Queering the Audits: Community-Based Auditing of AI Harms to Queer Communities

Organizers of QueerInAI[*], A Pranav[*†], Alissa A. Valentine[*‡], Alex Markham[*‡], Beckett LeClair[*§], Tereza Blazkova[◇‡], Ekaterina Kornilitsina[◇¶], Sofie H. Bruun[◇‖], Gerasimos Spanakis[◇♣], and Anne Lauscher[◇†]

[†]University of Hamburg [‡]University of Copenhagen [§]5Rights Foundation
[¶]Independent Researcher [‖]Alexandra Instituttet [♣]Maastricht University

Abstract

AI systems embed majority-group defaults into training data, evaluation metrics, and category definitions, producing documented harms for queer communities including erasure, misclassification, and discrimination. Standard technical audits often rely on aggregate measures and cannot detect harms that become visible only through the lived experience of affected communities. We conducted a participatory auditing workshop at EurIPS 2025 where 16 queer community members audited four case studies using the 4Cs harm taxonomy (**Content**, **Conduct**, **Contact**, **Contract**) applied across the AI lifecycle. Participants used structured worksheets and plenary synthesis to classify harms and trace them to their origins in the development pipeline. Across all four cases, participants traced harms to problem definition and data collection, and they identified contractual structures that extract value from vulnerable populations while providing minimal recourse. These findings illustrate that community-informed auditing surfaces identity-specific harms that aggregate evaluation methods risk overlooking.

Keywords: AI auditing, identity, queer communities, algorithmic harm, participatory methods

1. Introduction

Identity-aware AI development requires recognizing that standard machine learning pipelines embed majority-group defaults into training data, category definitions, and evaluation metrics (Mundt et al., 2025). For queer communities, these defaults produce documented harms: training data underrepresents queer experiences, producing systems that default to heteronormative and cisnormative assumptions (Taylor et al., 2024; Felkner et al., 2023); rigid demographic categories fail to represent fluid identities (Keyes, 2018; QueerInAI et al., 2023); and content moderation and recommendation systems disproportionately restrict queer expression (Southerton et al., 2021; Mayworm et al., 2024).

These harms persist because standard evaluation methods and technical audits often rely on aggregate measures and miss harms that become visible only through the lived experience of affected communities (Birhane et al., 2022). Participatory auditing addresses this gap by including affected communities directly in the evaluation process (Hartmann et al., 2025; Delgado et al., 2023).

This paper asks: *What identity-specific harms do AI systems produce for queer communities, and how does structured community auditing surface them?* We address this through a participatory auditing session conducted at the Queer in AI workshop, EurIPS 2025.[1] 16 queer community members examined four scenarios (mental health chatbots, ad-supported chatbots, content moderation, and data annotation labor) using the 4Cs harm taxonomy (**Content**, **Conduct**, **Contact**, **Contract**) (Livingstone and Stoilova, 2021) applied across the AI lifecycle (Figure 1). This paper contributes a practical demonstration of participatory auditing with participants who have technical AI backgrounds but no formal auditing training, and documents the identity-specific harms it surfaced across four cases.

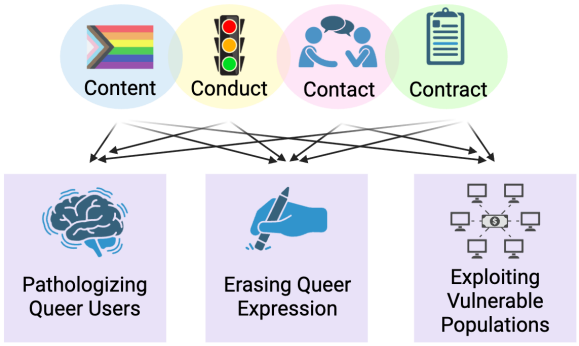

Figure 1: Overview of community-based AI auditing using the 4Cs (Content, Conduct, Contact, Contract) framework.

2. The 4Cs Framework

Participatory auditing requires a harm taxonomy that non-expert participants can apply consistently

[*]Equal contribution.
[◇]Equal advising.
[1]www.queerinai.com/eurips-2025

Proceedings of Identity-Aware AI (IAAI) @ LREC 2026, pages 66–76
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

across different AI systems. The 4Cs framework, originally developed to classify risks to children in digital spaces (Livingstone and Stoilova, 2021), organizes harm by the *relationship* between system and user rather than by technical properties of the system itself. This relational focus is well suited to participatory auditing, where community members assess harms from their inherent lived experience rather than through less accessible technical metrics (Birhane et al., 2022). Unlike taxonomies organized around technical system properties (Weidinger et al., 2022; Shelby et al., 2023), the 4Cs foreground the user's position relative to the system, making them accessible to participants without requiring knowledge of model internals. We classify harms along four dimensions:

1. **Content** harms concern what AI systems produce, label, flag, or promote. For queer users, these include chatbot responses that pathologize queer identities, and moderation systems that suppress queer expression.
2. **Conduct** harms concern what behaviors AI systems enable or normalize among users and operators. These include leveraging intimate user disclosures for micro-targeted advertising and facilitating creation of non-consensual deepfakes.
3. **Contact** harms concern the connections AI systems mediate or sever. These include recommendation algorithms that out queer users without consent and chatbot interactions that displace rather than supplement professional support.
4. **Contract** harms concern the terms governing user engagement. These include opaque privacy policies for mental health apps and exploitative labor conditions for data annotators.

