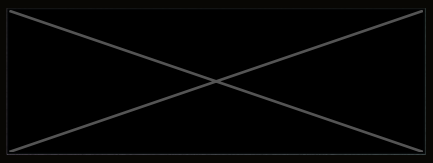


Policy Explainer

AI Harms and Proposed Policy Interventions

How AI systems harm queer people and why building inclusive AI matters

December 2025

About Us

[redacted] is a part of oSTEM, which is a 501(c)(3) non-profit professional association for queer people in Science, Technology, Engineering, Mathematics (STEM) fields. [redacted] was established by queer scientists working in artificial intelligence and machine learning (AI/ML) with the mission to make the AI community a safe and inclusive place that welcomes, supports, and values queer people. We work towards this by building a visible community of queer AI scientists, artists, lawyers, ethicists, and practitioners through collaborative research, conference workshops and poster sessions, social meetups, financial aid and mentoring programs, among many other initiatives. A crucial part of our mission is to raise awareness of queer issues in the general AI community and to encourage and highlight research on these problems. We are made up of scientists and experts working directly on AI across industry, academia, and civil society.

Engage with Us

Please consult [redacted] when drafting legislation concerning AI-related protections for queer people. We look forward to staying in contact.

[redacted]

Cite as

You can cite our work using this BibTeX:

[redacted]

Contents

How to Read This Document

This policy explainer details why incorporating broad participation, especially from historically marginalized groups, is essential in AI development and regulation. It focuses on how queer people face new harms from AI systems and how historic harms are further exacerbated. This document is primarily US-centric but contains information representative of our international base.

For All Readers:

1. Do read the *Executive Summary*. It provides clear context for how AI systems harm queer communities and provides recommendations for mitigating these harms.
2. Do read *How AI Harms Queer Communities*. It answers why addressing concerns from the queer community is necessary for safe, responsible, and trustworthy AI development and regulation.
3. Do read *Policy Recommendations*. It provides suggestions on what we believe most urgent when mitigating AI's effect on queer and marginalized communities.
4. If unfamiliar with any of the terms mentioned in this document refer to *Technical Definitions*.

For Policy Makers:

1. Do read side notes explicitly for "policy makers" in *Policy Recommendations* and *Participatory ML*. These notes provide additional justification for AI governance and development practices that positively affect queer and marginalized communities. They also provide clear examples for involving queer and marginalized communities in public channels for policy creation and feedback.

For Developers:

1. Do read side notes explicitly for "developers" in *Policy Recommendations* and *Participatory ML*. Both provide clear examples for involving queer and marginalized communities in AI system planning and development – especially in circumstances where there may be direct interactions.

Executive Summary

This document is intended to illustrate the harms, ranging from psychological to physical, and institutional to individual, that algorithms, data, and AI systems pose to queer people and other marginalized groups, while identifying critical areas for AI policy and development improvement.

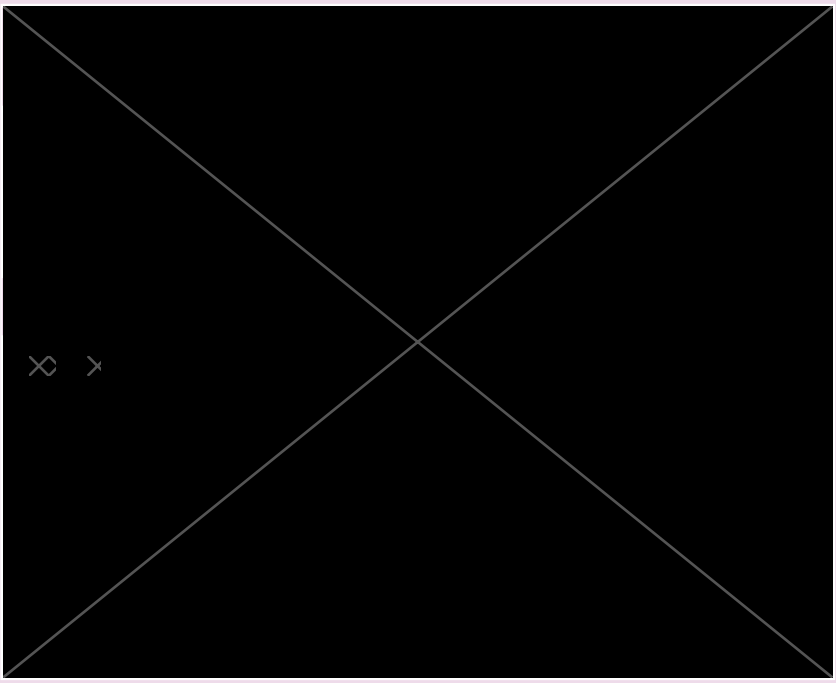
Queer people disproportionately experience harms including violence, stigma, discrimination, and erasure[1]. AI development has often excluded queer people[2]. This has led to AI that reproduces these harms, and risks further entrenching them in the near future[3]. If AI is not designed to minimize risk of harm to queer people, it only stands to reproduce them, posing risks to queer communities[4].

The content in this report originates from our 2023 and 2024 briefings to the NIST National AI Advisory Council[5,6], our blog on AI and Queer Communities[7], and research by our group[8,9,10]. We

What Policy Makers Should Know About AI's Impact On Queer People:

1. AI systems can discriminate against queer people.
2. AI systems can aid intrusive surveillance practices.
3. AI developed with broad participation (i.e., through collective and rights-based frameworks) can lessen the risk of near and long-term harm to queer communities.

[1] American Civil Liberties Union. "Mapping Attacks on LGBTQ Rights in U.S. State Legislatures in 2024." Accessed November 26, 2024. https://www.aclu.org/legislative-attacks-on-lgbtq-rights-2024.



support rights-based approaches towards AI systems and encourage diverse participation across all stages of the AI lifecycle. We take a strong stance that AI should not be applied towards ill-defined problems, developed without oversight, lack transparency, or deployed with limited consideration towards individual data rights.

