Algorithmic Exclusions, Collective Inclusions: Surveying Algorithmic Harms and Collective Action in LGBTQIA2S+ and Marginalized Communities

Anonymous authors
Paper under double-blind review

Abstract

LGBTQIA2S+ and marginalized communities often face harms when interacting with algorithmic systems, such as misgendering, content suppression, and other forms of exclusion. In this paper, we examine the social context that enables these harms in LGBTQIA2S+ spaces, summarize the existing literature on algorithmic harms, and explore how communities can leverage collective action to regain their agency. We survey methods that exploit properties of machine learning systems, such as data dependence, adversarial vulnerability to resist these harms through collective action. We categorize existing approaches to resisting these harms, organized by four collective motivations: reporting and contesting harms (Model Auditing and Challenging Algorithmic Decision Strategies), opting out of model training or decision-making (Algorithmic Opt-Out Strategies), actively intervening to shift model behavior (Collective Intervention Strategies), and seeking recommendations for favorable outcomes (Decision Modification and Recommendation Strategies). Through a mapping review, we systematically chart where LGBTQIA2S+ and other marginalized communities appear in this literature and where they are absent. Our mapping reveals that while these communities are well-represented in platform-based strategies such as folk theorization and data activism, they are nearly absent from model-based methods such as adversarial techniques and algorithmic collective action, where machine learning researchers have focused their efforts. These gaps highlight opportunities for ML researchers and developers to build community-focused tools and methods that enable collectives to coordinate responses to algorithmic harms and regain agency over the systems that affect them. By mapping resistance methods across data-based, model-based, and platform-based mechanisms, we identify where the current literature falls short in supporting the communities most affected by algorithmic harms.

1 Introduction

Algorithmic systems and tools to automate human decision-making have become part of everyday life, from social media feeds and job applications to assisting with healthcare diagnosis. However, since these systems are embedded in the social and cultural context in which they are deployed, they also carry a higher risk of reinforcing and amplifying harm to the communities they are trying to serve. Recent works on algorithmic harms (Diberardino et al., 2024; Shelby et al., 2023) have shown that algorithmic systems can often reinforce biases present in their training data. Moreover, the process of developing learning pipeline involves a series of design choices, whether intentional, conventional, or arbitrary, that can lead to multiple equally performing models whose outcomes may differ across communities (Ganesh et al., 2025). This means that even when training data is representative, the choices made during model development and selection can produce systems whose impacts differ across communities. These harms are diverse in nature and could result in adverse consequences and even result in diminishing opportunities for those affected.

Marginalized communities, such as LGBTQIA2S+, can be vulnerable to algorithmic harms. For example, In 2020, Uber's `Real-Time ID Check` facial-recognition system repeatedly deactivated transgender drivers

after they updated their profile photos following gender transition (*Incident 396*; (McGregor, 2021)). Such incidents shows how pre-existing biases in data can cause algorithmic harms and result in opportunity loss for individuals. Harms faced by LGBTQIA2S+ communities can appear in various ways, for example, misgendering in AI-based systems (Dev et al., 2021), algorithmic suppression of queer content on digital platforms (Snyder, 2009), and invasive security screenings of Trans travelers recommended by AI-based systems (Waldron & Medina, 2019).

Often, communities affected by these harms develop strategies and workarounds to address them. DeVrio et al. (2024)'s work shows how people affected by algorithmic harms develop *responses from below* to regain agency in algorithmic systems. Such responses are possible because the learning algorithms underlying these systems depend on user data not only for initial training but also for continued adaptation and personalisation, giving users a degree of influence over model behaviour that Vincent et al. (2021) term *data leverage*. Recent work has also demonstrated and studied how communities resist algorithmic harms through various tactics, including Algorithmic Collective Action (Hardt et al., 2023), Protective Optimization Technologies (POT) (Kulynych et al., 2020), and Folk Theory of algorithms (Xiao et al., 2025; DeVito, 2022).

These responses to harm caused by algorithmic systems can be individual or collective in nature. In this work, we examine harms through the lens of the LGBTQIA2S+ community and the social context in which they occur. We also draw from the literature of algorithmic harms and collective action from other marginalized groups, as the literature focusing on specific harms and collective action in LGBTQIA2S+ is scarce. We categorize collective responses to algorithmic harms according to four key motivations. When users experience harms and wish to report, discuss, or contest them, they can use `Model Auditing and Challenging Algorithmic Decision Strategies`. However, users, after coming across the harms or noticing the potential of harms, may also wish to opt out of the algorithmic decision-making process by either stopping their data contribution or making their data unusable for model training or inference. We group these strategies under `Algorithmic Opt-out strategies`. Users may also strategically change their behaviour when interacting with algorithmic systems to intentionally get a favorable outcome from the system. We group these strategies under `Collective Intervention Strategies`. Lastly, we classify patterns where users leverage algorithmic recourse tools to understand what modifications to their inputs or behaviors may yield the desired outcome from the system as `Decision Modification and Recommendation Strategies`. To identify where these strategies have been studied or adopted by different communities and where gaps remain, we conduct a mapping review that charts the intersection of resistance methods with community representation across LGBTQIA2S+, other marginalized, general population, and theoretical contexts.

This paper contributes to an ongoing conversation about the responsible development of AI systems by centering the experiences and agency of marginalized communities. While methods to repair algorithmic harms currently exist, most rely on top-down approaches that depend on service providers' goodwill. We argue that this is not the only path for affected communities such as LGBTQIA2S+ to participate in mitigating these harms. Expanding research on community-led, grassroots resistance offers pathways for both understanding and mitigating harms. We map collective response methods across data-based, model-based, and platform-based mechanisms and identify where LGBTQIA2S+ and other marginalised communities are represented in this literature and where they are absent. The gaps our mapping reveals point to opportunities for ML researchers and system developers to build tools, and infrastructure that can support community-led efforts toward more accountable AI systems.

2 Algorithms as Barriers to LGBTQIA2S+ Representation

Online platforms have become essential infrastructures for interpersonal connection by enabling individuals to maintain relationships, share information, and express themselves across social and geographical boundaries. Platforms often rely on AI algorithms that actively shape content visibility, identity representation, and community interactions online. In fact, these AI algorithms influence social realities by reinforcing or challenging systematic marginalization, which impacts digital inclusion and equity (Shelby et al., 2023). AI systems trained on datasets from everyday life replicate existing structural biases, which creates algorithmic biases. These algorithmic biases lead to systematic disadvantages for affected communities and reinforce hierarchies of power and exclusion.

Marginalized communities often face many types of discrimination, such as racism, homophobia, transphobia, or ableism. Many rely on online platforms as safer spaces because these environments can offer visibility, support, and connections (Lucero, 2017). LGBTQIA2S+ groups, in particular, are disproportionately affected, as they continue to face pervasive online discrimination (Fisher et al., 2019). In recent years, LGBTQIA2S+ communities, and specifically trans and non-binary youth, have increasingly come under attack, notably through legislative initiatives aimed at restricting access to gender-affirming care across North America and Europe (Kraschel et al., 2022; Breckenkamp et al., 2022). In the United States, several policy changes have introduced new barriers to transgender healthcare, further entrenching marginalization and stigma (Redfield & Chokshi, 2025). Similar efforts to limit access to gender-affirming medical care have emerged in Canada and the United Kingdom (Wright et al., 2021; Ross et al., 2023).

Algorithmic Harms in LGBTQIA2S+ Spaces. Shelby et al. (2023) proposed a five-part taxonomy of algorithmic harms: *representational*, *allocative*, *quality-of-service*, *interpersonal*, and *social system* harms. We refer the reader to Shelby et al. (2023) for details of this taxonomy. Each of these harm types can be seen in LGBTQIA2S+ contexts in different ways. For instance, *representational harms* occur when socially constructed beliefs and unjust hierarchies about social groups become part of model inputs and outputs; these manifest when systems fail to recognise non-cisnormative gender identities (Scheuerman et al., 2019) or when toxicity classifiers label LGBTQ speech as toxic (Dias Oliva et al., 2021). *Allocative harms* occur when algorithmic systems produce an uneven distribution of resources, information, and opportunities; these have been observed in the disproportionate moderation and demonetization of queer content (Haimson et al., 2021; Ungless et al., 2024). *Quality-of-service harms* arise when a given demographic group experiences performance disparities compared to others; for example, safety features in LLMs designed to prevent representational harms are disproportionately triggered for certain demographic groups, causing the model to refuse to answer their questions (Chehbouni et al., 2024). *Interpersonal harms* occur when algorithmic systems adversely shape relations between people or communities; these have been linked to hate speech exposure that negatively affects mental health and willingness to interact in digital spaces (Schmid et al., 2024; Zochniak et al., 2023). *Social system harms* affect users at the macro level, including information harms, cultural harms, and political and civic harms; examples include the spread of misinformation about gender-affirming care (McNamara et al., 2024; Shuster et al., 2025).

Algorithmic Gatekeeping and Queer Visibility This growing marginalization has increased the vulnerability of LGBTQIA2S+ individuals to online harassment (Kinney et al., 2022). This situation highlights the need to further study the experiences and challenges these communities face, especially related to algorithmic biases and visibility on online platforms. In fact, these online platforms facilitate connection, but they can also replicate dominant societal norms by limiting visibility to LGBTQIA2S+ identities considered acceptable or recognizable. Algorithms amplify sanitized, marketable images such as able-bodied, cisgender, and conventionally attractive gay couples while sidelining more complex or non-conforming representations such as disabled or trans and gender-nonconforming individuals (Wang & Zhou, 2022). To assure visibility, this forces marginalized platform users to constantly adapt their content to algorithmic states, which perpetuates conditional visibility and constraints genuine recognition and self-expression (Cotter, 2019; Bishop, 2018). Since psychological well-being depends on genuine connection and recognition (Pullen Sansfaçon et al., 2021), filtering visibility disadvantages those who do not match normative ideals. These findings suggest that algorithmic platforms decide which queer stories will have visibility, linking directly to algorithmic exclusion: visibility itself becomes a gate-kept resource reserved for identities that conform to platform norms.