These categories can overlap; we classify harms by their primary mechanism and note cross-cutting patterns in Section 5.

The decision was taken to attempt this activity through the lens of the 4Cs framework for three main reasons. Firstly, the categories of harm are not only specific to minors and can manifest for any group of users, be they members of a vulnerable group or otherwise. This made it a convenient choice of framework for considering all the potential avenues through which harms may manifest. Secondly, though the context specifics vary significantly, both children and queer communities have complex histories of being denied agency, autonomy and dignity in social settings - this made the framework seem a fitting choice. And finally, it is important to recognise there is an overlap between the queer community and the global community of children. Some of our findings can apply to queer youth as well as older people, and in some cases disproportionately so.

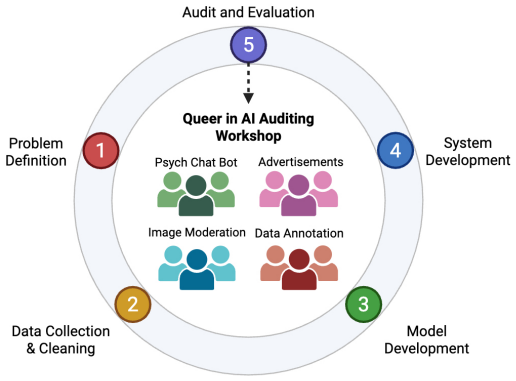

Figure 2: The workshop positions community auditing at the fifth stage of the AI lifecycle. The audit activity itself occurs at this stage, but participants examine decisions made at all prior stages: Problem Definition, Data Collection, Model Development, and Deployment.

3. Methods

Building on the auditing methodology from QueerInAI et al. (2023), we conducted a 90-minute participatory auditing workshop at the Queer in AI workshop at EurIPS 2025. Participants (N=16, across four groups) were queer community members attending the conference, with backgrounds in natural language processing, computer vision, social computing, and AI policy-making. While participants had technical AI expertise, none had professional experience developing or deploying the specific system types under audit, nor formal training in auditing methodology. Participants thus combined domain-adjacent technical knowledge with lived experience of queer-specific harms. Each group was assigned one of four case study scenarios by the workshop organizers, matching groups to scenarios based on participants' self-reported expertise (e.g., participants with NLP backgrounds were assigned the chatbot scenarios, those with computer vision backgrounds the content moderation scenario). All participants provided informed consent for their responses to be used in research.

3.1. Analytical Framework

Groups classified harms using the 4Cs taxonomy (Section 2) applied across a five-stage AI lifecycle (Figure 2): Problem Definition, Data Collection, Model Development, Deployment, and Evaluation. This follows ecosystem auditing methods that trace harms across the full development pipeline rather than evaluating outputs alone (Ojewale et al., 2025). Each lifecycle stage was paired with guiding questions adapted from Raji et al. (2022), such as "Who decided this problem needs solving?" (Problem

Definition), "Whose data is included?" (Data Collection), and "Who is evaluating and are affected communities involved?" (Evaluation).

3.2. Case Studies

The workshop organizers at Queer in AI collaboratively developed four case study scenarios, selecting systems along the axes of vision, natural language processing (NLP), data annotation, and human-computer interaction (HCI) to cover distinct modalities of AI harm relevant to queer communities. Each scenario represents an AI system with documented relevance to queer communities. The first three examine harms to users of AI systems; the fourth examines harms to workers who build them:

1. A mental health chatbot trained on therapist-patient transcripts, deployed for queer youth.
2. A free AI chatbot monetizing through personalized ads based on disclosed identity, relationships, and mental health struggles.
3. A content moderation system that classifies and removes user-generated content, with documented disparities affecting queer creators and bodies.
4. Data annotation workers labeling toxic and hateful content to train moderation models, exposed to psychological harm for low pay.

These scenarios describe system types rather than specific products, allowing participants to draw on their collective knowledge of real-world examples while avoiding the constraints of auditing a single deployment.