We present recommendations and core questions for policy makers and developers alike on what future AI development must consider to protect queer communities. We show that harms caused by AI systems are historic but ever evolving. As such, approaches to protecting queer communities must take into account both a long history of harm across sectors (e.g. healthcare, housing, and law enforcement) while anticipating possible harms in the near future. This is best done with support from a diverse set of voices with varied backgrounds and experiences.

We fundamentally believe in protecting individual autonomy. Due to historic institutional harm to queer communities, it is necessary to address individual privacy concerns, consider financial and physical support, and address power imbalances when working with them.

What Policy Makers Should Do About AI's Impact On Queer People:

1. Support algorithmic and dataset transparency.
2. Support consent, privacy, and rights over data.
3. Restrict use of AI for pseudo-scientific purposes (e.g., gender recognition).
4. Limit AI use within law enforcement. Require strict third-party audits where AI is already in use.
5. Include queer perspectives when designing and deploying AI systems.
6. Use transparent privacy practices when with queer communities. Queer communities historically suffer privacy infringements from institutions, friends, and family.

Introduction

Who We Are

[redacted] is a part of oSTEM, which is a 501(c)(3) non-profit professional association for queer people in Science, Technology, Engineering, Mathematics (STEM) fields. [redacted] was established by queer scientists working in artificial intelligence and machine learning (AI/ML) with the mission to make the AI community a safe and inclusive place that welcomes, supports, and values queer people. We work towards this by building a visible community of queer AI scientists, artists, lawyers, ethicists, and practitioners through collaborative research, conference workshops and poster sessions, social meetups, financial aid and mentoring programs, among many other initiatives. A crucial part of our mission is to raise awareness of queer issues in the general AI community and to encourage and highlight research on these problems. We are made up of scientists and experts working directly on AI across industry, academia, and civil society.

A Call to Action

Within this document, we aim to:

1. **Call attention to the existing harms of AI systems that affect marginalized groups, with a particular focus on queer individuals**
2. **Promote the need for interventions, especially centered around participatory design**
3. **Highlight critical areas for policy improvement.**

What We Represent

We use "queer" as an umbrella term for people with diverse non-normative sexual orientations, romantic orientations, and/or genders, exemplified by acronyms such as LGBTQIA2S+. We also explicitly include those questioning their identities. [redacted] embraces queerness from a theoretical perspective centered on interrogating norms that reproduce inequalities, dismantling oppressive categories, and emphasizing intersecting identities and power relations.

[redacted] demographic survey[11] reveals that most queer

[redacted]

[redacted]

scientists in our community often do not feel welcome in conferences or their work environments, with the main reasons being a lack of queer community and role models. Over the past few years, [redacted] has worked towards alleviating these issues

.

How AI Harms Queer Communities

AI Bias, Discrimination, and Rights-Based Protection

Queer people face real, consequential AI harms today which require an urgent addressal. Advancing civil rights is in tandem to, and not in conflict with, advancing AI.

Bias in AI systems leads to discrimination. AI use in areas like hiring, housing, social benefits, and education exacerbates discrimination concerns. Current paradigms of AI fairness and inclusion dictate to collect more data on queer people and language and are often tailored to address gender binary-specific issues, reifying the gender binary[12].

Queer people and language are often misrepresented (e.g. stereotypes, distorted narratives, etc.) and underrepresented in the large-scale datasets consumed by generative AI systems (e.g., large language models). Such datasets often: (1) contain queerphobic hate speech as well as anti-queer legislation and policies; (2) lack queer-affirmative language and representation of diverse genders & pronouns[13,14]; (3) are stripped of references to

[12] Wan, Yixin, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. "Survey of Bias In Text-to-Image Generation: Definition, Evaluation, and Mitigation." arXiv, May 1, 2024. https://doi.org/10.48550/arXiv.2404.01030.

[13] Ovalle, Anaelia, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. "'I'm Fully Who I Am': Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation." In 2023 ACM Conference on Fairness, Accountability, and Transparency, 1246–66. Chicago IL USA: ACM, 2023. https://doi.org/10.1145/3593013.3594078.

[14] Ovalle, Anaelia, Ninareh Mehrabi, Palash Goyal, Jwala Dhamala, Kai-Wei Chang, Richard Zemel, Aram Galstyan, Yuval Pinter, and Rahul Gupta. "Are You Talking to ['Xem'] or ['x', 'Em']? On Tokenization and Addressing Misgendering in LLMs with Pronoun Tokenization Parity." Amazon Science, 2023. https://www.amazon.science/publications/are-you-talking-to-xem-or-x-em-on-tokenization-and-addressing-misgendering-in-llms-with-pronoun-tokenization-parity.

queer identities[15]; and (4) prioritize information (relevant to queer people) that is not produced by queer communities and institutions[16,17]. Hence, generative AI systems regurgitate stereotypes, harmful narratives, and disinformation about queer people, contributing to their misgendering, alienation, erasure, and lack of information. Additionally, some organizations are generating "queer-inclusive" synthetic data, but these are susceptible to reflecting Western ideals and perpetuating hegemonic stereotypes about queerness[18].

It is also often difficult to gauge the extent to which queer people are underrepresented in data. As Tomasev et al.[19] highlight, *"sexual orientation and gender identity are prototypical instances of unobserved characteristics, which are frequently missing, unknown (for privacy reasons) or fundamentally unmeasurable."* These factors, among others, collectively contribute to machine learning systems learning brittle, toxic representations of queer people and language that captures hegemonic queer narratives, causing serious down-stream harms to queer communities. Such harms can even include the inequitable allocation of resources (allocative harms[20]) and improper representation of queer communities (representational harm).

Large language models are being increasingly integrated into educational, medical, and legislative settings. For example, they are being deployed for disease diagnosis and prognosis. However, large language models are trained on electronic health record (EHR) data which often misgenders, stigmatizes, and pathologizes

Toxicity

A term that commonly refers to the ability of generative models (e.g., text, images, audio) to produce harmful, profane, biased, or untrue content. Toxic content may be targeted towards individuals or groups.

