth labelling is also addressed in Panigutti et al.'s (2025) framework.

In relation to this challenge, there is also the issue of achieving the desired *representativeness*. For instance, if the selection of which parts of the system are to be analysed, such as those related to chosen keywords, is systematically biased in one way, this will hinder the representativeness regarding broader conclusions. The same can be said of data input into the system if it holds very specific characteristics. However, a selective analysis might be required to achieve representativeness for the posed statement. Concerning approaches incorporating real users, there is the possibility of selection bias with regard to the participants. For instance, only a certain subset of people might be willing to have their behaviour tracked through a browser extension (Metaxa et al., 2021). Therefore, when formulating conclusions regarding the scope of applicability, it is essential to consider the choices made and the potential limitations encountered (Meßmer and Degeling, 2023).

A further challenge, especially if causality is to be achieved, concerns the *control of confounding factors*, which is also reflected in Panigutti et al.'s (2025) framework. This is especially challenging in cases where data access is restricted, as it is more difficult to isolate the change in only the relevant attribute, and a multitude of potential confounding factors may be present. Consequently, a strategy for maintaining as much as possible constant, except for the relevant factor, must be developed, for instance, to create a control and treatment group in an experiment or to alter only one part of the algorithmic functioning. If this is not feasible, the collection of as much data as possible on potential confounders and controlling for them in the analysis might be beneficial, or the exploitation of rare exogenous events could be a possibility. Should sock puppets be utilised, their behaviour should be assessed for its alignment with the desired behaviour, ensuring that the analysis is not confounded by other unintended factors. In general, it might be beneficial to develop strategies for controlling for confounding factors with varying levels of control and realism, such that a clear identification of effects can be achieved with much control but abstraction, and better transfers to reality can be achieved with less control and more realism. This identified challenge also relates to Metaxa et al.'s (2021) key decision point of measuring personalisation, in which either strategies for avoiding personalisation of results to avoid the confounding effect of this, or, if personalisation is the variable of interest, strategies for isolating its effect should be set up.

Furthermore, the *isolation of the effects of the different components* in and surrounding algorithmic systems poses challenges. For instance, in the context of content moderation algorithms, the outcome may be a combination of different algorithm types, such as matching or classification algorithms, as well as human judgement. In the context of recommendation algorithms, it may be a combination of candidate selection and multiple ranking, respectively re-ranking algorithms. Should isolation of the effects of the different components be relevant to the research question, strategies for doing so should be developed.

However, depending on the level of access, such isolation might not be possible, thereby requiring this to be taken into account in the conclusions.

A general challenge associated with algorithms in sociotechnical systems is that they are *dynamic and ephemeral* (Metaxa et al., 2021). Given the prevalence of machine learning algorithms in sociotechnical platforms, their models and outputs are continuously updated in response to new data and user interactions, in addition to changes in algorithmic code. Consequently, any observation of the system might rapidly become outdated. Additionally, outputs such as search rankings or personalised feeds are typically not stored in a manner that allows for later retrieval, which hinders the reconstruction or audit of what users previously encountered, necessitating prior planning. Metaxa et al. (2021) thus advise temporal considerations to be made, such as regarding the frequency and timing of data collection, the examination of potential effects of current events, and the re-application of the audit procedure to analyse the effects of algorithm changes, similar to Zicari et al. (2022) who recommend ongoing monitoring.

The auditing process should only be conducted if adequate solutions to these challenges can be found.

5.8 Data analysis, robustness and reporting

The seventh step concerns the actual data work and analysis, which is represented in most frameworks. These tasks may encompass data cleansing or merging, metric calculation, and statistical testing. In certain situations, a comparison of the metrics with a baseline situation is necessary, for instance regarding the definition of an unbiased distribution of content sources, which must be carefully chosen for the drawing of correct conclusions, but which is often a non-trivial task (Metaxa et al., 2021; Bandy, 2021). Furthermore, when evaluating the metrics, it should be differentiated between a statistically significant result that carries little relevance in practice due to its small size and a statistically significant result that also has an economic relevance. Especially in the context of big data, a situation of statistical but not economic significance might arise (Wachter et al., 2021).