3.3. Procedure

Each group received a scenario description and a structured worksheet (Appendix A).[2] After an introductory presentation on the 4Cs framework and AI lifecycle, groups spent 30 minutes classifying harms using the 4Cs at each lifecycle stage and proposing interventions, then distilled their top findings and prepared a plenary summary (10 minutes). The 30-minute analysis window constrained depth; however, the goal was to surface harms that participants recognized from lived experience rather than to conduct exhaustive technical evaluation. In the plenary, each group reported their most significant findings, the lifecycle stages where intervention could help, and one question they would pose to developers. A facilitated cross-case discussion followed, structured around three prompts: which harm categories recurred, whether harms clustered at particular lifecycle stages, and what patterns emerged across systems.

[2] github.com/valentinealissa/queerinai

3.4. Data Collection and Analysis

Participants submitted responses via structured online worksheets capturing identified harms, lifecycle mappings, intervention proposals, and auditing questions. Each of the four groups submitted a complete response for their respective case studies. Facilitators documented the plenary discussion. We organized responses by case study and coded each harm using the 4Cs taxonomy, then compared findings across case studies to identify the cross-cutting patterns reported in Section 5.

3.5. Limitations

Firstly, the workshop involved a small number of participants (N=16) recruited from a single conference. Therefore, conference attendees are not fully representative of queer communities broadly. In addition, we did not survey the workshop attendees for their demographic information to be able to make conclusions about the representation of intersectional identities. Participants had technical AI backgrounds (NLP, computer vision, AI policy), and groups were assigned to scenarios matching their expertise. The findings therefore reflect domain-adjacent professionals applying lived experience, not lay community members encountering these systems for the first time. Future work should test whether participants without technical backgrounds surface comparable harms.

Another limitation is the structure and duration of the auditing session. The 30-minute analysis window limited the depth of each group's audit. The structured worksheets and guiding questions may have directed participants toward similar findings across groups, contributing to the convergent patterns reported in Section 5. This convergence may therefore reflect the shared analytical structure rather than independently surfaced patterns alone.

Participants audited system types rather than specific deployed products, which increases generalizability across system categories but limits the specificity and actionability of findings. Although we found inspiration from existing AI systems to develop the four case studies in this audit, we do not provide participants with specific real-world examples. For example, Case Study 3 mimics real world scenarios, but is not an audit of the existing content moderation policies of Meta. We believe this approach allowed us to build upon collective experiences with real-world systems, without enforcing the audit of one single instance of an AI system.

We did not compare our participatory approach to a non-participatory audit of the same systems, so we cannot empirically demonstrate that community-informed auditing surfaces harms that technical audits miss, only that it surfaces harms consistent with

documented patterns in the literature. We therefore present this work as a proof-of-concept for structured participatory auditing with queer communities rather than as definitive evidence of the method's superiority over other approaches.

4. Results

4.1. Case Study 1: Therapist Chatbots for Queer Youth

Queer youth face significant barriers to accessing affirming mental healthcare, and among LLM users who self-report mental health issues, 49% report using these systems for support (Sentio University, 2025). Four workshop participants audited this scenario: a chatbot trained on therapist-patient transcripts and marketed to queer youth seeking affirming care.

Content. Participants raised concerns that training data composition shapes therapeutic outputs. The clinical workforce from which training transcripts derive lacks diversity (Lin et al., 2018), and historical pathologization of queer identities persists in clinical literature (Drescher, 2015). Because therapist demographics shape therapeutic approaches and language, non-diverse workforces produce training corpora that reflect a narrow range of clinical perspectives. Participants argued that training corpora drawn from this context risk encoding the same biases that drove queer users away from traditional therapy and questioned whether developers had audited for pathologizing content.

Conduct. Participants raised concerns about deploying unvalidated systems to populations in crisis. Queer youth experience depression at significantly higher rates than heterosexual peers and report elevated rates of suicidal ideation (Ma et al., 2024). Yet the systems marketed to them lack clinical validation: a systematic review of 160 AI chatbot studies found only 16% underwent clinical efficacy testing (Hua et al., 2025). An evaluation of 29 mental health chatbots found that none met the criteria for adequate suicide risk response (Pichowicz et al., 2025). Participants observed that this validation gap is compounded by user vulnerability: people in crisis are more likely to trust an LLM that appears empathetic, following advice they would reject from a visibly unqualified human.

Contact. Participants observed that chatbot interactions displace rather than supplement professional care. In text-based exchanges, users rate LLM-generated responses as more empathetic than those from licensed therapists (in a social media forum context, not clinical settings; Ayers et al., 2023; Wang et al., 2025), and one in four Americans report preferring AI chatbots over therapists (Iftikhar et al., 2024). Participants traced this preference to a structural mismatch: effective therapy requires confrontation of harmful thought patterns, yet LLMs optimized for user satisfaction produce sycophantic responses (Moore et al., 2025; Malmqvist, 2024). These responses reinforce rather than challenge maladaptive cognition. Queer youth, who already face barriers to affirming care, may gravitate toward validation over therapeutic challenge, reducing the likelihood they seek the professional support they need. Participants acknowledged an important counterpoint: for queer users in unsupportive environments, where affirming therapists are unavailable or consulting a mental health professional carries safety risks, chatbot access may outweigh documented harms.