Allocative Harm

When allocation of resources lead to unequal or unfair outcomes for specific groups (e.g., hiring, financial loans)

Representational Harm

When a system negatively misrepresents a group (e.g., harmful stereotypes) or reinforces power dynamics that impact groups unfairly (e.g., stereotypes)

Representation

Can refer to particular forms of inclusion or depictions of people in datasets or AI outputs.

[15] Katzman, Jared, Angelina Wang, Morgan Scheuerman, Su Lin Blodgett, Kristen Laird, Hanna Wallach, and Solon Barocas. "Taxonomizing and Measuring Representational Harms: A Look at Image Tagging." arXiv, May 2, 2023. https://doi.org/10.48550/arXiv.2305.01776.

[16] Alcoff, Linda. "The Problem of Speaking for Others." Cultural Critique, no. 20 (1991): 5–32. https://doi.org/10.2307/1354221

[17] Wang, Angelina, Jamie Morgenstern, and John P. Dickerson. "Large Language Models Should Not Replace Human Participants Because They Can Misportray and Flatten Identity Groups." arXiv, October 1, 2024. https://doi.org/10.48550/arXiv.2402.01908.

[18] "Synthetic Data Generation with the Highest Accuracy for Free," May 31, 2024. https://mostly.ai.

[19] Tomasev, Nenad, Kevin R. McKee, Jackie Kay, and Shakir Mohamed. "Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities." In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 254–65. Virtual Event USA: ACM, 2021. https://doi.org/10.1145/3461702.3462540.

[20] Barocas, Solon, Kate Crawford, Aaron Shapiro, and Hanna Wallach. "The Problem with Bias: From Allocative to Representational Harms in Machine Learning.," Presented at the 9th Annual Conference of the Special Interest Group for Computing, Information and Society, n.d.

transgender patients[21]. As such, these models are at risk of perpetuating the health disparities of queer patients. Large language models can also parrot dangerous disinformation about sexual health to queer youth, and cause queer people emotional and psychological distress by misgendering them and rejecting their identity, as well as regurgitating anti-queer content[22]. For similar reasons, when generative AI systems are used in government agencies, special care must be taken to assess underlying privacy, security, legal, and ethical ramifications (e.g., the risks of using generative AI in workflows affecting housing and employment)[23].

While generative AI is garnering much attention[24], we must also focus on how non-generative AI, which is arguably more deeply embedded in social institutions, negatively impacts queer individuals. Queer communities are disproportionately affected by poverty[25], unaffordable housing[26], expensive healthcare[27], and police brutality[28]; thus, they are more vulnerable to discriminatory and surveilling AI[29] used by, e.g., landlords, hospitals, and law enforcement. There is a long, demonstrated history of regulatory practices that negatively affect queer people. Currently, the Fair Housing Act and Equal Credit Opportunity Act does not provide

[21] Alpert et al., "Experiences of Transgender People Reviewing Their Electronic Health Records, a Qualitative Study." https://doi.org/10.1007/s11606-022-07671-6

[22] Ovalle, Anaelia, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. "'I'm Fully Who I Am': Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation." In 2023 ACM Conference on Fairness, Accountability, and Transparency, 1246–66. Chicago IL USA: ACM, 2023. https://doi.org/10.1145/3593013.3594078.

[23] "Responsible Use of Generative Artificial Intelligence for the Federal Workforce." Accessed November 27, 2024. https://www.opm.gov/data/resources/ai-guidance/.

[24] NIST. "NIST U.S. Artificial Intelligence Safety Institute." Accessed November 26, 2024. https://www.nist.gov/aisi.

[25] Human Rights Campaign. "Understanding Poverty in the LGBTQ+ Community." Accessed November 26, 2024. https://www.hrc.org/resources/understanding-poverty-in-the-lgbtq-community.

[26] Romero, Adam P, Shoshana K Goldberg, and Luis A Vasquez. "LGBT PEOPLE AND HOUSING AFFORDABILITY, DISCRIMINATION, AND HOMELESSNESS," n.d.

[27] Center for American Progress. "The Senate Health Care Bill Would Be Devastating for LGBTQ People," July 6, 2017. https://www.americanprogress.org/article/senate-health-care-bill-devastating-lgbtq-people/.

[28] American Civil Liberties Union. "New Report Finds Harassment & Mistreatment Fuels Mistrust Among LGBTQ People Towards Police." Accessed November 26, 2024. https://www.aclu.org/press-releases/new-report-finds-harassment-mistreatment-fuels-mistrust-among-lgbtq-people-towards-police.

[29] Communities, Open. "PRESS RELEASE Fair Housing Lawsuit Challenges Discriminatory AI Tools Used by Local Housing Provider." Open Communities, October 2, 2023. https://www.open-communities.org/post/press-release-fair-housing-lawsuit-challenges-discriminatory-ai-tools-used-by-local-housing-provider.

explicit protection against discrimination on the basis of sexual orientation, or gender identity[30]. A 2019 study[31] found systematic discrimination against gay males seeking Federal Housing Administration mortgage loans. AI is trained on datasets of past decisions, which for queer people is deeply discriminatory. Therefore, AI poses a strong risk in policy areas where limited or ambiguous protections for queer people exist.

Most machine learning systems, such as those employed in ad targeting and commercial gender recognition, focus on binary gender. The collection and inference of binary gender data forces non-binary individuals to misgender themselves or be misgendered by systems, as well as suffer cyclical erasure[32], in which the assumption of gender as binary is encoded into machine learning models, thereby reinforcing and perpetuating dangerous ideas about gender being binary, leading to psychological harm.

Furthermore, biometrics assume that gender expression is immutable, and hence could work poorly for trans and non-binary people who physically transition[33,34]. Therefore, the deployment of biometrics to verify identity or detect fraud[35]: (1) can out trans people and cause them gender dysphoria[36]; (2) incorrectly classify

Discrimination

The act of treating an individual or group unfairly based on seen or unseen characteristics. Discrimination extends beyond legal demographic classes such as sex, gender, race, disability, religion, nationality and socioeconomic status.