This is followed by the eighth step of increasing the results' robustness through supplementary analyses and tests, which is also mentioned in Panigutti et al.'s (2025) framework. This may include varying model assumptions, specifications, and data filtering, and analysing the effects of this on the results, as well as analysing the out-of-sample fit of models, and conducting placebo and plausibility tests.

The final step in the process is the reporting of results, including any potential limitations and shortcomings, and the placing of these results into context (Metaxa et al., 2021; Meßmer and Degeling, 2023; Morales-Navarro et al., 2025). Moreover, the conducted data work should be communicated clearly, and the effects of any decisions made on the results stated, such that a replication by other researchers is possible (Metaxa et al., 2021; Meßmer and Degeling, 2023). The results may be contrasted and combined with previous findings.

6 Conclusion

Based on a structural analysis of existing empirical approaches to auditing algorithms—both with respect to the insufficient algorithmic moderation of illegal and harmful content and to potential self-preferencing in recommendation algorithms—combined with a further literature review on algorithmic auditing encompassing additional empirical studies, methodological approaches, frameworks, and best practices, a framework has been developed that sets out concrete, fine-grained steps for the practical implementation of algorithmic auditing under a variety of circumstances and objectives. This framework encompasses the steps of initial setup, determining whether causality in results is required, assessing whether platform support is available or necessary, selecting a focus on one of the identified methodological approaches of observational, input-output, user-centred or documentation analyses, determining metrics to be analysed, establishing how the required data should be collected, establishing strategies for handling implementation challenges, analysing the data, increasing its robustness and reporting the results. Common challenges encountered by researchers include ethical and legal risks, obtaining adequate data, achieving representativeness, controlling for confounding factors, isolating effects of certain algorithmic components, handling the dynamic nature of algorithms and choosing adequate metrics, for instance mirroring trade-offs or fairness considerations well. It has been demonstrated that existing frameworks can be well incorporated into the newly developed one, thereby substantiating its generalisability.

Furthermore, the analysis has identified numerous concrete approaches that can be employed for auditing algorithms and handling the various challenges, as well as the advantages and disadvantages of some differentiations, which can be utilised for future audits. The results have once again emphasised the importance of platform support, which is often indispensable for achieving causality in results and can enhance the efficiency, precision and scope of audits. The generated results are valuable for conducting future audits and can provide a scaffold for the analysis. Moreover, they facilitate the comprehension of conducted audits and their approaches, quality, and implications, thereby enabling observers to attain a more sophisticated understanding of the current state of research, as well as of specific results presented to them.

However, it should be noted that the present study is not without its limitations. As a systematic literature review has not been conducted, for instance following the PRISMA guideline, it is possible that some existing approaches have not been identified and the focus area of the approaches regarded in this paper might not be an exact representation of the underlying distribution. Moreover, the present study's in-depth analyses have been restricted to two domains, such that it might be the case that the framework can be less well applied to non-regarded domains. However, it should be noted that the results of frameworks regarding other domains, as well as some additional empirical literature, are also examined. Future research could further analyse how platform-supported audits, conducted by researchers with the platform's assistance rather than the platform itself, and documentation audits, could precisely be integrated into the framework, as the types of analysis have been less well represented in the examined empirical literature.

References

Agarwal, A., Dudík, M., & Wu, Z. S. (2019). Fair regression: Quantitative definitions and reduction-based algorithms. *Proceedings of the 36th International Conference on Machine Learning,* PMLR 97:120-129.

AlDahoul, N., Karim, H. A., Momo, M. A., Sy, M. A., & Tan, M. J. T. (2023). Evaluation of content moderation software for nudity and pornography detection in various scenarios. *MECON Multimedia University Engineering Conference.*

AlDahoul, N., Tan, M. J. T., Kasireddy, H. R., & Zaki, Y. (2024). *Advancing content moderation: Evaluating large language models for detecting sensitive content across text, images, and videos*. arXiv:2411.17123.