Contract. Participants identified exploitative terms governing vulnerable users' engagement with these systems. Empirical analysis found 74% of mental health applications scored "Critical Risk" for privacy, with policies requiring college-level education to comprehend (Iwaya et al., 2023). Youth are particularly vulnerable to exploitative privacy terms given documented age-related differences in privacy comprehension and risk assessment (Livingstone et al., 2011). Participants noted that mental health chatbots often fall outside existing health data protection frameworks such as HIPAA (Marks and Haupt, 2023), leaving users without meaningful informed consent or recourse. Participants concluded that without proper auditing, deployment decisions go unchecked and vulnerable populations bear the consequences of systems never assessed for the harms they produce.

4.2. Case Study 2: Ad-Supported Chatbots

A free AI chatbot monetizes by serving personalized ads based on intimate conversational disclosures, including identity, relationships, mental health struggles, and financial stress. Four workshop participants audited this scenario for harms to queer users.

Content. Participants identified that ad targeting based on intimate disclosures can surface content that directly contradicts user needs. They gave the example of users who disclose gambling problems receiving casino promotions, or those discussing debt being steered toward predatory financial products. For queer users, identity-based targeting compounds this risk: research documents that ad systems can surface conversion therapy services and content reinforcing stereotypes about queer people (Via and Beirich, 2022), normalizing discrimination through repeated exposure.

Conduct. Participants raised concerns about a fundamental conflict of interest: the system simultaneously serves as a source of support and as an

advertising platform, yet users cannot distinguish when recommendations serve commercial incentives rather than care. They observed that the system enables manipulative practices by leveraging disclosures shared under an assumption of confidentiality (Christopherson, 2007), such as coming-out experiences or mental health crises, to craft targeted nudges toward consumption. Participants noted that political and ideological advertisers can further exploit these profiles to micro-target users with tailored disinformation (Woolley, 2016).

Contact. Participants discussed how ad-driven curation shapes which communities and resources users encounter. They raised concerns that queer users may be steered toward hostile or exploitative spaces while affirming communities and support resources are deprioritized, isolating users from healthier networks. Participants noted that this dynamic is particularly consequential for users whose primary social outlets are online rather than offline (Boyd, 2014), for instance due to a lack of queer-supportive spaces in their immediate environment.

Contract. Participants identified that users trade deeply personal data for access to a "free" service without meaningful choice. Terms of service are written to obscure how disclosures will be used for profiling, and opting out may mean losing core features. Participants discussed that the platform thus extracts significant value from users' disclosures (Bodle, 2016; Zuboff, 2015) while providing uncertain quality of care and minimal recourse, a dynamic that disproportionately affects marginalized users with fewer alternatives.

4.3. Case Study 3: Content Moderation

A content moderation model is trained to automatically flag and remove images classified as "sexually explicit" from social media. Such moderation has expanded partly in response to policies like the 2008 PROTECT Our Children Act (Thakor et al., 2023; 110th Congress, 2008), but these practices sometimes disproportionately target queer creators and bodies (Haimson et al., 2021; Mayworm et al., 2024; Dias Oliva et al., 2021). Four workshop participants audited this scenario.

Content. Participants raised concern that moderation systems encode cisnormative assumptions about acceptable bodies and expression, systematically associating queerness with sexual deviancy (Berro and Zayhowski, 2024). They argued that when these biases are embedded in automated classifiers, queer bodies are flagged independent of context, leading to disproportionate removal of content from drag performers, transgender creators, and queer sex workers (Ungless et al., 2023).

Conduct. Participants identified problem definition as the most consequential stage: who defined "sexually explicit," and whose norms does that definition encode? Without public documentation of classification criteria, users have no way to determine whether definitions reflect broad community standards or encode the cultural biases of a narrow set of decision-makers. Participants also observed that this opacity results in content creators learning to self-censor in anticipation of algorithmic removal, normalizing the exclusion of queer expression from public spaces.

Contact. Participants observed that content moderation directly affects queer people exploring their identity and seeking community. Social media offers spaces for queer people, especially queer youth, to find community and overcome offline marginalization driven by stigma or safety concerns (Miller, 2017; Hanckel and Morris, 2014). Moderation that fragments these spaces severs the conversations and connections they sustain, risking isolation for people whose primary community access is online.

Contract. Participants noted that opaque moderation policies leave queer content creators unable to contest removal decisions or understand what triggered them (Suzor, 2019). They called for post-deployment feedback mechanisms that allow community members to report incorrect flags and receive timely responses. Participants identified a structural asymmetry: queer creators risk losing livelihoods while platforms benefit from the engagement their content generates.