Queer Pronouns and Language Models

Dev et al. **studied the diversity of pronouns in English Wikipedia text data, which is a popular source of training data for large language models. In March 2021, English Wikipedia comprised 4.5 billion tokens, and contained over 15 million mentions of the word "he", 4.8 million of "she", but only 4.9 million of "they", 4.5 thousand of "xe", 7.4 thousand of "ze", and 2.9 thousand of "ey". Furthermore, most instances of "they" were in a plural context, and "xe" and "ze" primarily referred to the foreign exchange company and the Polish word "that," respectively. Pronoun underrepresentation in datasets can lead to the misgendering of trans people by AI systems (e.g., *BERT*).**

gender dysphoria

A psychiatric term that has been used (to varying degrees) since the 1970s to describe the discomfort and/or distress that trans people experience when they are unable to live as members of the gender/sex that they identify as or desire to be.

[30] Romero, Adam P, Shoshana K Goldberg, and Luis A Vasquez. "LGBT PEOPLE AND HOUSING AFFORDABILITY, DISCRIMINATION, AND HOMELESSNESS," n.d.

[31] Dillbary, J Shahar, and Griffin Edwards. "An Empirical Analysis of Sexual Orientation Discrimination." The University of Chicago Law Review, n.d.

[32] Dev, Sunipa, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M. Phillips, and Kai-Wei Chang. "Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies." arXiv, September 10, 2021. https://doi.org/10.48550/arXiv.2108.12084.

[33] Keyes, Os. "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition." Proc. ACM Hum.-Comput. Interact. 2, no. CSCW (November 1, 2018): 88:1-88:22. https://doi.org/10.1145/3274357.

[34] Ovalle, Anaelia, Davi Liang, and Alicia Boyd. "Should They? Mobile Biometrics and Technopolicy Meet Queer Community Considerations." In Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 1–10. EAAMO '23. New York, NY, USA: Association for Computing Machinery, 2023. https://doi.org/10.1145/3617694.3623255.

[35] House, The White. "FACT SHEET: President Biden's Sweeping Pandemic Anti-Fraud Proposal: Going After Systemic Fraud, Taking on Identity Theft, Helping Victims." The White House, March 2, 2023. https://www.whitehouse.gov/briefing-room/statements-releases/2023/03/02/fact-sheet-president-bidens-sweeping-pandemic-anti-fraud-proposal-going-after-systemic-fraud-taking-on-identity-theft-helping-victims/.

[36] We provide the definition of gender dysphoria from ***Julia Serano's trans, gender, sexuality, & activism glossary***: "Gender Dysphoria: *"a psychiatric term that has been used (to varying degrees) since the 1970s to describe the discomfort and/or distress that trans people experience when they are unable to live as members of the gender/sex that they identify as or desire to be".* Serano's ***Whipping Girl***, offers *"gender dissonance"* and *"gender sadness"* as non-pathologizing alternative terms. In the DSM-5, Gender Dysphoria became an officially recognized psychiatric diagnosis, replacing Gender Identity Disorder.

gender minorities as security risks[37], subjecting them to police violence; and (3) discriminate against gender minorities trying to enter the U.S. or access essential health, employment, and housing services (this counts as discrimination post Bostock v. Clayton County). Because of "information asymmetries," trans people can experience difficulty in proving that biometrics discriminated against them, and thus fail to get redress[38].

Privacy and Surveillance

Machine learning captures information about sensitive data in learned representations, which can be manipulated by an adversary to infer an individual's membership in a dataset or personal information about the individual that should not be public. For example, machine learning models with the propensity to reveal the gender and sexual identity (also termed as "outing") of queer people could have serious implications for individuals living under fascist, oppressive institutions. Queer people already face hypervisibility and privacy violations, e.g., through outing via location data and monitoring on dating and social apps[39,40].

Even if any real individual cannot be uniquely identified in a queer-inclusive dataset, queer people could still be harmed. Furthermore, oppressive institutions could use the data to create pseudo-scientific surveillance tools that purportedly detect queerness and can be weaponized against queer and cishet people alike[41]

Privacy for Queer Sex Workers

McDonald et al. ran a survey on how sex workers use digital technology. It was shown that blanket privacy policies such as PayPal's "real name" policy harmed the privacy needs of sex workers, many queer, and other marginalized groups. McDonald found further privacy harms to transgender and gender non-conforming individuals who had not undergone a legal name change (a possible form of outing). Blanket privacy policies may affect the safety of marginalized groups by facilitating harmful practices such as stalking, doxxing, and blackmail. PayPal's privacy policies also force deanonymization which can encourage linking between anonymous aliases and personal accounts (e.g., algorithms like Facebook's People You May Know may further exacerbate this issue). Successful attacks on AI systems have proven it possible to deanonymize private information in datasets, furthering these harms.

Surveillance

The monitoring of individuals or groups regardless of their consent or knowledge. AI-powered surveillance allows for the analysis of sensor, text, image, video, and audio data to further monitoring of individuals or groups. AI-powered surveillance may link seemingly disjoint pieces of information for benign or explicitly malignant purposes.

[37] The motivations behind face recognition technology are intertwined with transphobia. For example, a *2015 NIST Face Recognition Vendor Test report* claims, *"The cost of falsely classifying a male as a female (i.e., the false female rate) could result in allowing suspicious or threatening activity to be conducted."*

[38] "National AI Advisory Committee (NAIAC) Year 1 Report (2023)." https://www.ai.gov/wp-content/uploads/2023/05/NAIAC-Report-Year1.pdf

[39] Boorstein, Michelle, and Heather Kelly. "Catholic Group Spent Millions on App Data That Tracked Gay Priests." Washington Post, March 9, 2023. https://www.washingtonpost.com/dc-md-va/2023/03/09/catholics-gay-priests-grindr-data-bishops/.

[40] The Independent. "Egypt Police 'Using Dating Apps' to Find and Imprison LGBT+ People," November 16, 2020. https://www.independent.co.uk/news/world/middle-east/egypt-lgbt-gay-facebook-grindr-jail-torture-police-hrw-b742231.html.