Algorithmic Literacy as Social Capital The uneven distribution of algorithmic literacy exacerbates these existing digital inequalities (Cotter & Reisdorf, 2020). In fact, the ability to effectively navigate platform-specific strategies such as optimal posting times, hashtag usage, and engagement tactics acts as social capital, which disproportionately benefits privileged users with greater access to digital resources while systematically disadvantaging marginalized users who may lack the same opportunities to understand or leverage these systems (Cotter, 2019). The notion of social capital, as conceptualized by Bourdieu, refers to the resources and advantages people can access through their membership in specific networks or group affiliations (Bourdieu, 1986). In the digital context, social capital refers to the ability to leverage knowl-

edge and connections specific to AI platforms to improve visibility and influence. Furthermore, persistent algorithmic issues such as involuntary inference of sexual orientation or gender identity (*SOGI*), biased moderation practices, misgendering, and cis-normative defaults further demonstrate the need for LGBTQIA2S+ communities to acquire algorithmic literacy as a tool for survival and visibility (Myles et al., 2023). For example, on TikTok, marginalized creators have faced content suppression based on factors such as queerness, race, and body size. Understanding this led communities to develop strategies, including camouflage and data observation, to counteract harmful algorithmic filtering (Ungless et al., 2024). This example shows the importance of algorithmic literacy because, to counteract these systematic disadvantages, communities must first recognize they are being marginalized, which enables them to mobilize collective action strategies to resist and reclaim their visibility.

Potential Issues with Relying on Repair from Above Regulatory processes may be slow or fail entirely. Canada's Bill C-27, which proposed a new regulatory framework for AI systems, was introduced on June 16, 2022, but died on January 6, 2025, when Parliament was prorogued (Gowling WLG, 2024). Relying on corporate goodwill is similarly limited: without legal obligations or independent oversight, companies often use *ethicswashing* strategies to fake ethical compliance (Schultz et al., 2025; Selbst, 2021).

Traditional algorithmic fairness approaches typically operate under a *provider-centric* model, where platforms are responsible for ensuring fair outcomes and users passively receive them. However, platforms may strategically manipulate fairness metrics to appear more equitable than they are. Aïvodji et al. (2019) introduced the concept of *fairwashing*, showing how explainability tools can be misused to construct a deceptive appearance of compliance. Similarly, Cirucci (2024) describes *privacy-washing*, where platforms project a misleading image of being privacy-friendly through ambiguous definitions and conflation of privacy with security. These examples highlight why users may reasonably doubt whether platforms will implement fairness methods in ways that truly align with community interests. These limitations underscore the need for more participatory approaches in which affected communities are actively involved in shaping AI systems.

3 Methods of Resisting Algorithmic Harms

When users encounter algorithmic harm, they can engage in various forms of resistance, individually or collectively. Users can do this by collectively making sense of the black-box algorithm, changing their behaviour when interacting with algorithms, modifying their data, or seeking algorithmic recourse that provides recommendations on the changes they need to make to their data to get favorable outputs. In this section, we define the resistance methods referenced throughout the paper, organised by the type of mechanism they employ. Some methods span multiple mechanism types; we define each once in its primary category and note where it also relates to others. Section 5 discusses how these methods map onto different collective goals and presents the evidence for their adoption by LGBTQIA2S+ and other marginalised communities (Table 1).

3.1 Data-Based Mechanisms

Data-based mechanisms exploit the dependency of AI systems on user-generated data. By modifying, withholding, redirecting, or producing data, users can influence what models learn and how they behave.

Data Leverage. AI systems are fundamentally data-driven, often depending on user data for both training and inference. This dependency creates an opportunity for users to exert influence over these systems by altering their data-related contributions. Vincent et al. (2021) refer to this leverage as *data leverage*, outlining several strategies through which users can exercise it. These include withholding or ceasing data contributions (*data strikes*), deliberately manipulating data (*data poisoning*), and purposefully sharing data with platforms that align with their values (*conscious data contribution*). Data leverage reframes the conversation around data-driven systems, centering the agency of the users in these systems to take action.

Data Refusal. While data strikes frames non-participation as an instrumental strategy to affect model performance, *data refusal* based on feminist and Indigenous scholarship to position non-participation as a political act that challenges the authority of data collectors (Zong & Matias, 2024). Zong & Matias (2024)

introduce a framework for *data refusal from below*, writing from the standpoint of people who refuse rather than the institutions that seek their compliance. They characterize refusal strategies across four facets: *autonomy* (whether refusal accounts for individual or collective interests), *time* (whether it reacts to past harm or proactively prevents future harm), *power* (the extent to which refusal makes change possible), and *cost* (whether refusal can reduce or redistribute penalties experienced by refusers). Data refusal encompasses but is not limited to data strikes; it also includes practices such as Indigenous data sovereignty, and legislative campaigns to ban harmful technologies.

Data Poisoning. Data poisoning attacks involve perturbing input samples so that, when they are used for training AI models, they negatively affect the model's performance (Biggio et al., 2012). These attacks can be categorized as untargeted or targeted poisoning, depending on the adversary's goal. Untargeted data poisoning aims to decrease the model's overall accuracy by injecting a small fraction of corrupted training samples, whereas targeted poisoning seeks to cause specific instances to be misclassified as an attacker-chosen target label rather than their true label. When coordinated by a collective, poisoning can shift model decisions in ways that benefit the group. Data poisoning also relates to model-based mechanisms (Section 3.2), as it exploits vulnerabilities in how models learn from training data.

Unlearnable Examples. To prevent unauthorized data exploitation, Huang et al. (2021) introduce the concept of *unlearnable examples*, which add error-minimizing noise to data samples so that machine learning models cannot extract meaningful patterns from them. These perturbations cause the model to treat the protected data as uninformative. Their method demonstrates strong effectiveness in image classification tasks and shows that the generated noise can transfer across datasets. This transferability makes the approach particularly valuable for user-driven resistance: communities can use noise generated from public datasets to protect their own private data.

Data Defences. *Data defences* (Agnew et al., 2024) enable data owners to prevent large language models (LLMs) from inferring personally identifying information (PII) from textual content. These defenses operate by embedding adversarial prompt injections, specifically crafted to prevent PII extraction, within the original text. Given that LLMs have been shown to memorize user characteristics and pose privacy risks (Staab et al., 2024), this strategy may be valuable for users who wish to publish text online while limiting how their information is used or profiled by AI systems.

Data Archival. Data archiving can serve as an act of grassroots resistance. Currie & Paris (2018) argue that preserving data over the long term is itself an activist project, discussing two literatures that have largely developed in isolation. They identify several shared affinities between archival activism and data activism: both respond to institutional neglect of marginalized perspectives, both seek to make overlooked issues visible and taken seriously in public discourse, and both push beyond standard ways of recording history and presenting statistical evidence. Related practices of counter-data production involve communities assembling their own case records and statistics not only to fill gaps left by official sources but also to reframe public narratives, influence policy, and support affected communities (D'Ignazio et al., 2025). These archival and counter-data practices are particularly relevant for communities whose experiences are systematically undercounted or misrepresented in training data, as they ensure that grassroots evidence is not only produced but also stewarded for future use.

3.2 Model-Based Mechanisms

Adversarial attacks are techniques that deliberately manipulate data or model behavior to exploit security properties such as integrity (causing incorrect predictions), availability (disrupting system functionality), or confidentiality (extracting sensitive information about the model or training data). Albert et al. (2021) discusses how such attacks have potential for good, especially in cases where these systems cause harm. They position their work as a call to action for the adversarial ML community to also investigate how the attacks can be repurposed as tools of resisting algorithmic harms.

Confidentiality Attacks. Confidentiality attacks compromise a platform's ability to secure its data and models. Platforms developing AI systems often treat their models as proprietary assets, making them accessible only through paid APIs. However, these models remain vulnerable to *model extraction attacks*, in which an adversary with black-box access to a prediction API attempts to reconstruct the underlying

model by using its predictions to train a substitute model (Tramèr et al., 2016). Aïvodji et al. (2020) further demonstrates that adversaries can leverage explanations provided by AI systems to make model-extraction attacks significantly more query-efficient. Model extraction can be repurposed for beneficial goals: affected users could employ an extracted model to facilitate independent auditing, or to enable conscious data contribution campaigns (Vincent & Hecht, 2021) by training a more community-aligned model (Libon et al., 2025). Moreover, an extracted model, now available in a white-box setting, can facilitate independent auditing, enabling greater scrutiny of model behaviour, biases, and potential harms.

Integrity Attacks. Integrity attacks aim to subvert or change the behaviour of algorithmic systems. Users wishing to correct algorithmic behaviour without relying on platforms may be motivated to repurpose these attacks to either evade algorithmic decisions or modify data to get favorable outcomes. *Adversarial examples* are perturbed inputs designed to "trick" machine learning models into making incorrect predictions at inference time (Goodfellow et al., 2015; Szegedy et al., 2014). These perturbations are often small, sometimes imperceptible to humans, yet sufficient to cause misclassification. They are also transferable, meaning that examples crafted to fool one model may also generalize to other models. One motivation for repurposing this technique is the use of adversarial examples to evade facial recognition systems (Shen et al., 2019). *Style cloaking* applies similar perturbations to creative works before they are shared online, preventing models from learning artistic styles or voice characteristics; tools such as Glaze (Shan et al., 2023) and AntiFake (Yu et al., 2023) implement this for unauthorized speech synthesis.