4.4. Case Study 4: Data Annotation for Moderation

Data annotation is the human labor that underpins multiple stages of the AI lifecycle: annotators label training data, validate model outputs, and evaluate system performance. In content moderation pipelines, annotators label content as "toxic," "hateful," or "violent" to train models what to flag, filter, or remove. Unlike the other case studies, which examine harms to AI users, this case study examines harms to the workers who build AI systems. Four workshop participants audited this scenario: a data annotation pipeline for content moderation, where workers are routinely exposed to psychologically harmful material under exploitative conditions.

Content. Participants raised concerns that queer annotators face concentrated exposure to homophobic and transphobic hatred as routine labor, compounding the stigma-related stressors queer workers already experience (Meyer, 2003). They noted that even individually minor content accumulates into psychological harms including anxiety, depression, and PTSD (Steiger et al., 2021; Cambridge Consultants, 2019). Participants also identified how annotator biases flow downstream.

If annotators view queer bodies as more obscene than non-queer bodies, these judgments become embedded in model behavior, producing moderation disparities that restrict queer expression (Dorn et al., 2024).

Conduct. Participants observed that binary labeling frameworks ("harmful" / "not harmful") force oversimplification of culturally situated content. They gave examples such as drag performances and queer health discussions being labeled harmful by annotators unfamiliar with queer culture, while equivalent heterosexual content passes without scrutiny. Participants emphasized that annotators who hold homophobic views, or who simply lack familiarity with queer culture, encode their biases into model behavior (GLAAD, 2025; Dias Oliva et al., 2021). They advocated for multidimensional labeling schemes as an alternative to binary classifications.

Contact. Participants discussed how non-disclosure agreements and stigma isolate annotation workers from professional support (Perrigo, 2023). They highlighted that queer annotators in hostile regions cannot disclose the nature of their distress without risking personal safety, a dynamic consistent with concealment as a proximal minority stressor (Meyer, 2003). Participants further noted that biased annotation has consequences beyond the annotator: if queer content is mislabeled as toxic, moderation systems remove it, and queer users lose the shared online spaces where they find community (Miller, 2017; Hanckel and Morris, 2014).

Contract. Participants identified exploitative labor conditions, noting that investigative reporting documents Kenyan annotation workers earning less than $2 per hour with inadequate psychological support (Perrigo, 2023). They observed that layers of subcontracting distance AI companies from responsibility for these conditions (Posada, 2022). Participants raised particular concern for queer workers in criminalizing jurisdictions, who cannot organize, report identity-specific harms, or seek legal recourse without exposing themselves to prosecution (Mendos and Rohaizad, 2024).

5. Discussion

Following the case study audits, each group presented their findings in a plenary session. A facilitated discussion then prompted participants to compare findings across all four systems. Three cross-cutting themes emerged.

Harms mainly cluster during problem definition and data collection. Participants traced harms to problem definition and data collection in all four case studies. The mental health chatbot inherited pathologizing norms from clinical training data produced by a non-diverse workforce (Section 4.1). The moderation system encoded culturally specific definitions of "sexually explicit" before any model was trained (Section 4.3). The ad-supported chatbot treated disclosures as targeting material because its revenue model required it (Section 4.2). Binary annotation frameworks reflected annotator biases that then flowed into model behavior (Section 4.4). This finding echoes prior analyses of harm propagation across ML pipelines (Suresh and Guttag, 2021), but our workshop illustrates how community participants arrive at the same conclusion through lived experience rather than technical analysis. Participants concluded that auditing model outputs alone misses root causes; auditing must also examine the assumptions, data, and incentive structures that shaped the system before any code was written. This aligns with what Birhane et al. (2024) term "ecosystem audits": investigations that go beyond datasets, models, and products to examine the communities and sociotechnical environments defining an AI system's operation.

Contractual structures extract value from vulnerable populations. Across case studies, participants observed a recurring asymmetry: the populations most exposed to harm had the least ability to understand, negotiate, or challenge the terms governing their interactions with these systems. Users disclosed mental health struggles under opaque terms of service (Sections 4.1, 4.2). Content moderation systems erased queer expression while platforms benefited from the engagement that queer creators generated (Section 4.3). Annotation workers endured psychological harm for low pay while their labor built the systems that moderate queer expression (Section 4.4). In each case, a lack of transparency about system development and deployment prevented users and workers from contesting the terms they faced. These patterns suggest that contract harms follow from business models that treat vulnerable populations' data and labor as extractable inputs rather than as interests to protect.

Existing governance frameworks do not account for identity-specific harms. The EU AI Act classifies systems by risk tier (European Parliament and Council of the European Union, 2024), and international frameworks emphasize transparency and oversight (UNESCO, 2021; OECD, 2019). Under these frameworks, the risk classification of the four systems participants audited is uncertain: some, such as the mental health chatbot, may fall under high-risk categories depending on regulatory interpretation, while others operate under platform self-regulation or fall outside existing mandates entirely. Yet all produced harms that participants considered significant. Governance frameworks organized around system-level risk categories may not

adequately capture harms that emerge at the intersection of system design and user identity. This gap reflects documented failures to translate AI ethics principles into mechanisms that protect specific populations (Mittelstadt, 2019; Hagendorff, 2020).