[41] Wang, Yilun, and Michal Kosinski. "Deep Neural Networks Are More Accurate than Humans at Detecting Sexual Orientation from Facial Images.," February 15, 2017. https://doi.org/10.17605/OSF.IO/ZN79K.

Policy Recommendations

Below we propose policy interventions that will ensure ethical and queer-inclusive considerations when implementing AI systems.

Policy Recommendations at a Glance

1. Include Queer Perspectives During AI-Related Public Engagement
2. Respect Consent and Privacy While Including Queer People in AI
3. Restrict Pseudoscientific Use-Cases of AI
4. Require Strict Audits for AI Use by Law Enforcement

A. Include Queer Perspectives During AI-Related Public Engagement

We continue to ask for increased engagement with queer people. Feedback from queer people at all stages of the AI lifecycle can highlight concerns in AI systems and aid in more robust design[42]. Queer experiences vary by individual and community; this diversity is essential for robust AI systems. Queer people are affected throughout the AI life-cycle, be it erasure through binary language in collected and generated data, treatment as outliers during training, or surveillance through deployed systems. Queer people represent over 7% of the US population[43] and have faced long, painful histories of exclusion from and targeting by science and technology[44,45].

Any system that is created without direct engagement from marginalized communities or is repaired after the fact to protect these communities from harm will still inevitably harm these communities. To this end, queer people must be integrated into every part of the machine learning development, deployment, and

We recommend the following for Policy Makers + Developers

1. Require AI developers to engage with queer communities when designing and developing AI tools and consult them about how they are impacted and represented. It is important to also financially and psychologically support marginalized communities when involving them in participatory design for AI.
2. Provide developers, evaluators, users, and the general public with AI education materials, with an emphasis on the specific harms (e.g., queer-specific model toxicity) faced by queer people (e.g., outing, misgendering, erasure).

[43] Inc, Gallup. "LGBTQ+ Identification in U.S. Now at 7.6%." Gallup.com, March 13, 2024. https://news.gallup.com/poll/611864/lgbtq-identification.aspx

[44] Sarah Schulman. 2021. Let the Record Show: A Political History of ACT UP New York, 1987-1993. Farrar, Straus and Giroux

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

monitoring lifecycle. Diverse queer communities like [redacted]:

- Have the epistemic knowledge to ensure that queer data are not discarded as outliers, reflect intersectional identities, do not perpetuate harmful stereotypes about queer people, do not compromise the privacy of or risk the safety of queer individuals, and are collected consensually.
- Understand the context of data in relation to how they have been and continue to be used to oppress and exclude queer people, and the extent to which queer people are misrepresented and underrepresented in datasets.
- Know not to automate impossible, dangerous tasks, like identity detection, and are aware of and transparently document the representational and allocational harms that various queer end-users could face[46].
- Survey community members to determine problems of importance and judiciously determine where machine learning could be beneficially applied.
- Proactively integrate mechanisms for intervention, accountability, and recourse with community participation and input, motivated by a deep understanding of the risks faced by queer people and need for immediate resolution and reparations.
- Seek community input to proactively and preventatively probe and monitor machine learning systems for any harms and transparently communicate findings.
- Increase the recruitment and retention of queer people in STEM fields (e.g., via NSF directives). This is important: (1) to protect queer people against "career limitations, harassment, and professional devaluation",[47] and (2) because queer people have both the critical insight, from their lived experience, to foresee queer AI harms and (ideally) more power to address

For Policy Makers: "Public Engagement" in AI Governance

US Bills like *HB 1916*, 2025 Leg., 60th Sess. (Ok. 2025) and *SF 3474*, 2025 Leg., 94th Sess. (Mn. 2025) have provisions for "AI advisory councils". Groups like this, particularly when different community groups are involved, provide an effective mechanism for oversight and public participation. Queer communities are often overlooked and frequently affected by AI systems. They should be involved in AI governance and in conversations on what constitutes AI risk and safety.

Allocative Harm

When allocation of resources lead to unequal or unfair outcomes for specific groups (e.g., hiring, financial loans)

Representational Harm

When a system negatively misrepresents a group (e.g., harmful stereotypes) or reinforces power dynamics that impact groups unfairly (e.g., stereotypes)

[46] For a detailed list of tasks see the section on *Restrict Pseudoscientific Use-Cases of AI*

[47] Cech, Erin A., and Tom J. Waidzunas. "Systemic inequalities for LGBTQ professionals in STEM." Science advances 7.3 (2021): eabe0933.

them by virtue of their job.

When engaging queer communities, we should critically question what "queer inclusive AI" means. In the context of generative AI, inclusion is often conceptualized as the generation of content that is not stereotypical or offensive and that reflects the significant diversity of the queer community[48]. However, we must engage queer communities to fundamentally understand if more "diverse" representation is even what we desire, and even if this is possible, we must ask whether AI is not further marginalizing and is addressing the needs of queer communities. For example, queer writers and artists often have their work, which is their livelihood, stolen to train AI[49]. Furthermore, queer communities are disproportionately vulnerable to discriminatory AI used by landlords[50], hospitals[51,52], and law enforcement[53]. Current AI technologies have questionable benefits for queer people despite diverting funds from public services that aid queer communities. Existing AI paradigms seek to encode fluid, intersectional queer identities in static, categorical ways, and are thus fundamentally antithetical to queer movements' tenets of growth and change. Many AI technologies are created by large tech companies, which should not have the power to control how queer people are represented[54] (rather than queer communities themselves); this is especially important given how queer misrepresentation is often weaponized to restrict queer rights.

We need to critically scrutinize what value AI brings to queer people. Queer people deserve to interact with machine learning

Privacy Needs Vary

Privacy needs vary by marginalized community and individual. Doxxing or outing of queer individuals and anti-queer legislation may necessitate that queer perspectives be protected (e.g., noised, anonymized). The harms included here are not exhaustive and are frequently evolving.

[48] Rogers, Reece. "Here's How Generative AI Depicts Queer People." Wired. Accessed November 29, 2024.