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How Facebook's ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1–30.

Ami, R. (2025). AI in automated content moderation on social media. *International Journal of Artificial Intelligence and Machine Learning in Engineering, 21*(3).

Bandy, J. (2021). Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1–34.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research, 50*(1), 3–44.

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., … Goodrow, C. (2019). Fairness in recommendation ranking through pairwise comparisons. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2212–2220).

Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. *Companion Proceedings of the 2019 World Wide Web Conference* (pp. 491–500).

Boroughf, B. (2015). The next great YouTube: Improving Content ID to foster creativity, cooperation, and fair compensation. *Albany Law Journal of Science & Technology, 25*, 95.

Boshmaf, Y., Perera, I., Kumarasinghe, U., Liyanage, S., & Al Jawaheri, H. (2023). Dizzy: large-scale crawling and analysis of onion services. *Proceedings of the 18th International Conference on Availability, Reliability and Security* (pp. 1-11).

Bucknall, B. S., & Trager, R. F. (2023). *Structured access for third-party research on frontier AI models: Investigating researchers' model access requirements*. Whitepaper, October 2023.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the Conference on Fairness, Accountability and Transparency* (pp. 77–91). PMLR.

Carugati, C. (2022). *How to implement the self-preferencing ban in the European Union's Digital Markets Act*. Policy Contribution 22/2022, Bruegel.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., … Hadfield-Menell, D. (2024). Black-box access is insufficient for rigorous AI audits. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2254–2272).

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports, 12*(1), Article 4209.

Chen, N., & Tsai, H. T. (2024). Steering via algorithmic recommendations. *The RAND Journal of Economics, 55*(4), 501–518.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163.

Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., & Goel, S. (2023). The measure and mismeasure of fairness. *Journal of Machine Learning Research, 24*(312), 1–117.

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1571–1583).

Cure, M., Hunold, M., Kesler, R., Laitenberger, U., & Larrieu, T. (2022). Vertical integration of platforms and product prominence. *Quantitative Marketing and Economics, 20*(4), 353–395.

Dash, A., Chakraborty, A., Ghosh, S., Mukherjee, A., & Gummadi, K. P. (2024). Investigating Nudges toward Related Sellers on E-commerce Marketplaces: A Case Study on Amazon. *Proceedings of the ACM on Human-Computer Interaction, 8*(CSCW2), 1-31.

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. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 873-884).

Datta, A., Tschantz, M. C., & Datta, A. (2014). *Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination*. arXiv:1408.6491.

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity* (ITASEC17) (pp. 86–95).

Dendorfer, F., & Seibel, R. (2025). *What's In the Box? The Effect Of Self-Preferencing On Amazon*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5219585.

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 67–73).

Do, V., Corbett-Davies, S., Atif, J., & Usunier, N. (2022). Online certification of preference-based fairness for personalized recommender systems. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(6), Article 6. https://doi.org/10.1609/aaai.v36i6.20606

Dreier, L., & Pirker, J. (2023). Toxicity in Twitch live stream chats: Towards understanding the impact of gender, size of community, and game genre. *2023 IEEE Conference on Games* (CoG) (pp. 1–4). IEEE.

Eltaher, F., Gajula, R. K., Miralles-Pechuán, L., Crotty, P., Martínez-Otero, J., Thorpe, C., & McKeever, S. (2025). *Protecting young users on social media: Evaluating the effectiveness of content moderation and legal safeguards on video-sharing platforms*. arXiv:2505.11160.

Farronato, C., Fradkin, A., & MacKay, A. (2023). Self-preferencing at Amazon: evidence from search rankings. *AEA Papers and Proceedings* (Vol. 113, pp. 239-243).

Farronato, C., Fradkin, A., & MacKay, A. (2025). *Vertical integration and consumer choice: Evidence from a field experiment*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5381505.

Fink, N. C., & Saltman, E. (2025). *Fighting terror with tech: The evolution of the Global Internet Forum to Counter Terrorism. Trust, Safety, and the Internet We Share: Multistakeholder Insights* (forthcoming).