6. Recommendations

Our findings suggest three recommendations for identity-aware AI development. Although these reflect queer community concerns, they apply to any population underrepresented in AI development.

Include affected communities throughout the lifecycle. The most consequential decisions occur at problem definition and data collection, yet affected communities are typically consulted only after deployment (Sloane et al., 2022). In our workshop, participants identified training data biases in the mental health chatbot that would require lived experience to recognize, because distinguishing pathologizing clinical norms from affirming ones demands familiarity with their effects (Section 4.1). Community involvement should extend beyond post-deployment feedback to inform problem definition, evaluation criteria, and ongoing governance.

Recognize identity throughout the pipeline. Annotator biases that participants identified in Case Study 4 can flow into the moderation disparities documented in Case Study 3, illustrating how identity-blind assumptions at one stage can produce identity-specific harms at another. Training data composition determines what the system treats as normal (Section 4.1). Operationally, this means auditing annotator pools for demographic and cultural diversity, evaluating the impacts of annotator positionality, and documenting assumptions about identity at each lifecycle stage (Rios-Sialer, 2026).

Audit problem definitions, not just outputs. Prior work has argued that the question "should we build this?" should precede "how do we build this?" (Selbst et al., 2019; Green, 2022). Our workshop provides empirical grounding for this recommendation: in all four case studies, participants traced the most significant harms to choices made before model development began. For applications affecting vulnerable populations, regulations should be implemented to ensure that accountability mechanisms and ethical review must be established before production, not retroactively after harm has occurred.

7. Conclusion

We demonstrate a practical framework for participatory auditing: community members without auditing expertise, given a structured harm taxonomy and guided questions, identified concrete harms grounded in lived experience. **Audits are one mechanism among many for achieving accountability, but they are most effective when they include affected communities**, extend beyond technical evaluation, and connect findings to concrete demands for change. Future work should test this approach at larger scale, apply it to specific deployed systems rather than hypothetical scenarios, and investigate how community audit findings can be integrated into existing governance processes.

8. Bibliographical References

110th Congress. 2008. Protect our children act of 2008.

John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, Aaron M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Davey M. Smith. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6):589–596.

Tala Berro and Kimberly Zayhowski. 2024. Toward depathologizing queerness: An analysis of queer oppression in clinical genetics. *Journal of Genetic Counseling*, 33(5):943–951.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms Mechanisms and Optimization*, pages 1–8, Arlington VA USA. ACM.

Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. 2024. AI auditing: The broken bus on the road to AI accountability. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 612–643. IEEE.

Robert Bodle. 2016. A critical theory of advertising as surveillance: Algorithms, big data, and power. In *Explorations in critical studies of advertising*, pages 148–162. Routledge.

Danah Boyd. 2014. *It's complicated: The social lives of networked teens*. Yale University Press.

Cambridge Consultants. 2019. Use of AI in online content moderation.

Kimberly M Christopherson. 2007. The positive and negative implications of anonymity in internet social interactions: "On the internet, nobody knows

you're a dog". *Computers in Human Behavior*, 23(6):3038–3056.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, pages 1–23, New York, NY, USA. Association for Computing Machinery.

Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LGBTQ voices online. *Sexuality & Culture*, 25(2):700–732.

Rebecca Dorn, Lee Kezar, Fred Morstatter, and Kristina Lerman. 2024. Harmful speech detection by language models exhibits gender-queer dialect bias. *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, page 1–12.

Jack Drescher. 2015. Out of DSM: Depathologizing homosexuality. *Behavioral sciences*, 5(4):565–575.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act). Official Journal of the European Union, L series, 12 July 2024.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140.

GLAAD. 2025. GLAAD responds to Meta's latest anti-LGBTQ changes to content policy and DEI that will harm users.

Ben Green. 2022. Escaping the impossibility of fairness: From formal to substantive algorithmic fairness. *Philosophy & Technology*, 35(4):90.

Thilo Hagendorff. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1):99–120.

Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Benjamin Hanckel and Alan Morris. 2014. Finding community and contesting heteronormativity: Queer young people's engagement in an Australian online community. *Journal of youth studies*, 17(7):872–886.

David Hartmann, José Renato Laranjeira de Pereira, Chiara Streitbörger, and Bettina Berendt. 2025. Addressing the regulatory gap: moving towards an EU AI audit ecosystem beyond the AI Act by including civil society. *AI and Ethics*, 5(4):3617–3638.