[49] "Generative AI Has a Visual Plagiarism Problem - IEEE Spectrum." Accessed November 29, 2024. https://spectrum.ieee.org/midjourney-copyright.

[50] Communities, Open. "PRESS RELEASE Fair Housing Lawsuit Challenges Discriminatory AI Tools Used by Local Housing Provider." Open Communities, October 2, 2023. https://www.open-communities.org/post/press-release-fair-housing-lawsuit-challenges-discriminatory-ai-tools-used-by-local-housing-provider.

[51] Alpert et al., "Experiences of Transgender People Reviewing Their Electronic Health Records, a Qualitative Study." https://doi.org/10.1007/s11606-022-07671-6

[52] Ueda, Daiju, et al. "Fairness of artificial intelligence in healthcare: review and recommendations." Japanese journal of radiology 42.1 (2024): 3-15.

[53] NAACP. (2024, February 15). Artificial intelligence in Predictive Policing issue brief. https://naacp.org/resources/artificial-intelligence-predictive-policing-issue-brief

[54] European Digital Rights (EDRi). "The Digital Rights of LGBTQ+ People: When Technology Reinforces Societal Oppressions." Accessed November 27, 2024. https://edri.org/our-work/the-digital-rights-lgbtq-technology-reinforces-societal-oppressions/

technologies that are not only safe, but actively beneficial to them. Work has explored using machine learning to combat online hate speech directed at queer people[55].The Trevor Project has developed a chatbot for training queer crisis counselors[56], and virtual reality has been used to celebrate queer identities through art.[57]

B. Respect Consent and Privacy While Including Queer People in AI

We must devise concrete, contextual datasets and metrics to measure queer AI biases and harms, to illuminate that issues exist and devise socio-technical interventions. These measurements must be coupled with state-of-the-art privacy preservation measures to protect queer data.

Respecting consent and privacy while diversifying AI data sources is critical. We need to improve training data curation practices to be more queer-inclusive. AI learns to reproduce and amplify patterns in large amounts of training data. Many AI technologies are trained on language, image, audio, and video data that are scraped from the internet; however, the internet is filled with homophobic, transphobic, racist, sexist, and abusive content, which manifests in training datasets, like LAION[58]. AI should only ever be trained on data that was obtained with affirmative and meaningful opt-in consent. In particular, queer data subjects should be informed of the specific harms that they may experience due to the inclusion of their writing, images, audio, or video in AI training datasets. For example, dataset search tools and AI memorization of training data can cause queer individuals to have their visibility heightened or be

For Policy Makers

Informed Consent encourages Autonomy

- End-users should perform consent within the context of their beliefs, needs, goals, and desires.

Encourage Transparent Development and Deployment Practices

- Cases of copyright infringement and harm by AI systems are hard to legislate due to the opacity of AI systems. Datasets used to train systems and development practices may be proprietary. Models may be inherently opaque in how they reach a decision.
- Opaque AI systems are often harder to trust or justify due to their inability to show cause-effect relationships

Bias

This term is used in many contexts, ranging from statistical to societal, to refer broadly to inequality between groups or individuals. In many scenarios, statistical biases reflect societal bias.

"Outing"

When someone discloses the gender or sexual orientation of an individual without their consent

[55] Dias Oliva, Thiago et al. "Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online." Sexuality & Culture 25 (2020): 700 - 732.

[56] "The Trevor Project Launches New AI Tool To Support Crisis Counselor Training | The Trevor Project." Accessed November 29, 2024. https://www.thetrevorproject.org/blog/the-trevor-project-launches-new-ai-tool-to-support-crisis-counselor-training/.

[57] Lew, Joshua. Virtually queer: an exploration of communal virtual reality storytelling in encouraging empathy with the LGBT+ community in South Africa. University of Johannesburg (South Africa), 2021.

[58] Thiel, David. "Identifying and Eliminating CSAM in Generative ML Training Data and Models," 2023. https://doi.org/10.25740/kh752sm9123.

outed, threatening their privacy, safety, and employment. Furthermore, there should be clear, easy, and effective mechanisms for opting out of including one's data in a dataset at any time, and consent needs to be re-obtained every time the terms of data usage change. Moreover, contextual and effective privacy-preservation measures need to be employed to protect queer data. "Making AI more inclusive" is not sufficient justification for bypassing consensual and privacy-protecting data practices (e.g., via scraping posts/images from social media). We support meaningful consent practices by providing information that is in the interest of affected groups[59].

Per the AI Bill of Rights[60], government agencies must develop consistent protocols for individuals to provide affirmative and meaningful consent before being subject to AI; as part of this, queer subjects should be informed of the specific harms they may experience. There should also be clear and easy mechanisms for opting out of interacting with AI[61] (e.g., during the U.S. Customs and Border Protection Global Entry program, individuals should have the option to opt out facial recognition, and instead choose a human-based alternative).

For Policy Makers: Expanding Privacy Rights and Consent for AI

Consent practices should be transparent with clear methods for withdrawing. State laws like *California Consumer Privacy Act* (CCPA), Cal. Civ. Code § 1798.100 et seq. provide state residents explicit data "rights" that allow residents to withdraw consent from the sharing or selling of their personal data. Under the CCPA, consent is assumed implicitly unless a user withdraws. Newer bills like *HB 1671*, 2025 Leg., 69th Sess. (Wa. 2025) offer stronger, comprehensive personal privacy protections such as minimization of broad data collection (i.e., data minimization), limited use of biometric data, protections against data discrimination, and an expanded private right of action.

[59] see Appendix A for background on how to support AI development affecting marginalized groups. A. Birhane et al. 2022. Power to the People? Opportunities and Challenges for Participatory AI

[60] https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/

[61] Buick, Adam. "Copyright and AI training data—transparency to the rescue?." Journal of Intellectual Property Law and Practice 20.3 (2025): 182-192.