Ford, G. S. (2023). *Can self-preferencing by an online retailer be detected? A Monte Carlo simulation*. Phoenix Center Policy Bulletin, (64).

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM, 64*(4), 136–143.

Gerards, J., & Borgesius, F. Z. (2022). Protected grounds and the system of non-discrimination law in the context of algorithmic decision-making and artificial intelligence. *Colorado Technology Law Journal, 20*, 1.

Gerards, J., & Xenidis, R. (2020). *Algorithmic discrimination in Europe: Challenges and opportunities for gender equality and non-discrimination law*. Publications Office of the European Union, 2021. https://data.europa.eu/doi/10.2838/544956.

Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019). Fairness-aware ranking in search and recommendation systems with application to LinkedIn Talent Search. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2221–2231). https://doi.org/10.1145/3292500.3330691

Goldman, E. (2021). Content moderation remedies. *Michigan Technology Law Review, 28*, 1.

Gómez, E., Contreras, D., Boratto, L., & Salamó, M. (2025). Enhancing recommender systems with provider fairness through preference distribution-awareness. *International Journal of Information Management Data Insights, 5*(1), 100311.

Gómez, E., Contreras, D., Boratto, L., & Salamó, M. (2025). Enhancing recommender systems with provider fairness through preference distribution-awareness. *International Journal of Information Management Data Insights*, *5*(1), 100311.

Gomez, J. F., Machado, C., Paes, L. M., & Calmon, F. (2024). Algorithmic arbitrariness in content moderation. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2234–2253).

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society, 7*(1), Article 2053951719897945.

Gosztonyi, G., Gyetván, D., & Kovács, A. (2025). Theory and Practice of Social Media's Content Moderation by Artificial Intelligence in Light of European Union's AI Act and Digital Services Act. *European Journal of Law and Political Science 4*(1):33-42.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All you need is "love": Evading hate speech detection. *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security* (pp. 2–12).

Gursoy, F., & Kakadiaris, I. A. (2022). *Error parity fairness: Testing for group fairness in regression tasks*. arXiv:2208.08279.

Gutfeter, W., Gajewska, J., & Pacut, A. (2023, September). Detecting sexually explicit content in the context of child sexual abuse material (CSAM): End-to-end classifiers and region-based networks. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 154–168).

Haimson, O. L., Delmonaco, D., Nie, P., & Wegner, A. (2021). Disproportionate removals and differing content moderation experiences for conservative, transgender, and Black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2)*, 1–35.

Hartzell, O. (2025). *Platform Preferencing and Price Competition II: Evidence From Amazon*. SSRN 5126920.

Hartzell, O., & Haupt, A. (2025). *Platform preferencing and price competition I: Evidence from Amazon*. SSRN 5126918.

Hasan, A., Brown, S., Davidovic, J., Lange, B., & Regan, M. (2022). Algorithmic bias and risk assessments: Lessons from practice. *Digital Society, 1*(2), Article 14.

Hilbert, M., Thakur, A., Flores, P. M., Zhang, X., Bhan, J. Y., Bernhard, P., & Ji, F. (2024). 8–10% of algorithmic recommendations are 'bad', but… An exploratory risk–utility meta-analysis and its regulatory implications. *International Journal of Information Management, 75*, 102743.

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). *Deceiving Google's Perspective API built for detecting toxic comments*. arXiv:1702.08138.

Hunold, M., Kesler, R., & Laitenberger, U. (2020). Rankings of online travel agents, channel pricing, and consumer protection. *Marketing Science, 39*(1), 92–116.

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), 1–27.

Imana, B., Korolova, A., & Heidemann, J. (2021). *Auditing for discrimination in algorithms delivering job ads*. *Proceedings of the Web Conference 2021* (pp. 3767–3778).

Imana, B., Korolova, A., & Heidemann, J. (2023). Having your privacy cake and eating it too: Platform-supported auditing of social media algorithms for public interest. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), 1–33.

Imana, B., Korolova, A., & Heidemann, J. (2024). Auditing for racial discrimination in the delivery of education ads. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2348–2361).