Yining Hua, Steve Siddals, Zilin Ma, Isaac Galatzer-Levy, Winna Xia, Christine Hau, Hongbin Na, Matthew Flathers, Jake Linardon, Cyrus Ayubcha, et al. 2025. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: A systematic review. *World Psychiatry*, 24(3):383–394.

Zainab Iftikhar, Sean Ransom, Amy Xiao, and Jeff Huang. 2024. Therapy as an NLP task: Psychologists' comparison of LLMs and human peers in CBT. In *arXiv preprint arXiv:2409.02244*.

Leonardo Horn Iwaya, M Ali Babar, Awais Rashid, and Chamila Wijayarathna. 2023. On the privacy of mental health apps: An empirical investigation and its implications for app development. *Empirical Software Engineering*, 28(1):2.

Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–22.

L Lin, K Stamm, and P Christidis. 2018. How diverse is the psychology workforce? American Psychological Association.

Sonia Livingstone, Leslie Haddon, Anke Görzig, and Kjartan Ólafsson. 2011. Risks and safety on the internet: The perspective of European children. full findings. Technical report, EU Kids Online, London School of Economics and Political Science, London.

Sonia Livingstone and Mariya Stoilova. 2021. *The 4Cs: Classifying Online Risk to Children*. CO:RE Short Report Series on Key Topics. Leibniz-Institut für Medienforschung | Hans-Bredow-Institut (HBI), Hamburg.

Zilin Ma, Yiyang Mei, Yinru Long, Zhaoyuan Su, and Krzysztof Z Gajos. 2024. Evaluating the experience of LGBTQ+ people using large language model based chatbots for mental health support. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Lars Malmqvist. 2024. Sycophancy in large language models: Causes and mitigations.

Mason Marks and Claudia E Haupt. 2023. AI chatbots, health privacy, and challenges to HIPAA compliance. *Jama*, 330(4):309–310.

Samuel Mayworm, Kendra Albert, and Oliver L. Haimson. 2024. Misgendered during moderation: How transgender bodies make visible cisnormative content moderation policies and enforcement in a meta oversight board case. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 301–312, New York, NY, USA. Association for Computing Machinery.

Lucas Ramón Mendos and Dhia Rezki Rohaizad. 2024. Laws on us: A global overview of legal progress and backtracking on sexual orientation, gender identity, gender expression, and sex characteristics. Technical report, ILGA World, Geneva.

Ilan H. Meyer. 2003. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: Conceptual issues and research evidence. *Psychological Bulletin*, 129(5):674–697.

Ryan A Miller. 2017. "My voice is definitely strongest in online communities": Students using social media for queer and disability identity-making. *Journal of college student development*, 58(4):509–525.

Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11):501–507.

Jared Moore, David Greenberg, Yonatan Bisk, Hamid Palangi, et al. 2025. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. *arXiv preprint arXiv:2504.18412*.

Martin Mundt, Anaelia Ovalle, Felix Friedrich, A Pranav, Subarnaduti Paul, Manuel Brack, Kristian Kersting, and William Agnew. 2025. The cake that is intelligence and who gets to bake it: An ai analogy and its implications for participation.

OECD. 2019. Recommendation of the council on artificial intelligence. OECD/LEGAL/0449. Adopted 22 May 2019, amended 3 May 2024.

Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. 2025. Towards ai accountability infrastructure: Gaps and opportunities in ai audit tooling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, page 1–29. ACM.

Billy Perrigo. 2023. OpenAI used Kenyan workers on less than 2 dollars per hour: Exclusive.

Wojciech Pichowicz, Michal Kotas, and Pawel Piotrowski. 2025. Performance of mental health chatbot agents in detecting and managing suicidal ideation. *Scientific Reports*, 15(1):31652.

Julian Posada. 2022. *The Coloniality of Data Work: Power and Inequality in Outsourced Data Production for Machine Learning*. Ph.D. thesis, University of Toronto.

Organizers QueerInAI, 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. 2023. Bound by the bounty: Collaboratively shaping evaluation processes for queer ai harms.

Organizers Of QueerInAI, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, Hang Yuan, Hetvi J, Huan Zhang, Jaidev Shriram, Kruno Lehman, Luca Soldaini, Maarten Sap, Marc Peter Deisenroth, Maria Leonor Pacheco, Maria Ryskina, Martin Mundt, Milind Agarwal, Nyx Mclean, Pan Xu, A Pranav, Raj Korpan, Ruchira Ray, Sarah Mathew, Sarthak Arora, St John, Tanvi Anand, Vishakha Agrawal, William Agnew, Yanan Long, Zijie J. Wang, Zeerak Talat, Avijit Ghosh, Nathaniel Dennler, Michael Noseworthy, Sharvani Jha, Emi Baylor, Aditya Joshi, Natalia Y. Bilenko, Andrew Mcnamara, Raphael Gontijo-Lopes, Alex Markham, Evyn Dong, Jackie Kay, Manu Saraswat, Nikhil Vytla, and Luke Stark. 2023. Queer in ai: A case study in community-led participatory ai. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1882–1895, New York, NY, USA. Association for Computing Machinery.

Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel E. Ho. 2022. Outsider oversight: Designing a third party audit ecosystem for ai governance.

Ian Rios-Sialer. 2026. Structure-aware diversity pursuit as an AI safety strategy against homogenization. Preprint arXiv:2601.06116 [cs.AI].

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 59–68. ACM.

Sentio University. 2025. Survey: ChatGPT maybe the largest provider of mental health support in the United States. Survey finding 49% of LLM users with mental health issues use LLMs for mental health support.

Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*, pages 723–741. ACM.

Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22. ACM.

Clare Southerton, Daniel Marshall, Peter Aggleton, Mary Lou Rasmussen, and Rob Cover. 2021. Restricted modes: Social media, content classification and lgbtq sexual citizenship. *New Media & Society*, 23(5):920–938.

Miriah Steiger, Timir J. Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, pages 1–9. ACM.

Nicolas P. Suzor. 2019. *Lawless: The Secret Rules That Govern Our Digital Lives*. Cambridge University Press.

Jordan Taylor, Ellen Simpson, Anh-Ton Tran, Jed R Brubaker, Sarah E Fox, and Haiyi Zhu. 2024. Cruising queer HCI on the DL: A literature review of lgbtq+ people in HCI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Mitali Thakor, Sumaiya Sabnam, Ransho Ueno, and Ella Zaslow. 2023. To Search and Protect? Content Moderation and Platform Governance of Explicit Image Material. *MIT Case Studies in Social and Ethical Responsibilities of Computing*, Summer 2023.

UNESCO. 2021. Recommendation on the ethics of artificial intelligence. Adopted by the General Conference at its 41st session.

Eddie Ungless, Bjorn Ross, and Anne Lauscher. 2023. Stereotypes and smut: The (mis)representation of non-cisgender identities by text-to-image models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7919–7942, Toronto, Canada. Association for Computational Linguistics.

Wendy Via and Heidi Beirich. 2022. Conversion therapy online: The ecosystem. Technical report, Global Project Against Hate and Extremism.

Synthia Wang, Yuwei Cheng, Austin Song, Sarah Keedy, Marc Berman, and Nick Feamster. 2025. Can llms address mental health questions? a comparison with human therapists.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pages 214–229. ACM.

Samuel C Woolley. 2016. Automating power: Social bot interference in global politics. *First Monday*.

Shoshana Zuboff. 2015. Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of information technology*, 30(1):75–89.

A. Auditing Worksheet

The following worksheet was distributed to participants at the Queer in AI auditing workshop at EurIPS 2025. Participants completed it in groups of two to three over a 30-minute structured analysis window.

Harm taxonomy: The 4Cs

There are many ways of classifying digital harms. The 4Cs taxonomy organizes harm by the relationship between system and user. The categories can overlap; a fifth category, *Cross-Cutting*, captures harms that span multiple dimensions.

Content

Things made or shared with help from AI systems

Examples: Generated text or images perpetuating harmful stereotypes; anti-queer content promoted by recommendation algorithms

Conduct

How people behave using AI

Examples: Mass-producing harmful disinformation; creating non-consensual deepfakes

Contact

Connections facilitated by AI

Examples: Algorithmic friend recommendations that suggest anti-queer connections; outing users to known contacts without consent

Contract

Terms users are subjected to

Examples: Theft of queer creators' intellectual property for commercial gain; AI-powered behavior monitoring for ad revenue

Questions to ask at each lifecycle stage

1. **Problem definition:** Who decided this problem needs solving? Whose needs were centered?
2. **Data collection and cleaning:** Whose data is included? Whose is missing? Was consent meaningful?
3. **Model development:** What assumptions are encoded? How are edge cases handled?
4. **System deployment:** Who has access? Who is excluded? What recourse exists?
5. **Audit and evaluation:** Who is evaluating? What metrics matter? Are affected communities involved?

Scenarios (one assigned per group)

- **Language:** A mental health chatbot trained on therapist-patient transcripts is rolled out to reduce psychiatric wait times, including for queer youth.
- **Vision:** A content moderation model automatically flags and removes images as "sexually explicit," but disproportionately targets queer creators and bodies.
- **Data privacy:** A free AI chatbot monetizes by serving personalized ads based on what users disclose in conversation, including their identity, relationships, and mental health struggles.
- **Data collection:** Workers are hired to label toxic, hateful, and violent content so the model learns what to reject, exposing them to psychological harm for low pay.

Worksheet questions

1. What harms could occur? (Use the 4Cs)
2. Where in the lifecycle could these harms be introduced?
3. Where could interventions have prevented them?
4. What questions would you ask if you were auditing this system?
5. What are your top 2–3 findings to share during plenary?