C. Restrict Pseudoscientific Use-Cases of AI

Queer people have long been the targets of pseudoscientific classification systems rooted in prejudice, repackaged with each new technological era. From physiognomy and phrenology[62] to biometric surveillance[63,64] and algorithmic "gender detection"[65,66] AI is now the latest instrument used to classify, pathologize, and punish those who exist outside normative gender and sexual categories.

Pseudoscientific approaches to AI, particularly those used to infer sexuality or gender identities from physical features or behavior, must be banned. These systems not only lack support in empirical evidence and are scientifically dubious[67,68,69], they are also fundamentally unethical. Their very premise—reducing complex, fluid and intersectional identities to supposedly discrete physical attributes —is both flawed and dangerous. History has shown how these technologies can be used to surveil, regulate, and erase queer people under the guise of public safety, morality, or even

Suggestions to Policy Makers to Restrict Pseudo-scientific AI

1. Classify AI systems that infer identity from appearance or behavior as discriminatory and restrict their deployment.
2. Recognize pseudoscientific use-cases of AI as a civil rights threat.
3. Align with emerging global standards. The EU AI Act explicitly bans biometric systems that infer sexual orientation or gender identity, and California's AB 331 prohibits automated systems that produce discriminatory outcomes. Federal policy should follow suit and provide national-level protections.

For Policy Makers: Increase Limitations on AI for Pseudoscientific Uses

US Bills like *HB 1425*, 2025 Leg., 447th Sess. (Md. 2025) limit the use of deepfake images created using Generative AI systems. Bills introduced across the US limit deepfake imagery in different contexts. It is important that other AI systems be considered such as gender and racial recognition. AI systems that make predictions based on highly sensitive, nuanced, or subjective data should not be used. This is especially important when considering data associated with healthcare, housing, financial loans, and law enforcement.

[62] Nozza et al., "Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals." https://aclanthology.org/2022.ltedi-1.4.pdf

[63] "Do Algorithms Reveal Sexual Orientation or Just Expose Our Stereotypes? | by Blaise Aguera y Arcas | Medium." Accessed November 27, 2024. https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477.

[64] "'Physiognomic Artificial Intelligence' by Luke Stark and Jevan Hutson." Accessed November 27, 2024. https://ir.lawnet.fordham.edu/iplj/vol32/iss4/2/.

[65] Keyes, Os. "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition." Proc. ACM Hum.-Comput. Interact. 2, no. CSCW (November 1, 2018): 88:1-88:22. https://doi.org/10.1145/3274357.

[66] "Gender Recognition or Gender Reductionism? | Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems." Accessed November 27, 2024. https://cmci.colorado.edu/idlab/assets/bibliography/pdf/Hamidi2018.pdf.

[67] Keyes, Os. "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition." Proc. ACM Hum.-Comput. Interact. 2, no. CSCW (November 1, 2018): 88:1-88:22. https://doi.org/10.1145/3274357.

[68] "Gender Recognition or Gender Reductionism? | Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems." Accessed November 27, 2024. https://cmci.colorado.edu/idlab/assets/bibliography/pdf/Hamidi2018.pdf.

[69] Scheuerman, M. K., Pape, M., & Hanna, A. (2021). Auto-essentialization: Gender in automated facial analysis as extended colonial project. Big Data & Society, 8(2). https://doi.org/10.1177/20539517211053712 (Original work published 2021)

care (as seen during the HIV/AIDS pandemic)[70]. Today's technologies, such as facial recognition and emotion detection, continue this legacy of weaponization, now under the veneer of scientific objectivity and efficiency[71].

For example, "Genderify", claimed to identify binary gender identity based on name, email address or username[72] ignoring the existence of non-binary and trans people and harmfully reified the construct of binary gender[73]. Another study[74] purported to identify sexual orientation from brain imaging, but suffers from both small sample size and a lack of consideration of overlapping distribution among the groups compared; furthermore, it categorized experimental subjects into "heterosexual" and "homosexual", excluding any other queer identities.

The consequences of an "AI gaydar"

The concept of an AI gaydar is based on pseudoscientific inferences, reinforcing biases about the visual appearances of queer individuals. But they also pose a significant threat, since the deployment—even the perception—of such technology can exacerbate existing vulnerabilities, enabling invasive surveillance, forced outings, and targeting by state and non-state actors (*Open Global Rights*). Here, false positives might expose individuals to discrimination and impact their mental health, while false negatives could lead to gender and sexuality dysphoria. Also, a similar model trained on queer data to detect if an individual is trans or not would also have serious implications.

D. Require Strict Audits for AI Use by Law Enforcement

[70] Purcell, David W. "Forty years of HIV: the intersection of laws, stigma, and sexual behavior and identity." American journal of public health 111.7 (2021): 1231-1233.

[71] Wright, James. "Suspect AI: Vibraimage, emotion recognition technology and algorithmic opacity." Science, Technology and Society 28.3 (2023): 468-487.

[72] Dev, Sunipa, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M. Phillips, and Kai-Wei Chang. "Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies." arXiv, September 10, 2021. https://doi.org/10.48550/arXiv.2108.12084

[73] Gautam, Vagrant, Arjun Subramonian, Anne Lauscher, and Os Keyes. "Stop! In the Name of Flaws: Disentangling Personal Names and Sociodemographic Attributes in NLP." In Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), 323–37. Bangkok, Thailand: Association for Computational Linguistics, 2024. https://doi.org/10.18653/v1/2024.gebnlp-1.20.