Imana, B., Korolova, A., & Heidemann, J. (2025). Auditing for bias in ad delivery using inferred demographic attributes. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2640–2656).

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 (*pp. 1–27).

Juneja, P., Bhuiyan, M. M., & Mitra, T. (2023). Assessing enactment of content regulation policies: A post hoc crowd-sourced audit of election misinformation on YouTube. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–22).

Jürgensmeier, L., & Skiera, B. (2023). *Measuring self-preferencing on digital platforms*. arXiv:2303.14947.

Kaplan, L., Gerzon, N., Mislove, A., & Sapiezynski, P. (2022). Measurement and analysis of implied identity in ad delivery optimization. *Proceedings of the 22nd ACM Internet Measurement Conference* (pp. 195–209).

Kim, M. P., Korolova, A., Rothblum, G. N., & Yona, G. (2019). *Preference-informed fairness*. arXiv:1904.01793.

Kingsley, S., Wang, C., Mikhalenko, A., Sinha, P., & Kulkarni, C. (2020). *Auditing digital platforms for discrimination in economic opportunity advertising*. arXiv:2008.09656.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent trade-offs in the fair determination of risk scores*. arXiv:1609.05807.

Klonick, K. (2017). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review, 131*, 1598.

Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., et al. (2024). Towards algorithm auditing: Managing legal, ethical and technological risks of AI, ML and associated algorithms. *Royal Society Open Science, 11*(5), 230859.

Kumar, D., AbuHashem, Y. A., & Durumeric, Z. (2024). Watch your language: Investigating content moderation with large language models. *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 18, pp. 865–878).

Lahti, H., Kokkonen, M., Hietajärvi, L., Lyyra, N., & Paakkari, L. (2024). Social media threats and health among adolescents: evidence from the health behaviour in school-aged children study. *Child and adolescent psychiatry and mental health*, *18*(1), 62.

Lam, M. S., Gordon, M. L., Metaxa, D., Hancock, J. T., Landay, J. A., & Bernstein, M. S. (2022). End-user audits: A system empowering communities to lead large-scale investigations of harmful algorithmic behavior. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW2), 1–34.

Lam, M. S., Pandit, A., Kalicki, C. H., Gupta, R., Sahoo, P., & Metaxa, D. (2023). *Sociotechnical audits: Broadening the algorithm auditing lens to investigate targeted advertising*. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW2), 1–37.

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.

Lee, K. H., & Musolff, L. (2025). *Two-Sided Markets Shaped by Platform-Guided Search*.

Lees, A., Tran, V. Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., & Vasserman, L. (2022). A new generation of perspective api: Efficient multilingual character-level transformers. *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 3197-3207).

Leung, T. C., Qi, S., & Strumpf, K. (2025). *Dissecting Netflix's self-preferencing: Evidence from viewer-level data.* Working Papers 25-08, NET Institute.

Lykouris, T., & Weng, W. (2024). *Learning to defer in content moderation: The human–AI interplay*. arXiv:2402.12237.

Mahomed, Y., Crawford, C. M., Gautam, S., Friedler, S. A., & Metaxa, D. (2024). Auditing GPT's content moderation guardrails: Can ChatGPT write your favorite TV show? *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 660–686).

Majumdar, S., Pendleton, B., & Gupta, A. (2025). *Red teaming AI red teaming*. arXiv:2507.05538.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology, 4*(1), 103–120.

Meßmer, A.-K., & Degeling, M. (2023). *Auditing recommender systems: Putting the DSA into practice with a risk-scenario-based approach*. Policy Brief, Interface.

Metaxa, D., Park, J. S., Robertson, R. E., Karahalios, K., Wilson, C., Hancock, J., & Sandvig, C. (2021). Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction*, *14*(4), 272-344.

Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & Elish, M. C. (2021). Algorithmic impact assessments and accountability: The co-construction of impacts. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 735–746).

Mökander, J. (2023). Auditing of AI: Legal, ethical and technical approaches. *Digital Society, 2*(3), Article 49.