Availability Attacks. Availability attacks aim to compromise the reliability of a system or hinder users' access to it, effectively degrading its normal functionality. Shumailov et al. (2021) introduce *sponge examples*, inputs crafted to drastically increase energy consumption during inference. In autoregressive transformer pipelines, inference complexity grows with the number of input and output tokens. Sponge examples exploit this property by inflating tokenization and sequence lengths, thereby forcing models to perform significantly more computation than expected.

Strategic Classification. Individuals interacting with algorithmic systems could also try to "game" the system by modifying their data to get a favorable outcome. Hardt et al. (2016) formalize this idea within the framework of *strategic classification*, where individuals leverage information about the classifier to manipulate their attributes to achieve preferred decisions. The problem is modeled as a sequential game: a decision-maker publishes a classifier, and strategic individuals can then modify their inputs, at some cost, to obtain more favorable outcomes.

Algorithmic Collective Action. Algorithmic Collective Action (ACA), introduced by Hardt et al. (2023), provides a principled framework for analyzing how coordinated data modifications by groups of individuals can influence the behavior of deployed models. Given a collective data-modification strategy, the ACA framework can determine the minimum collective size required to achieve a desired level of success, as well as how success scales with collective size. Recent works have extended ACA to settings with multiple competing collectives (Karan et al., 2025) and differential privacy constraints (Solanki et al., 2025). Although ACA is grounded in how models respond to data modifications (Section 3.1), we place it here because the framework centres on exploiting model learning dynamics rather than on data practices themselves.

3.3 Platform-Based Mechanisms

Platform-based mechanisms are practices that emerge from users' everyday interactions with algorithmic systems. They typically require low technical skill but depend on shared knowledge and collective participation.

Folk-Theorisation and Sensemaking. When users encounter an algorithmic system, they can form an experiential understanding of these systems based on their interactions. This understanding, termed as *folk-theorisation* (Karizat et al., 2021), could involve trying various inputs, creating community knowledge, and trying to understand the working of the algorithms through trial and error. Folk theorisation could be used to form a shared understanding of the system. It could also be used by a collective to influence the outcomes of the algorithms they use to address the harms they face, serving as a tool for community resilience.

Everyday Algorithmic Auditing. Users, during their everyday interaction with algorithmic systems, may notice and report inaccuracies, discriminatory behavior, or harms that algorithms cause on the platform. Shen et al. (2021) describe the process where users of a system detect, understand, and interrogate harms of the system from their everyday interactions as *"everyday algorithm auditing"*. It offers a user-driven solution for communities to observe and report harms they encounter in their regular interaction with algorithmic systems.

Platform Migration and Conscious Data Contribution. Platform migration involves users collectively moving to alternative platforms, either as protest or to seek environments better aligned with their values. Migration can function as a form of data strike when it withdraws user-generated content and engagement from an incumbent platform, while simultaneously serving as conscious data contribution to a competitor. Platform migration thus also operates as a data-based mechanism (Section 3.1), as it directly affects the data available to both the incumbent and competing platforms.

3.4 Harm Reporting Infrastructure

Harm reporting infrastructure provides channels through which users can document and aggregate evidence of algorithmic failures. This includes platform-internal reporting mechanisms (e.g., flagging or appeals processes), as well as external databases such as the AI Incident Database (McGregor, 2021), which catalogs real-world AI failures to prevent their recurrence. Recent work has proposed reporting-based frameworks for identifying systematic algorithmic harms from individual user reports (Dai et al., 2025). Unlike everyday algorithmic auditing, which describes an organic user-driven process, harm reporting infrastructure refers to the *systems and tools* that facilitate the collection and aggregation of such evidence.

3.5 User-Facing Resistance Tools

Several of the mechanisms described above have been packaged into user-facing tools that lower the technical barrier to adoption. These tools bridge the gap between academic research on adversarial techniques and practical community use. Examples include *Glaze* (Shan et al., 2023) and *Nightshade* (Shan et al., 2024), which allow artists to apply style cloaking and data poisoning to their work before sharing it online; *Fawkes* (Shan et al., 2020) and *LowKey* (Cherepanova et al., 2021), which enable users to cloak photos against facial recognition systems; and *AdNauseam* (Howe & Nissenbaum, 2017), a browser extension that disrupts ad-tracking by automatically clicking ads in the background, effectively poisoning the data that targeted advertising algorithms rely on. The availability of such tools is a key factor in determining whether a resistance strategy can move from theoretical possibility to community adoption.

3.6 Legal, Institutional, and Policy Mechanisms

Algorithmic Abandonment. When users or media highlight discriminatory behavior, this can build public pressure on the platforms to correct this behaviour. In some cases, platforms might also decide to discontinue harmful algorithms. Johnson et al. (2024) define algorithmic abandonment as a decision made by actors with jurisdiction over the system to discontinue the process of developing, deploying, or using the algorithm due to its (potential) harms. They analyze 40 cases of real-world discontinuation and propose a six-stage taxonomy (*Discovery, Diagnosis, Dissemination, Dialogue, Decision, Death*). While abandonment campaigns often originate from platform-based practices such as everyday auditing and collective sensemaking (Section 3.3), the decision to discontinue a system is ultimately an institutional one.

Demanding Less Discriminatory Algorithms. AI systems often exhibit model multiplicity (Black et al., 2022), where several models achieve similar accuracy yet differ in their individual predictions or aggregate behaviors. If multiple models perform equally well, yet some produce less harm to affected groups, communities can legitimately demand that such alternatives be adopted. Black et al. (2023) argue that service providers should have a legal obligation to search for and identify less discriminatory algorithms (LDAs) that reduce disparate impact.

Algorithmic Recourse. Algorithmic recourse allows users who receive unfavorable outcomes not only to obtain explanations for those decisions but also to receive actionable recommendations for improving future

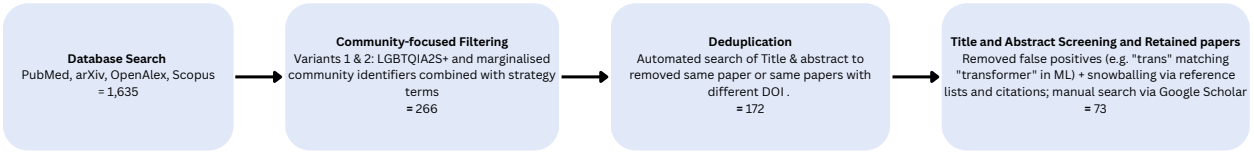

Figure 1: Flow diagram of the mapping review process. Community-focused queries (variants 1 and 2) yielded 266 results, which were deduplicated, screened for false positives, and merged with manually identified papers to produce 73 retained papers.

results (Karimi et al., 2023). A widely used form is *counterfactual explanations* (Wachter et al., 2018), which provide "what-if" scenarios by identifying minimal changes to an individual's input that would have led to a different decision. Creager & Zemel (2021) extend the concept of recourse by framing it as a form of collective action, enabling groups of users to coordinate feature modifications to shift system behavior for entire communities.

4 Methodology

To identify how the resistance strategies discussed in Section 3 have been studied and adopted in communities, we conduct a mapping review (Campbell et al., 2023). Our goal is to systematically chart the available evidence on the methods discussed in Section 3, surfacing where LGBTQIA2S+ and marginalised communities are represented in this literature and where gaps remain.

We searched four databases (PubMed, arXiv, OpenAlex, and Scopus) using programmatically generated queries. We first compiled a structured set of synonym lists for each of the resistance strategy categories (see rows of Table 1), drawing on the terminology used in the works discussed in Section 3. These synonym lists were stored in a structured configuration file and used as input to a search script that automatically constructed database-specific queries. The script adapted query syntax per database, wrapping Scopus queries in `TITLE-ABS-KEY()` and appending `[tiab]` field tags for PubMed, to restrict matches to titles, abstracts, and keywords. For each strategy, three query variants were generated: (1) strategy terms combined with LGBTQIA2S+ identifiers (e.g., `LGBT, queer, sexual minority", gender non-conforming"`), (2) strategy terms combined with broader marginalised-community identifiers, and (3) strategy terms combined with a sociotechnical relevance filter to capture general-population work related to algorithmic harms or resistance. An illustrative Scopus query is:

```
TITLE-ABS-KEY(("data strike" OR "data boycott" OR "data withholding" OR "data
refusal" OR "data protest" OR "data leverage") AND (LGBT OR LGBTQ OR "LGBTQ+"
OR LGBTQIA OR queer OR "sexual minority" OR "gender minority" OR "gender
non-conforming"))
```

The synonym lists were iteratively refined to remove terms that introduced false positives or conflated distinct concepts. For example, "machine unlearning" (removing data influence from trained models) was excluded from the unlearnable examples category (adding noise to prevent learning). The initial search yielded 144 from LGBTQIA2S+ queries, 122 from marginalised-community queries, and 1,363 from topic-only queries. The 1,363 topic-only results, which were screened separately, were used to identify representative papers for the General Population and Theoretical columns of Table 1. The 266 LGBTQIA2S+ and marginalised-community results were deduplicated by DOI and title similarity, producing 172 unique candidates. These were screened by title and abstract, removing papers that matched only due to false-positive keyword collisions. Some of the common source of false positives was use of search terms in different domains for example the term "trans" in LGBTQIA2S+ identifiers matching "transformer" or "transistor" in ML papers. After initial categorisation, we reviewed the reference lists of included papers as well as subsequent works that cited them, particularly in areas where our mapping revealed gaps, to identify additional relevant evidence that the database searches may have missed. After screening and including manually identified papers from

Google Scholar 73 papers were retained. Figure 1 shows the methodology flow chart. The search scripts, configuration files, and full discovery data will be released as an open-source toolkit on GitHub upon publication to support reproducibility and enable other researchers to extend this mapping review to additional communities or resistance strategies.