Mökander, J., & Floridi, L. (2023). Operationalising AI governance through ethics-based auditing: An industry case study. *AI and Ethics, 3*(2), 451–468.

Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021). Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics, 27*(4), Article 44.

Morales-Navarro, L., Kafai, Y. B., Vogelstein, L., Yu, E., & Metaxa, D. (2025). Learning about algorithm auditing in five steps: Scaffolding how high school youth can systematically and critically evaluate machine learning applications. *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 28, pp. 29186–29194).

Mouton, J., & Rottembourg, B. (2024). *Auditing the ranking strategy of a marketplace's algorithm in the frame of competition law commitments with surrogate models: The Amazon Buy Box case*. GREDEG Working Papers 2024-27.

Muralikumar, M. D., Yang, Y. S., & McDonald, D. W. (2023). A human-centered evaluation of a toxicity detection API: Testing transferability and unpacking latent attributes. *ACM Transactions on Social Computing, 6*(1–2), 1–38.

Murthy, D. (2021). Evaluating platform accountability: Terrorist content on YouTube. *American Behavioral Scientist, 65*(6), 800–824.

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society,* Article 146144481877305.

National Institute of Standards and Technology. (2024). *Artificial intelligence risk management framework: Generative artificial* intelligence profile (NIST AI 600-1). https://doi.org/10.6028/NIST.AI.600-1

Nguyen, T. T., Wilson, C., & Dalins, J. (2023). *Fine-tuning LLaMA 2 large language models for detecting online sexual predatory chats and abusive texts*. arXiv:2308.14683.

Nogara, G., Pierri, F., Cresci, S., Luceri, L., Törnberg, P., & Giordano, S. (2025). Toxic bias: Perspective API misreads German as more toxic. *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 19, pp. 1346–1357).

Ojewale, V., Steed, R., Vecchione, B., Birhane, A., & Raji, I. D. (2024). *Towards AI accountability infrastructure: Gaps and opportunities in AI audit tooling*. arXiv:2402.17861.

Panigutti, C., Yela, D. F., Porcaro, L., Bertrand, A., & Garrido, J. S. (2025). How to investigate algorithmic-driven risks in online platforms and search engines? A narrative review through the lens of the EU Digital Services Act. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 828–839).

Pattrn Analytics & Intelligence (2023). *Evaluating recommender systems in relation to the dissemination of illegal and harmful content in the UK*. Ofcom.

Perera, A., Aleti, A., Tantithamthavorn, C., Jiarpakdee, J., Turhan, B., Kuhn, L., & Walker, K. (2022). Search-based fairness testing for regression-based machine learning systems. *Empirical Software Engineering, 27*(3), Article 79.

Pierri, F., Luceri, L., Jindal, N., & Ferrara, E. (2023). Propaganda and misinformation on Facebook and Twitter during the Russian invasion of Ukraine. *Proceedings of the 15th ACM Web Science Conference 2023* (pp. 65–74).

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., et al. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44).

Raval, D. (2022). *Steering in one click: Platform self-preferencing in the amazon buy box*. Unpublished manuscript.

Reimers, I., & Waldfogel, J. (2023). *A framework for detection, measurement, and welfare analysis of platform bias.* NBER Working Paper No. w31766. National Bureau of Economic Research.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). *Auditing algorithms: Research methods for detecting discrimination on Internet platforms*. Paper presented to "Data and Discrimination: Converting Critical Concerns into Productive Inquiry," a preconference at the 64th Annual Meeting of the International Communication Association.

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* (pp. 1668–1678).

Sapiezynski, P., Ghosh, A., Kaplan, L., Rieke, A., & Mislove, A. (2022). Algorithms that "don't see color": Measuring biases in lookalike and special ad audiences. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 609–616).

Shen, H., DeVos, A., Eslami, M., & Holstein, K. (2021). Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2), 1–29.

Shukla, P., Chong, W. Y., Patel, Y., Schaffner, B., Pruthi, D., & Bhagoji, A. (2025). *Silencing empowerment, allowing bigotry: Auditing the moderation of hate speech on Twitch*. arXiv:2506.07667.