We categorized them into: (a) which resistance strategy it addresses, (b) whether LGBTQIA2S+ or marginalized communities are explicitly mentioned or studied, and (d) whether the paper documents real-world adoption or only academic/theoretical usage. The resulting examples are presented in Table 1.

Table 1: Literature Mapping Results: resistance methods × community representation. Representative papers per cell (max 3). Unmarked citations directly study the listed community; † indicates the community is mentioned or applicable but not the primary focus. Dashes indicate gaps in the literature.

Method	**LGBTQIA2S+**	**Other Marginalised**	**General Population**	**Theoretical**
Data-Based Mechanisms				
Data Leverage, Strikes & Refusal	Stevens & Doğan (2025); Doğan et al. (2025)	Zong & Matias (2024) †, Garcia et al. (2022)	Vincent et al. (2019)†, Schmitz & Samory (2025)	Vincent et al. (2021; 2019); Zong & Matias (2024)
Data Poisoning	—	—	Shan et al. (2024)	Biggio et al. (2012); Gupta et al. (2024)
Unlearnable Examples	—	—	Huang et al. (2021); Sun et al. (2022)	Huang et al. (2021)
Data Defences	—	—	Agnew et al. (2024) †	Agnew et al. (2024)
Data Archival & Data Activism	Doğan et al. (2025); Dig; Felkner et al. (2023)	D'Ignazio et al. (2025)	Currie & Paris (2018)†	Currie & Paris (2018)
Model-Based Mechanisms				
Integrity Attacks	—	Rosenberg et al. (2022)†, Khorzooghi et al. (2025)†	Shan et al. (2020)	Goodfellow et al. (2015); Szegedy et al. (2014)
Confidentiality Attacks	—	—	—	Tramèr et al. (2016); Aïvodji et al. (2020)
Availability Attacks	—	—	—	Shumailov et al. (2021)
Strategic Classification	—	—	—	Hardt et al. (2016)
Algorithmic Collective Action	Cho (2022)	Ben-Dov et al. (2025) †	Sigg et al. (2025)†, Xiao et al. (2025); Marasciulo (2022)	Hardt et al. (2023); Karan et al. (2025); Solanki et al. (2025)
Platform-Based Mechanisms				

Method	LGBTQIA2S+	Other Marginalised	General Population	Theoretical
Folk Theorisation & Sensemaking	DeVito (2022); Monea (2023)	Williams et al. (2025); Kojah et al. (2025)	Xiao et al. (2025)	Simpson et al. (2022); DeVito (2021)
Everyday Algorithmic Auditing	Dennler et al. (2023), Li et al. (2023)†	Li et al. (2023)†,	Shen et al. (2021); DeVos et al. (2022); Attenberg et al. (2015)	Shen et al. (2021)†
Conscious Data Contribution & Platform Migration	Pan et al. (2025)	Pan et al. (2025)†	Yuan et al. (2025); Quelle et al. (2026); Schmitz & Samory (2025)	Vincent & Hecht (2021)†
Harm Reporting Infrastructure				
Incident Databases & Reporting Frameworks	—	—	McGregor (2021)	Dai et al. (2025)
User-Facing Resistance Tools				
User-Facing Resistance Tools	—	Jiang et al. (2026); Howe & Nissenbaum (2017)	Shan et al. (2023; 2024; 2020)	—
Legal, Institutional & Policy Mechanisms				
Less Discriminatory Algorithms or Legal Contestation via Class Action Lawsuits	Divino Group LLC et al. (2019)	United States Department of Justice, Civil Rights Division (2023); U.S. Equal Employment Opportunity Commission (2024)	Schor (2024); Wodecki (2024)	Black et al. (2023)
Algorithmic Abandonment	—	Buolamwini & Gebru (2018)	Johnson et al. (2024)†	Johnson et al. (2024)
Algorithmic Recourse	—	—	—	Karimi et al. (2023); Creager & Zemel (2021)

Summary of the gaps and evidence. Several patterns emerge from Table 1. LGBTQIA2S+ communities are well-represented in platform-based mechanisms, particularly folk theorisation, and in data-based strategies such as data refusal and data activism. These are strategies that emerge organically from everyday platform interactions and could explain wider adoption by various communities. Model-based mechanisms have little to no LGBTQIA2S+ evidence; the only exception is Cho (2022), which documents a form of algorithmic platform resistance rather than direct use of adversarial ML techniques. Several model-based methods, including sponge examples, model extraction, and strategic classification, remain purely theoretical

across all community columns, with no documented real-world adoption. These gaps point to a need for research that bridges the divide between the technical methods studied in adversarial ML and the community practices already in use, as well as for tools and educational resources that can make technical resistance strategies more accessible to affected communities.

5 Motivation and Goals of The Collective

While individuals may develop informal understandings of how algorithms function, collectives can engage in more coordinated efforts to influence these systems. The methods defined in Section 3 serve different collective goals depending on the community's intent: making harms visible, opting out of harmful systems, actively intervening to change model behaviour, or seeking recommendations for more favourable outcomes. Table 1 presents our evidence and gap map, revealing where these strategies have been studied or adopted across LGBTQIA2S+ communities, other marginalised groups, general populations, and academic contexts. A striking pattern emerges: LGBTQIA2S+ communities are well-represented in platform-based and data-based strategies (folk theorisation, content evasion, data activism) but almost entirely absent from model-based mechanisms (adversarial examples, algorithmic collective action, recourse). This gap aligns with the skill barriers summarised in Table 2.

5.1 Model Auditing and Challenging Algorithmic Decisions

When a group of people with similar protected attributes receive adverse outputs compared to others, the resulting disparate impact can include disproportionate content moderation or demonetisation of queer content (Haimson et al., 2021; Ungless et al., 2024) and AI systems trained with cis-normative data failing for transgender individuals (Scheuerman et al., 2019). Mechanisms such as everyday algorithmic auditing (Section 3.3) enable users to surface patterns of harm and collectively document adverse impacts. The evidence gathered through these practices can then be used to contest harmful outcomes and demand changes from service providers.

Recent work shows that AI systems often exhibit model multiplicity (Black et al., 2022; Ganesh et al., 2025), where several models achieve similar accuracy yet differ in their predictions. This arbitrariness provides leverage for collectives: if multiple models perform equally well yet some produce less harm, communities can demand that less discriminatory alternatives be adopted (Black et al., 2023). Contestation can also result in algorithmic abandonment (Johnson et al., 2024). A notable example occurred in 2020, when Twitter users observed that the platform's image-cropping algorithm disproportionately favored white faces. Users collectively examined the issue and built on one another's findings (Hern, 2020; Shen et al., 2021), and Twitter ultimately decided to stop using the algorithm (Chowdhury, 2021).

5.2 Algorithmic Opt-Out Strategies

Collectives may also wish to take preventative steps to reduce harms. The widespread use of publicly available user data to train models has raised concerns about privacy (Press, 2025; Rucker, 2025), copyright infringement (Brittain, 2025; Schor, 2024; Wodecki, 2024), and the potential misuse of personal or creative content. Opt-out strategies are exemplified by tools such as Glaze (Shan et al., 2023), which applies style cloaks to artworks before they are shared online, making them difficult for models to learn from while remaining visually similar to human viewers. For many artists, posting their work online is essential for visibility, so fully withholding art from public platforms is often not viable; style cloaking and related techniques (Section 3.2) offer an alternative. Different opt-out strategies are shown in Figure 2.

5.3 Collective Intervention Strategies

These tactics help users prevent misuse of their data, but they may not be sufficient to address or correct harmful decisions made by deployed models. Collectives can also use coordinated interventions to steer algorithmic systems toward more equitable outcomes. The ACA framework (Hardt et al., 2023) provides theoretical tools to analyze how such interventions unfold: factors such as users' access to the system, the