Smahel, D., Machackova, H., Mascheroni, G., Dedkova, L., Staksrud, E., Ólafsson, K., ... & Hasebrink, U. (2020). *EU Kids Online 2020: Survey results from 19 countries*. Doi: 10.21953/lse.47fdeqj01ofo.

Steinberg, D., Reid, A., & O'Callaghan, S. (2020). *Fairness measures for regression via probabilistic classification*. arXiv:2001.06089.

Steinebach, M. (2024). Robustness and collision-resistance of PhotoDNA. *Journal of Cyber Security and Mobility, 13*(3), 541–564.

Stray, J. (2022). Designing recommender systems to depolarize. *First Monday, 27*(5). https://doi.org/10.5210/fm.v27i5.12604

Tagharobi, H., & Simbeck, K. (2022). Introducing a framework for code based fairness audits of learning analytics systems on the example of Moodle learning analytics. *Proceedings of the 14th International Conference on Computer Supported Education – Volume 2: CSEDU* (pp. 45–55). SciTePress.

Teng, X. (2022). *Self-preferencing, quality provision, and welfare in mobile application markets* (No. 10042). CESifo Working Paper.

Timmaraju, A. S., Mashayekhi, M., Chen, M., Zeng, Q., Fettes, Q., Cheung, W., ... & Roudani, R. (2023). Towards fairness in personalized ads using impression variance aware reinforcement learning. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 4937-4947).

Urman, A., Makhortykh, M., & Hannak, A. (2024a). *Mapping the field of algorithm auditing: A systematic literature review identifying research trends, linguistic and geographical disparities*. arXiv preprint.

Urman, A., Smirnov, I., & Lasser, J. (2024b). The right to audit and power asymmetries in algorithm auditing. *EPJ Data Science, 13*(1), Article 19.

Ustun, B., Liu, Y., & Parkes, D. (2019). Fairness without harm: Decoupled classifiers with preference guarantees. *Proceedings of the International Conference on Machine Learning* (pp. 6373–6382). PMLR.

van Boheemen, P., Bösch, M., Costanzo, G., Divon, T., Frühwirth, L., Hammelburg, E., Klüser, J., Postma, L., Ring, E., Steffen, N., & Wang, X. (2025). *Live-Streaming: Mapping Networks of Influence and (Dis)information Flow.* Web publication or website. The Digital Methods Initiative.

Vitorino, P., Avila, S., Perez, M., & Rocha, A. (2018). Leveraging deep neural networks to fight child pornography in the age of social media. *Journal of Visual Communication and Image Representation, 50,* 303–313.

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review, 41*, 105567.

Waldfogel, J. (2024). *Amazon self-preferencing in the shadow of the digital markets act.* NBER Working Paper No. w32299. National Bureau of Economic Research.

Wang, H., Li, Y., Huang, R., & Mi, X. (2025). Detecting and understanding the promotion of illicit goods and services on Twitter. *Proceedings of the ACM Web Conference 2025* (pp. 3389–3404).

Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019) *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research*. Nuffield Foundation.

Ye, Z., Maragheh, R. Y., Morishetti, L., Vashishtha, S., Cho, J., Nag, K., … Achan, K. (2023). *Seller-side outcome fairness in online marketplaces*. arXiv:2312.03253.

Yesilada, M., & Lewandowsky, S. (2022). YouTube recommendations and problematic content: A systematic review. *Internet Policy Review, 11*(1), Article 1652.