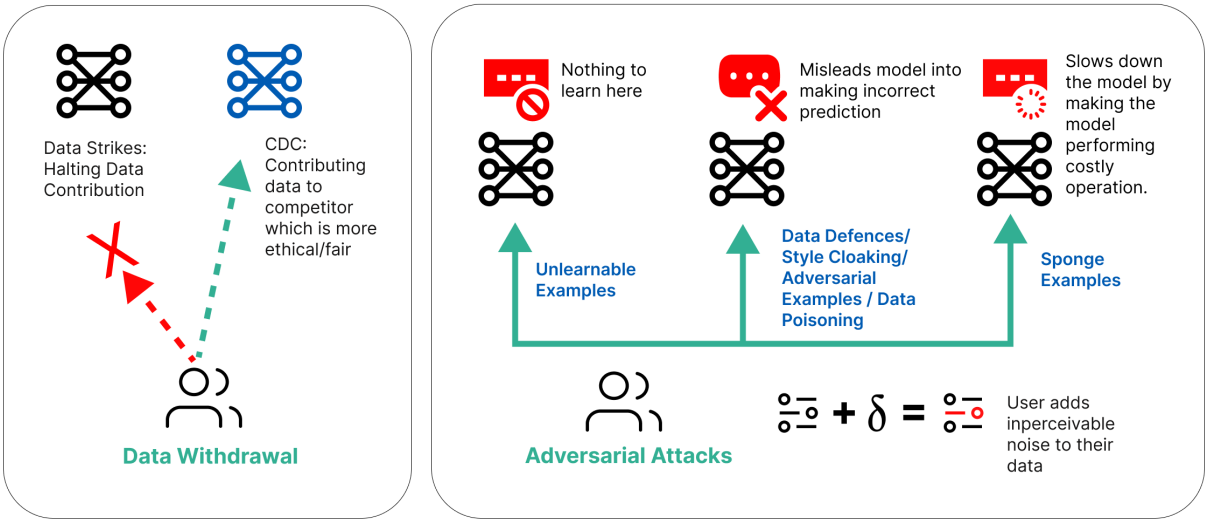

Figure 2: Data withdrawal and adversarial-attack strategies. Left: *Data withdrawal*, where users halt or redirect their data contributions through data strikes (stopping data contribution) or conscious data contribution (sending data to a more ethical or fair competitor). Right: *Adversarial attacks*, where users add a small perturbation δ to their data to (a) create unlearnable examples that provide "nothing to learn," (b) use data defences / style cloaking / adversarial examples / data poisoning to mislead the model's predictions, and (c) craft sponge examples that slow down the model by making inference more computationally costly.

type of model deployed, and the structural characteristics of the collective all influence the critical mass required for successful action. Individual users may also seek to "correct, shift, or expose harms" without relying on service providers to intervene (Kulynych et al., 2020).

These forms of collective intervention already appear in real-world cases. The Decline Now campaign (Sigg et al., 2025) organized DoorDash drivers to collectively decline low-payout orders, thereby increasing future payouts through coordinated behavior. DeVrio et al. (2024) document similar "responses from below" to algorithmic harm. Another example is AdNauseam (Howe & Nissenbaum, 2017), a browser extension that disrupts ad-tracking by automatically clicking ads, effectively poisoning the data that targeted advertising algorithms rely on.

5.4 Decision Modification and Recommendation Strategies

Algorithmic recourse (Section 3.6) allows users who receive unfavorable outcomes to receive actionable recommendations for improving future results. Creager & Zemel (2021) extend recourse to a collective setting, where coordinated feature modifications can shift the system's behavior for entire communities. Beyond recourse, collectives can repurpose model extraction techniques (Aïvodji et al., 2020) to reconstruct an approximate version of a deployed model, enabling more informed and strategic forms of resistance.

These practices already manifest in real-world contexts: job seekers routinely optimize their résumés to be "ATS-friendly" (Purcell, 2025). In marginalized communities such as LGBTQIA2S+ groups, coordinated strategies for decision modification and collective recourse can counteract content moderation bias, shadow banning, and algorithmic suppression. Instead of relying solely on individual appeals, collective recourse offers a means of pushing platforms to become more inclusive and responsive to marginalized expressions.

Table 2 summarizes how different resistance methods align with (i) the level of technical or legal skill required (Column: `Low Skill Barrier?`), (ii) whether they can be meaningfully used by individuals acting alone (Column: `Co-ordination required?`), and (iii) who primarily benefits from the method — the individual

participant, the broader group, or both (Column: `Primarily Benefits`). For example, everyday algorithmic auditing requires low skill but is not effective as a purely individual strategy, whereas data poisoning has a higher technical barrier and is most effective when a coordinated group is willing to modify their data in ways that may not immediately benefit each person individually but can shift model behavior in favor of the collective.

Table 2: Ways in which collective resistance methods and strategies align with the goals, skills, and access of affected collectives.

Methods and Strategies	**Objective / Goal**	**Low Skill Barrier?**	**Co-ordination required?**	**Primarily Benefits**
Data-Based Mechanisms				
Data Strikes (Vincent et al., 2019)	Opt-out	Yes	Yes	Both
Data Refusal (Zong & Matias, 2024)	Opt-out	Yes	Partially	Both
Data Poisoning (Suya et al., 2021; Vincent et al., 2021)	Opt-out & Intervention	No	Partially	Both
Unlearnable Examples (Huang et al., 2021)	Opt-out	No	No	Individual
Data Defences (Agnew et al., 2024)	Opt-out	Partially	No	Individual
Counter-Data & Data Archival (D'Ignazio et al., 2025; Currie & Paris, 2018)	Intervention	Partially	Yes	Collective
Model-Based Mechanisms				
Adversarial Examples (Goodfellow et al., 2015; Szegedy et al., 2014)	Opt-out & Intervention	No	No	Individual
Style Cloaking (Shan et al., 2023; Yu et al., 2023)	Opt-out	Partially	No	Individual
Sponge Examples (Shumailov et al., 2021)	Opt-out	No	Yes	Collective
Using Model Extraction for generating explanations (Aïvodji et al., 2020)	Decision Modification	No	No	Both
Algorithmic Collective Action (Hardt et al., 2023; Karan et al., 2025)	Intervention	No	Yes	Collective
Platform-Based Mechanisms				
Folk Theorisation & Sensemaking (Karizat et al., 2021)	Auditing & Challenging	Yes	Partially	Both

Continued on next page

Methods and Strategies	Objective / Goal	Low Skill Barrier?	Co-ordination required?	Primarily Benefits
Everyday Algorithmic Auditing (Shen et al., 2021)	Auditing & Challenging	Yes	Yes	Collective
Platform Migration	Opt-out	Yes	Partially	Both
Content Evasion / Algospeak	Opt-out	Yes	No	Individual
Harm Reporting Infrastructure				
Incident Databases & Reporting (McGregor, 2021)	Auditing & Challenging	Yes	Partially	Collective
User-Facing Resistance Tools				
Privacy & Anti-Tracking Tools (Howe & Nissenbaum, 2017)	Opt-out	Yes	No	Individual
Creative Protection Tools (e.g Glaze) (Shan et al., 2023)	Opt-out	Yes	No	Individual
Legal, Institutional & Policy Mechanisms				
Algorithmic Abandonment (Johnson et al., 2024)	Auditing & Challenging	Yes	Yes	Collective
Less Discriminatory Algorithms (Black et al., 2023)	Auditing & Challenging	No	Yes	Collective
Legal Contestation & Appeals	Decision Modification	Partially	No	Individual
Algorithmic Recourse (Creager & Zemel, 2021)	Decision Modification	Partially	Partially	Both

Primarily Benefits: *Individual* = the acting user is the primary beneficiary; *Collective* = benefits accrue to the group or community, often requiring coordination to be effective; *Both* = can benefit the individual directly while also producing collective effects at scale.

The table summarises how different resistance methods align with (i) the level of technical or legal skill required, (ii) whether coordination among multiple participants is needed for the method to be effective, and (iii) who primarily benefits when the method succeeds.

6 Discussion

6.1 Resilience and resistance in the face of AI: drawing from AI itself

Strategies for collective resistance against algorithmic harms empower individuals, especially those from marginalized communities, to push back against AI systems that treat them unfairly. They can adopt a variety of tactics, including flagging biased content, flooding training datasets with counterexamples, and more. Some strategies even co-opt the logic of AI systems themselves, using their internal dynamics to challenge and expose biases, thereby compelling platforms to acknowledge and address algorithmic injustices.

However, simply knowing and applying these strategies does not offer a complete or lasting solution. Because AI systems are continually evolving, these forms of resistance often encounter opposition and degradation over time. For instance, platforms can update their defenses, neutralize known resistance patterns, or adjust model parameters to reassert control. Additionally, bad actors may appropriate ACA-inspired tactics for malicious purposes, undermining their legitimacy and impact. As a result, resilience and adaptability are critical. Communities must continuously recalibrate their resistance strategies to keep pace with shifting technological dynamics and ensure the ongoing protection of their rights and interests. This highlights the need for sustained collective effort, infrastructure, and knowledge-sharing practices to support long-term resistance against algorithmic exclusion and harms.

6.2 Challenges and risks of collective action

These strategies, while designed to empower communities, can also be used by malicious actors seeking to exploit or harm collectives rather than protect them, undermining trust, safety, and legitimacy of these strategies within affected communities. Furthermore, when multiple collectives employ ACA tactics with competing goals or conflicting methods, this may lead to fragmentation or counterproductive interference between groups (Karan et al., 2025). Another critical challenge lies in the varying efficacy of ACA tactics across different AI systems. Their effectiveness may be compromised by factors such as the system's adaptability to interventions, or characteristics of the underlying training data (Solanki et al., 2025). Tactics that succeed against one model may fail or be neutralized in others. Importantly, many ACA techniques overlap with those used in malicious data poisoning.

While collectives may use model manipulation to resist harm, similar strategies can be deployed with harmful intent. Conversely, defenses developed to counter adversarial or targeted data poisoning may inadvertently suppress legitimate collective action, limiting users' ability to contest algorithmic decisions. Lastly, ACA efforts are also susceptible to the free-rider problem (Olson, 1971), where some benefit from the collective effort without contributing to it. This raises important questions about whether and how the classic dynamics of free-riding manifest in the context of algorithmic resistance, and what mechanisms might encourage sustained and equitable participation.