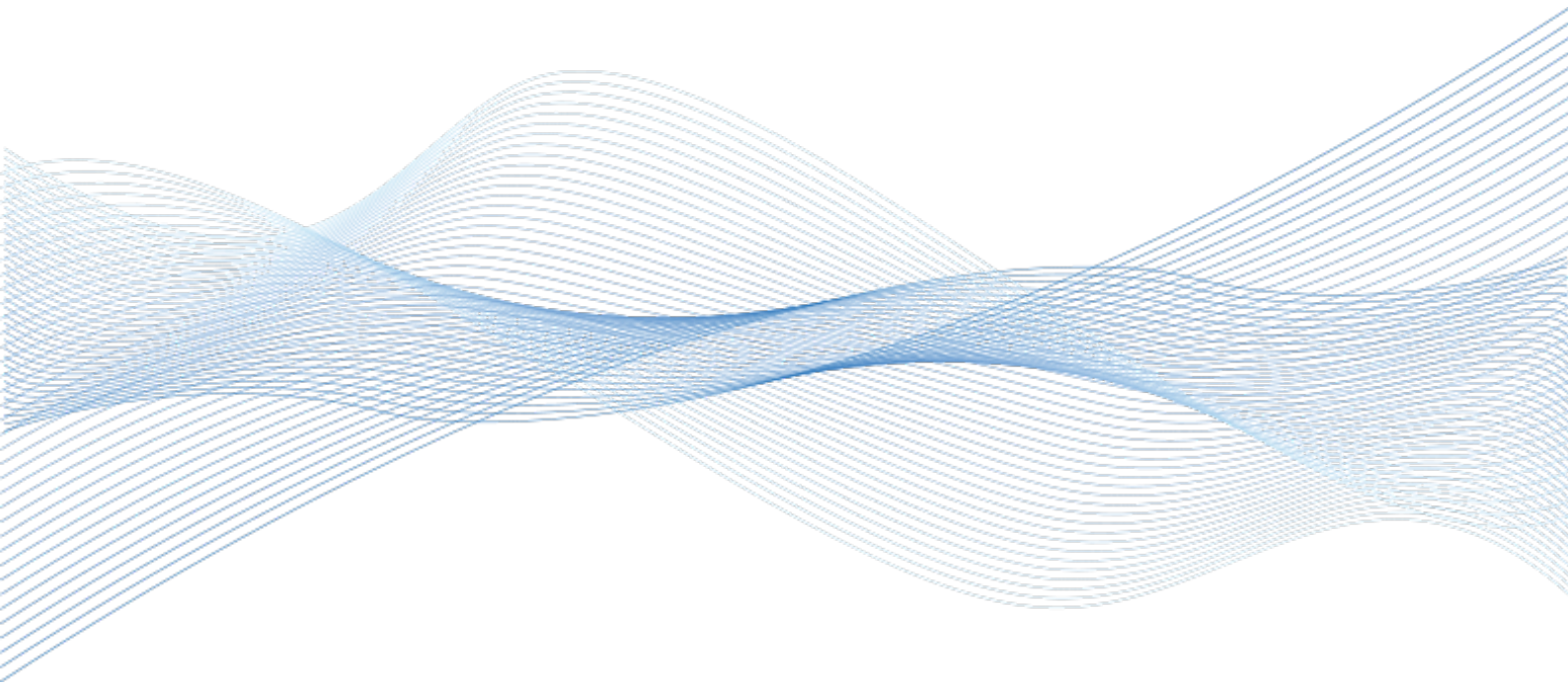
Zaccour, J., Binns, R., & Rocher, L. (2025). *Access denied: Meaningful data access for quantitative Algorithmus audits*. arXiv:2502.00428.

Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., & Weller, A. (2017). From parity to preference-based notions of fairness in classification. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.

Zehlike, M., Yang, K., & Stoyanovich, J. (2021). *Fairness in ranking: A survey*. arXiv:2103.14000.

Zehlike, M., Yang, K., & Stoyanovich, J. (2022). Fairness in ranking, part II: Learning-to-rank and recommender systems. *ACM Computing Surveys, 55*(6), 1–41.

Zicari, R. V., Amann, J., Bruneault, F., Coffee, M., Düdder, B., Hickman, E., et al. (2022). *How to assess trustworthy AI in practice.* arXiv:2206.09887.



WIK Wissenschaftliches Institut für
Infrastruktur und Kommunikationsdienste GmbH
Rhöndorfer Str. 68
53604 Bad Honnef, Germany
www.wik.org

ISSN 2750-5448 (Online)