6.3 Balancing allyship and risks: research responsibility when research involves marginalized communities

Faced with the complexities of this situation, we argue that researchers must navigate the right balance between transparency and protection of tactics. In fact, we argue that there is a tension between democratizing knowledge about ACAs and the need to protect the communities these same strategies aim to defend. In fact, there is a risk that sharing these ACA tactics could expose marginalized communities to retaliation or harm. This concern is particularly relevant for the LGBTQIA2S+ communities because they face targeted surveillance and exploitation in online spaces (Tanni et al., 2024). Different groups within the LGBTQIA2S+ who are especially targeted for online surveillance, harassment, and deplatforming include transgender, two-spirit, and racialised individuals (Turner et al., 2024).

We argue that, as researchers aiming to act as allies, we must be vigilant to ensure our research is not used to undermine or exploit the very communities we aim to support. We argue that following ethical guidelines might help us better navigate these challenges. For example, the Canadian Professional Association for Transgender Health (CPATH) offers principles of ethical guidelines for research involving transgender communities that can help guide research practices involving marginalized communities. CPATH's ethical guidelines explain how to engage communities in participatory research (Bauer et al., 2019).

6.4 Participatory research as a form of empowerment

An important consideration for future work is democratizing research by engaging marginalized communities to actively participate. Engaging LGBTQIA2S+ communities not as subjects but as co-researchers ensures that their lived experiences and expertise directly inform the research. This consideration is essential in research on resistance strategies because it is grounded in the principle of empowerment.

6.5 Contributing to empowerment: building algorithmic literacy

We argue that a concrete next step would be the development and diffusion of accessible educational tools or instructional resources that help build algorithmic literacy. In fact, systematic biases and barriers like algorithmic harms, algorithmic gatekeeping, and unequal representation cannot be identified or challenged without a baseline understanding of how algorithmic systems operate (Gagrčin et al., 2024). Algorithmic literacy can equip individuals from marginalized communities with the skills and tools to interpret and recognize how algorithmic systems work against them. Algorithmic literacy helps in revealing these systemic biases by understanding AI systems, but provides communities with means to act on them, including strategically adapting their practices, engaging in collective forms of resistance, and advocating for more accountability and fairness within AI systems.

However, many factors that influence accessibility to these different forms of resistance need to be taken into consideration, such as digital access, age, education level, socioeconomic status, and much more. These different factors influence how individuals can engage with algorithmic systems. In addition, we argue that it is necessary to recognize and support different forms of resistance. While some forms of resistance are more technical, like data poisoning, other techniques are discursive, like reclaiming hashtags, documenting algorithmic harms, or building solidarity networks. Recognising the different ways collectives can resist is important and validates different ways marginalized communities can protect themselves against algorithmic harms.

7 Conclusion and Future Work

In conclusion, in this paper, we looked at how AI systems shape the online experiences of LGBTQIA2S+ communities by looking at the types of harms they experience and how they resist them by using tactics of ACAs. An important reflection that emerged from our analysis was the fact that to be able to better resist harms, affected communities need the knowledge and tools to recognize these harms and navigate these AI systems. These findings highlight the importance of algorithmic literacy.

Nevertheless, to be able to build algorithmic literacy, there is first the need to understand where people currently stand in their understanding of algorithmic systems and how they can affect them. What emerged in our research is a clear gap. In fact, while many tactics of resistance are documented, there is a lack of literature on how communities currently understand these AI systems. This lack of knowledge regarding the current state of algorithmic literacy makes it difficult to adapt tactics of resistance to make them more accessible and effective.

To address this, future research could explore the current state of algorithmic literacy regarding ACAs, feeling of efficacy among populations, access to tools, and resistance tactics. This would allow the development of a baseline of knowledge to use as a foundation for future work aiming at expanding algorithmic literacy. This type of research can contribute to empowering marginalized communities and allow them to develop the tools needed to better resist and shape algorithmic systems.

References

Digital Transgender Archive. https://www.digitaltransgenderarchive.net/.

Today's Gender Is No: Genderbot's Algorithmic Platform Resistance: TOPIA: Vol 48. *TOPIA: Canadian Journal of Cultural Studies*, February 2022.

W. Agnew, H. H. Jiang, C. Sum, M. Sap, and S. Das. Data defenses against large language models, 2024.

U. Aïvodji, A. Bolot, and S. Gambs. Model extraction from counterfactual explanations, 2020.

Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing: The risk of rationalization. In *International Conference on Machine Learning*, pp. 161–170. PMLR, 2019.

Kendra Albert, Maggie Delano, Bogdan Kulynych, and Ram Shankar Siva Kumar. Adversarial for Good? How the Adversarial ML Community's Values Impede Socially Beneficial Uses of Attacks. In *ICML 2021 Workshop on Adversarial Machine Learning*, June 2021.

Joshua Attenberg, Panos Ipeirotis, and Foster Provost. Beat the Machine: Challenging Humans to Find a Predictive Model's "Unknown Unknowns". *J. Data and Information Quality*, 6(1):1:1–1:17, March 2015. ISSN 1936-1955. doi: 10.1145/2700832.

G. Bauer, A. Devor, M. Heinz, Z. Marshall, A. Pullen Sansfaçon, and J. Pyne. CPATH ethical guidelines for research involving transgender people & communities. http://cpath.ca/en/resources/, 2019.

Omri Ben-Dov, Samira Samadi, Amartya Sanyal, and Alexandru Ţifrea. Fairness for the People, by the People: Minority Collective Action, November 2025.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 1467–1474, Madison, WI, USA, June 2012. Omnipress. ISBN 978-1-4503-1285-1.

Sophie Bishop. Anxiety, panic and self-optimization: Inequalities and the YouTube algorithm. *Convergence: The International Journal of Research into New Media Technologies*, 24(1):69–84, February 2018. ISSN 1354-8565, 1748-7382. doi: 10.1177/1354856517736978.

E. Black, M. Raghavan, and S. Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 850–863, 2022. doi: 10.1145/3531146.3533149.

Emily Black, John Logan Koepke, Pauline Kim, Solon Barocas, and Mingwei Hsu. Less Discriminatory Algorithms. October 2023. doi: 10.2139/ssrn.4590481.

P. Bourdieu. The forms of capital. In J. G. Richardson (ed.), *Handbook of Theory and Research for the Sociology of Education*, pp. 241–258. Greenwood Press, 1986.

J. Breckenkamp, J. Thirugnanamohan, A. Stern, O. Razum, and Y. Namer. Trans* people's access to gender-affirming health care: A European comparison. In *European Journal of Public Health*, volume 32, 2022. doi: 10.1093/eurpub/ckac129.070.

B. Brittain. Anthropic wins key US ruling on AI training in authors' copyright lawsuit. June 2025.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81, pp. 77–91. PMLR, 2018. doi: 10.48550/arxiv.1712.05382.

Fiona Campbell, Andrea C. Tricco, Zachary Munn, Danielle Pollock, Ashrita Saran, Anthea Sutton, Howard White, and Hanan Khalil. Mapping reviews, scoping reviews, and evidence and gap maps (EGMs): The same but different- the "Big Picture" review family. *Systematic Reviews*, 12(1):45, March 2023. ISSN 2046-4053. doi: 10.1186/s13643-023-02178-5.

Khaoula Chehbouni, Megha Roshan, Emmanuel Ma, Futian Wei, Afaf Taik, Jackie Cheung, and Golnoosh Farnadi. From Representational Harms to Quality-of-Service Harms: A Case Study on Llama 2 Safety Safeguards. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15694–15710, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.927.

Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John Dickerson, Gavin Taylor, and Tom Goldstein. LowKey: Leveraging Adversarial Attacks to Protect Social Media Users from Facial Recognition, January 2021.

R. Chowdhury. Sharing learnings about our image cropping algorithm. X Engineering Blog, May 2021.

Angela M. Cirucci. Oversharing the super safe stuff: "Privacy-washing" in Apple iPhone and Google Pixel commercials. *First Monday*, May 2024. ISSN 1396-0466. doi: 10.5210/fm.v29i5.13321.

K. Cotter. Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. *New Media & Society*, 21(4):895–913, 2019. doi: 10.1177/1461444818815684.

K. Cotter and B. C. Reisdorf. Algorithmic knowledge gaps: A new dimension of (digital) inequality. *International Journal of Communication*, 14:745–765, 2020.

E. Creager and R. Zemel. Online algorithmic recourse by collective action, 2021.

Morgan E. Currie and Britt S. Paris. Back-ups for the future: Archival practices for data activism. *Archives and Manuscripts*, 46(2):124–142, May 2018. ISSN 0157-6895, 2164-6058. doi: 10.1080/01576895.2018.1468273.

Jessica Dai, Paula Gradu, Inioluwa Deborah Raji, and Benjamin Recht. From Individual Experience to Collective Evidence: A Reporting-Based Framework for Identifying Systemic Harms. In *Forty-Second International Conference on Machine Learning*, June 2025.

Nathan Dennler, Anaelia Ovalle, Ashwin Singh, Luca Soldaini, Arjun Subramonian, Huy Tu, William Agnew, Avijit Ghosh, Kyra Yee, Irene Font Peradejordi, Zeerak Talat, Mayra Russo, and Jess De Jesus De Pinho Pinhal. Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pp. 375–386, New York, NY, USA, August 2023. Association for Computing Machinery. ISBN 979-8-4007-0231-0. doi: 10.1145/3600211.3604682.

S. Dev, M. Monajatipoor, A. Ovalle, A. Subramonian, J. Phillips, and K.-W. Chang. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1968–1994, 2021. doi: 10.18653/v1/2021.emnlp-main.150.

Michael Ann DeVito. Adaptive Folk Theorization as a Path to Algorithmic Literacy on Changing Platforms. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2):339:1–339:38, October 2021. doi: 10.1145/3476080.

Michael Ann DeVito. How transfeminine TikTok creators navigate the algorithmic trap of visibility via folk theorization. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), November 2022. doi: 10.1145/3555105.

Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *CHI Conference on Human Factors in Computing Systems*, pp. 1–19, New Orleans LA USA, April 2022. ACM. ISBN 978-1-4503-9157-3. doi: 10.1145/3491102.3517441.

A. DeVrio, M. Eslami, and K. Holstein. Building, shifting, & employing power: A taxonomy of responses from below to algorithmic harm. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1093–1106, 2024. doi: 10.1145/3630106.3658958.

T. Dias Oliva, D. M. Antonialli, and A. Gomes. Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LGBTQ voices online. *Sexuality & Culture*, 25(2):700–732, 2021. doi: 10.1007/s12119-020-09790-w.

Nathalie Diberardino, Clair Baleshta, and Luke Stark. Algorithmic Harms and Algorithmic Wrongs. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1725–1732, Rio de Janeiro Brazil, June 2024. ACM. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3659001.

Catherine D'Ignazio, Isadora Cruxên, Angeles Martinez Cuba, Helena Suárez Val, Amelia Dogan, and Natasha Ansari. Geographies of missing data: Spatializing counterdata production against feminicide. *Environment and Planning D: Society and Space*, 43(1):29–50, February 2025. ISSN 0263-7758, 1472-3433. doi: 10.1177/02637758241275961.

Divino Group LLC et al. Divino group LLC et al. v. google/YouTube: Class action complaint, 2019.

Amelia Lee Doğan, Nikko Stevens, and Catherine D'Ignazio. Trans data: A research and design agenda from trans activists' transformative data science. *Proc. ACM Hum.-Comput. Interact.*, 9(7), October 2025. doi: 10.1145/3757682.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. WinoQueer: A community-in-the-loop benchmark for anti-LGBTQ+ bias in large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9126–9140, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.507.

C. M. Fisher, M. R. Woodford, R. E. Gartner, P. R. Sterzing, and B. G. Victor. Advancing research on LGBTQ microaggressions: A psychometric scoping review of measures. *Journal of Homosexuality*, 66(10): 1345–1379, 2019. doi: 10.1080/00918369.2018.1539581.

E. Gagrčin, T. K. Naab, and M. F. Grub. Algorithmic media use and algorithm literacy: An integrative literature review. *New Media & Society*, 2024. doi: 10.1177/14614448241291137.

Prakhar Ganesh, Afaf Taik, and Golnoosh Farnadi. Systemizing Multiplicity: The Curious Case of Arbitrariness in Machine Learning. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8 (2):1032–1048, October 2025. ISSN 3065-8365. doi: 10.1609/aies.v8i2.36609.

Patricia Garcia, Tonia Sutherland, Niloufar Salehi, Marika Cifor, and Anubha Singh. No! Re-imagining data practices through the lens of critical refusal. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), November 2022. doi: 10.1145/3557997.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations 2015*. arXiv, March 2015. doi: 10.48550/arXiv.1412.6572.

Gowling WLG. Bill C-27: Timeline of developments. 2024.

Isha Gupta, Hidde Lycklama, Emanuel Opel, Evan Rose, and Anwar Hithnawi. Fragile Giants: Understanding the Susceptibility of Models to Subpopulation Attacks, October 2024.

O. L. Haimson, D. Delmonaco, P. Nie, and A. Wegner. 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, 2021. doi: 10.1145/3479610.

M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pp. 111–122, 2016. doi: 10.1145/2840728.2840730.

Moritz Hardt, Eric Mazumdar, Celestine Mendler-Dünner, and Tijana Zrnic. Algorithmic collective action in machine learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pp. 12570–12586, Honolulu, Hawaii, USA, July 2023. JMLR.org.

A. Hern. Twitter apologises for "racist" image-cropping algorithm. September 2020.

D. C. Howe and H. Nissenbaum. Engineering privacy and protest. In *3rd International Workshop on Privacy Engineering (IWPE 2017)*, volume 1873 of *CEUR Workshop Proceedings*, pp. 57–64, 2017.

H. Huang, X. Ma, S. M. Erfani, J. Bailey, and Y. Wang. Unlearnable examples: Making personal data unexploitable, 2021.

Harry H. Jiang, Jordan Taylor, and William Agnew. How Professional Visual Artists are Negotiating Generative AI in the Workplace, March 2026.

N. Johnson, S. Moharana, C. N. Harrington, N. Andalibi, H. Heidari, and M. Eslami. The fall of an algorithm: Characterizing the dynamics toward abandonment. In *FAccT 2024*, pp. 337–358, 2024. doi: 10.1145/3630106.3658910.

A. Karan, N. Vincent, K. Karahalios, and H. Sundaram. Algorithmic collective action with two collectives. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2025. doi: 10.1145/3715275.3732098.

A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2023. doi: 10.1145/3527848.

N. Karizat, D. Delmonaco, M. Eslami, and N. Andalibi. Algorithmic folk theories and identity: How TikTok users co-produce knowledge of identity and engage in algorithmic resistance. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):305:1–305:44, 2021. doi: 10.1145/3476046.

Seyyed Mohammad Sadegh Moosavi Khorzooghi, Poojitha Thota, Mohit Singhal, Abolfazl Asudeh, Gautam Das, and Shirin Nilizadeh. FairDeFace: Evaluating the Fairness and Adversarial Robustness of Face Obfuscation Methods, March 2025.

M. K. Kinney, T. E. Pearson, and J. Ralston Aoki. Improving "life chances": Surveying the anti-transgender backlash, and offering a transgender equity impact assessment tool for policy analysis. *Journal of Law, Medicine & Ethics*, 50(3):489–508, 2022. doi: 10.1017/jme.2022.89.

Sena A. Kojah, Ben Zefeng Zhang, Carolina Are, Daniel Delmonaco, and Oliver L. Haimson. "Dialing it Back:" Shadowbanning, Invisible Digital Labor, and how Marginalized Content Creators Attempt to Mitigate the Impacts of Opaque Platform Governance. *Proceedings of the ACM on Human-Computer Interaction*, 9(1):1–22, January 2025. ISSN 2573-0142. doi: 10.1145/3701191.

K. L. Kraschel, A. Chen, J. L. Turban, and I. G. Cohen. Legislation restricting gender-affirming care for transgender youth: Politics eclipse healthcare. *Cell Reports Medicine*, 3(8), 2022.

B. Kulynych, R. Overdorf, C. Troncoso, and S. Gürses. POTs: Protective optimization technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 177–188, 2020. doi: 10.1145/3351095.3372853.

Rena Li, Sara Kingsley, Chelsea Fan, Proteeti Sinha, Nora Wai, Jaimie Lee, Hong Shen, Motahhare Eslami, and Jason Hong. Participation and Division of Labor in User-Driven Algorithm Audits: How Do Everyday Users Work together to Surface Algorithmic Harms? In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, Hamburg Germany, April 2023. ACM. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3582074.

Lena Libon, Meghana Bhange, Rushabh Solanki, Elliot Creager, and Ulrich Aïvodji. Conscious data contribution via community-driven chain-of-thought distillation. In *NeurIPS 2025 Workshop on Algorithmic Collective Action*, 2025.

L. Lucero. Safe spaces in online places: Social media and LGBTQ youth. *Multicultural Education Review*, 2017.

Marília Marasciulo. How Anitta megafans gamed Spotify to help create Brazil's first global chart-topper. April 2022.

Sean McGregor. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 15458–15463, May 2021. doi: 10.1609/aaai.v35i17.17817.

M. McNamara, Q. McLamore, N. Meade, M. Olgun, H. Robinson, and A. Alstott. A thematic analysis of disinformation in gender-affirming healthcare bans in the United States. *Social Science & Medicine*, 351: 116943, 2024. doi: 10.1016/j.socscimed.2024.116943.

Alexander Monea. Cruising Tiktok: Using algorithmic folk knowledge to evade cisheteronormative content moderation. *AoIR Selected Papers of Internet Research*, December 2023. ISSN 2162-3317, 2162-3317. doi: 10.5210/spir.v2023i0.13466.

D. Myles, S. Duguay, and L. Flores Echaiz. Mapping the social implications of platform algorithms for LGBTQ+ communities. *Journal of Digital Social Research*, 5(4):1–26, 2023. doi: 10.33621/jdsr.v5i4.162.

M. Olson. *The Logic of Collective Action: Public Goods and the Theory of Groups.* Harvard University Press, 1971.

Ziqi Pan, Runhua Zhang, Jiehui Luo, Yuanhao Zhang, Yue Deng, and Xiaojuan Ma. From Platform Migration to Cultural Integration: The Ingress and Diffusion of #wlw from TikTok to RedNote in Queer Women Communities. In *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing*, pp. 227–232, Bergen Norway, October 2025. ACM. ISBN 979-8-4007-1480-1. doi: 10.1145/3715070.3749230.

A. Press. Reddit sues AI company Anthropic for allegedly 'scraping' user comments to train chatbot. June 2025.

A. Pullen Sansfaçon, M. A. Gelly, and K. Ens Manning. Affirmation and safety: An intersectional analysis of trans and nonbinary youths in quebec. *Social Work Research*, 45(3):207–219, 2021. doi: 10.1093/swr/svab009.

K. Purcell. ATS resume: How to create a resume that gets you noticed. Jobscan, September 2025.

Dorian Quelle, Frederic Denker, Prashant Garg, and Alexandre Bovet. Simple contagion drives population-scale platform migration, February 2026.

E. Redfield and I. Chokshi. Impact of the executive order redefining sex on transgender, non-binary, and intersex people. January 2025.

Harrison Rosenberg, Brian Tang, Kassem Fawaz, and Somesh Jha. Fairness Properties of Face Recognition and Obfuscation Systems, September 2022.

M. B. Ross, H. Jahouh, M. G. Mullender, B. P. C. Kreukels, and T. C. van de Grift. Voices from a multidisciplinary healthcare center: Understanding barriers in gender-affirming care–a qualitative exploration. *International Journal of Environmental Research and Public Health*, 20(14):6367, 2023. doi: 10.3390/ijerph20146367.

K. Rucker. LinkedIn sued over privacy violations, data collection for AI models. January 2025.

M. K. Scheuerman, J. M. Paul, and J. R. Brubaker. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–33, 2019. doi: 10.1145/3359246.

U. K. Schmid, A. S. Kümpel, and D. Rieger. How social media users perceive different forms of online hate speech: A qualitative multi-method study. *New Media & Society*, 26(5):2614–2632, 2024. doi: 10.1177/14614448221091185.

Andreas Schmitz and Mattia Samory. From Volunteerism to Corporatization: Analyzing Participation in the 2015 and 2023 Reddit Blackouts. *Social Media + Society*, 11(1):20563051241309497, January 2025. ISSN 2056-3051. doi: 10.1177/20563051241309497.

Z. Schor. Andersen v. stability AI: The landmark case unpacking the copyright risks of AI image generators. December 2024.

Mario D Schultz, Ludovico Giacomo Conti, and Peter Seele. Digital ethicswashing: A systematic review and a process-perception-outcome framework. *AI and Ethics*, 5(2):805–818, 2025.

Andrew D. Selbst. An institutional view of algorithmic impact assessments. *Harvard Journal of Law & Technology*, 35(1):117–192, 2021.

Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *Proceedings of the 29th USENIX Conference on Security Symposium*, SEC'20, USA, 2020. USENIX Association. ISBN 978-1-939133-17-5.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. Glaze: Protecting artists from style mimicry by text-to-image models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, pp. 2187–2204, USA, August 2023. USENIX Association. ISBN 978-1-939133-37-3.

Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y. Zhao. Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models. pp. 807–825. IEEE Computer Society, May 2024. ISBN 979-8-3503-3130-1. doi: 10.1109/SP54263.2024.00207.

R. Shelby, S. Rismani, K. Henne, Aj. Moon, N. Rostamzadeh, P. Nicholas, N. Yilla-Akbari, J. Gallegos, A. Smart, E. Garcia, and G. Virk. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 723–741, 2023. doi: 10.1145/3600211.3604673.

H. Shen, A. DeVos, M. Eslami, and K. Holstein. 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, 2021. doi: 10.1145/3479577.

M. Shen, Z. Liao, L. Zhu, K. Xu, and X. Du. VLA: A practical visible light-based attack on face recognition systems in physical world. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–19, 2019. doi: 10.1145/3351261.

Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, and Ross Anderson. Sponge Examples: Energy-Latency Attacks on Neural Networks. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 212–231, Vienna, Austria, September 2021. IEEE. ISBN 978-1-6654-1491-3. doi: 10.1109/EuroSP51992.2021.00024.

Stef M. Shuster, Grant Bunn, Kenneth Joseph, and Celeste Campos-Castillo. How False Information Is Used Against Sexual and Gender Minorities and What We Can Do About It. *Sex & Sexualities*, 1(1):59–66, May 2025. ISSN 3033-3717, 3033-3717. doi: 10.1177/3033371251329532.

Dorothee Sigg, Moritz Hardt, and Celestine Mendler-Dünner. Decline Now: A Combinatorial Model for Algorithmic Collective Action. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pp. 1–17, New York, NY, USA, April 2025. Association for Computing Machinery. ISBN 979-8-4007-1394-1. doi: 10.1145/3706598.3713966.

Ellen Simpson, Andrew Hamann, and Bryan Semaan. How to Tame "Your" Algorithm: LGBTQ+ Users' Domestication of TikTok. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–27, January 2022. ISSN 2573-0142. doi: 10.1145/3492841.

C. Snyder. Amazon “glitch” delists gay-themed books, interwebs cry foul. 2009.

Rushabh Solanki, Meghana Bhange, Ulrich Aïvodji, and Elliot Creager. Crowding Out The Noise: Algorithmic Collective Action Under Differential Privacy. In *NeurIPS 2025 Workshop on Algorithmic Collective Action*, November 2025.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond Memorization: Violating Privacy Via Inference with Large Language Models, May 2024.

Nikko Stevens and Amelia Lee Doğan. Trans data epistemologies: Transgender ways of knowing with data. *Big Data & Society*, 12(4):20539517251381694, December 2025. ISSN 2053-9517. doi: 10.1177/20539517251381694.

Zhensu Sun, Xiaoning Du, Fu Song, Mingze Ni, and Li Li. CoProtector: Protect open-source code against unauthorized training usage with data poisoning. In *Proceedings of the ACM Web Conference 2022*, Www '22, pp. 652–660, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9096-5. doi: 10.1145/3485447.3512225.

Fnu Suya, Saeed Mahloujifar, Anshuman Suri, David Evans, and Yuan Tian. Model-Targeted Poisoning Attacks with Provable Convergence. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 10000–10010. PMLR, July 2021.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, February 2014.

T. Tanni, M. Akter, J. Anderson, M. Amon, and P. Wisniewski. Examining the unique online risk experiences and mental health outcomes of LGBTQ+ versus heterosexual youth. In *Proceedings of the 2024 ACM Conference on Human Factors in Computing Systems*, pp. 1–21, 2024. doi: 10.1145/3613904.3642509.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing Machine Learning Models via Prediction APIs, October 2016.

H. A. Turner, D. Finkelhor, K. Mitchell, and D. Colburn. Prevalence of technology-facilitated abuse among sexual and gender minority youths. *JAMA Network Open*, 7(2):e2354485, 2024. doi: 10.1001/jamanetworkopen.2023.54485.

E. L. Ungless, N. Markl, and B. Ross. Experiences of censorship on TikTok across marginalised identities, 2024.

United States Department of Justice, Civil Rights Division. Louis et al. v. SafeRent et al. (D. Mass.), 2023.

U.S. Equal Employment Opportunity Commission. Mobley v. workday, inc., 2024.

N. Vincent, H. Li, N. Tilly, S. Chancellor, and B. Hecht. Data leverage: A framework for empowering the public in its relationship with technology companies. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 215–227, 2021. doi: 10.1145/3442188.3445885.

Nicholas Vincent and Brent Hecht. Can" conscious data contribution" help users to exert" data leverage" against technology companies? *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–23, 2021.

Nicholas Vincent, Brent Hecht, and Shilad Sen. “Data strikes”: Evaluating the effectiveness of a new form of collective action against technology companies. In *The World Wide Web Conference*, Www '19, pp. 1931–1943, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-6674-8. doi: 10.1145/3308558.3313742.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR, March 2018.

Lucas Waldron and Brenda Medina. When transgender travelers walk into scanners, invasive searches sometimes wait on the other side. August 2019.

S. Wang and O. T. Zhou. Being recognized in an algorithmic system: Cruel optimism in gay visibility on Douyin and Zhihu. *Sexualities*, 2022. doi: 10.1177/13634607221106912.

Gianna Williams, Natalie Chen, Michael Ann DeVito, and Alexandra To. Why Can't Black Women Just Be?: Black Femme Content Creators Navigating Algorithmic Monoliths. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, Yokohama Japan, April 2025. ACM. ISBN 979-8-4007-1394-1. doi: 10.1145/3706598.3713842.

B. Wodecki. Artists copyright claims against AI image generators advance in court. August 2024.

T. Wright et al. Accessing and utilising gender-affirming healthcare in the UK: Experiences of trans and non-binary individuals. *BMC Health Services Research*, 21:667, 2021. doi: 10.1186/s12913-021-06661-4.

Qing Xiao, Yuhang Zheng, Xianzhe Fan, Bingbing Zhang, and Zhicong Lu. Let's Influence Algorithms Together: How Millions of Fans Build Collective Understanding of Algorithms and Organize Coordinated Algorithmic Actions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pp. 1–19, New York, NY, USA, April 2025. Association for Computing Machinery. ISBN 979-8-4007-1394-1. doi: 10.1145/3706598.3713279.

Zhiyuan Yu, Shixuan Zhai, and Ning Zhang. AntiFake: Using Adversarial Audio to Prevent Unauthorized Speech Synthesis. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 460–474, Copenhagen Denmark, November 2023. ACM. ISBN 979-8-4007-0050-7. doi: 10.1145/3576915.3623209.

Kangyu Yuan, Li Zhang, Hanfang Lyu, Ziqi Pan, Yuanhao Zhang, Junze Li, Bingcan Guo, Jiaxiong Hu, Qingyu Guo, and Xiaojuan Ma. "I Love the Internet Again": Exploring the Interaction Inception of "TikTok Refugees" Flocking into RedNote. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–8, Yokohama Japan, April 2025. ACM. ISBN 979-8-4007-1395-8. doi: 10.1145/3706599.3719738.

K. Zochniak, O. Lewicka, Z. Wybrańska, and M. Bilewicz. Homophobic hate speech affects well-being of highly identified LGBT people. *Journal of Language and Social Psychology*, 42(4):453–463, 2023. doi: 10.1177/0261927X231174569.

Jonathan Zong and J. Nathan Matias. Data refusal from below: A framework for understanding, evaluating, and envisioning refusal as design. *ACM J. Responsib. Comput.*, 1(1), March 2024. doi: 10.1145/3630107.