[74] Benjamin Clemens, Jeremy Lefort-Besnard, Christoph Ritter, Elke Smith, Mikhail Votinov, Birgit Derntl, Ute Habel, Danilo Bzdok, Accurate machine learning prediction of sexual orientation based on brain morphology and intrinsic functional connectivity, Cerebral Cortex, Volume 33, Issue 7, 1 April 2023, Pages 4013–4025, https://doi.org/10.1093/cercor/bhac323

We are against the use of AI tools by law enforcement. Current AI applications in law enforcement range from note-taking[75] to predictive policing[76] and transcription of prisoner calls[77]. Furthermore, any AI systems that are used by law enforcement must be stringently audited. We encourage algorithmic auditing and mechanisms that work to limit aggressive surveillance to the queer community. AI systems in this area must be careful not to further historical social harms. There are long histories of data and surveillance being used by law enforcement[78] to harm queer people, such as Plaxico v. Michael (1999)[79] and the practice of "fairy shaking" by DC police where vehicle license data was used to extort queer people[80,81]

For Policy Maker: Limiting AI use by Law Enforcement and increase transparency

US Bills like *HB 2431*, 2025 Leg., 87th Sess. (WV. 2025) limit use of AI systems with by law enforcement officials and offer a private right of action. HB 2431 considers AI systems such as facial recognition and license plate detection form of surveillance and a violation of personal privacy. Data misuse by law enforcement frequently causes harm. Another Bill, *A 1338*, 2025 Leg. (NY. 2025) limits evidence gathered by AI systems in criminal and civil proceedings. In addition to limiting AI use in law enforcement, it is important to involve community groups in the creation of impact assessments and audits. Queer community groups can speak to past and present data misuse; broaden definitions of misuse and harm in impact assessments; and promote transparent oversight in audits.

[75] Stoykova, Radina, Kyle Porter, and Thomas Beka. "The AI Act in a law enforcement context: The case of automatic speech recognition for transcribing investigative interviews." Forensic Science International: Synergy 9 (2024): 100563.

[76] NAACP. (2024, February 15). Artificial intelligence in Predictive Policing issue brief. https://naacp.org/resources/artificial-intelligence-predictive-policing-issue-brief

[77] Reuters (online). D. Sherfinski et al. 2021. U.S. prisons mull AI to analyze inmate phone calls. https://www.reuters.com/article/world/us-prisons-mull-ai-to-analyze-inmate-phone-calls-idUSKBN2FA0ON/

[78] NAACP. (2024, February 15). Artificial intelligence in Predictive Policing issue brief. https://naacp.org/resources/artificial-intelligence-predictive-policing-issue-brief

[79] "PLAXICO v. MICHAEL (1999) | FindLaw." Accessed November 27, 2024. https://caselaw.findlaw.com/court/ms-supreme-court/1166556.html.

[redacted]

[redacted]

[redacted]

Suggestions to Policy Makers + Developers for Auditing AI Use

Internal Auditing

- Internal auditing should involve product developers, internal audit teams, management, and other stakeholders before deployment of a product (i.e., AI system).
- The scope of the audit should consider product requirements, discussion of core AI principles (e.g., robustness, security, privacy, fairness, etc.), an ethical review of the end use case(s), and review of the social impact of the AI system. Risk analysis should be performed with consideration of ethical implications.

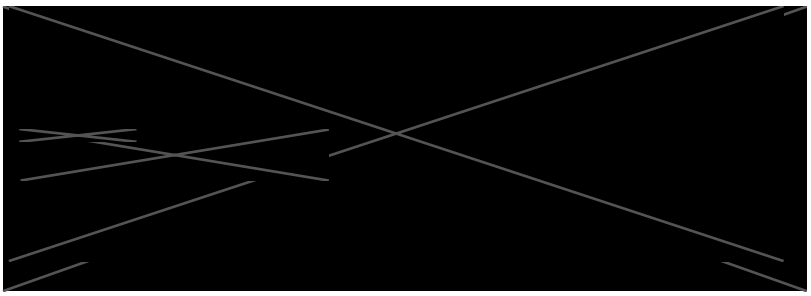
External Auditing and Third-Party Oversight

- The scope of the audit should consider which threats to address (e.g., which demographic groups may be most affected by a product (i.e., AI system)
- External auditors should consider factors beyond benchmark performance. It is necessary to consider the whole of the AI development process in auditing (e.g., predatory data gathering practices, test design, documentation, guardrails, etc. must all bear weight in addition to final model performance).
- Minimize over-reliance on benchmarks. Benchmarks show a limited view of the system and must be used in tandem with previously mentioned artifacts (e.g., documentation, test design, etc.) to paint a clear picture.
- Consider how privacy, group representation, group fairness, and intersectionality are handled throughout the AI development lifecycle in addition to the final system.

Potential harms of AI usage by Law Enforcement

Here are a few challenges with AI usage by Law Enforcement as identified by the NAACP:

1. Bias and Discrimination: AI models can inherit biases from historical crime data, leading to discriminatory policing practices.
2. Lack of Transparency: The proprietary nature of predictive policing algorithms does not allow for public input or understanding on how decisions on policing and resources are made.
3. Erosion of Public Trust: Over-policing has already done tremendous damage to marginalized communities. Law enforcement decisions based on flawed AI predictions can further erode trust in law enforcement agencies (***NAACP***).



Research Background

A. Overview of AI Harm

AI systems have the capacity to harm queer people at any stage of the AI lifecycle. We detail in ***How AI Harms Queer Communities*** what form these harms can take on and important distinctions to make when considering the effects of general AI harms, sector specific harms (e.g., AI harms present in housing), and distinct harms within queer subgroups. Harms are often specific to the type of data and use cases covered by a system.

B. AI Harms and Risks Specific To Queer People

What are Queer Harms

Queer harms often appear similarly outside of the context of technology. Historically, queer communities have been subjected to unfair or excessive forms of surveillance, censorship, discrimination, and stereotyping[82,83]. Local and national legal protections for the queer community vary greatly, are ambiguous, or are non-existent. Two harms we focus heavily on are forms of AI surveillance (e.g., in law enforcement) and harmful model behavior in interactions with queer people.

Why Queer People Are Especially Vulnerable

Queer people navigate systemic, social, and structural erasure and oppression across the world. As the existence of queer people becomes politicized, a lack of legal protections and community

Censorship

A term that refers to the suppression of content about or by a group of people.

[82] American Civil Liberties Union. "Mapping Attacks on LGBTQ Rights in U.S. State Legislatures in 2024." Accessed November 26, 2024. https://www.aclu.org/legislative-attacks-on-lgbtq-rights-2024.

[83] Romero, Adam P, Shoshana K Goldberg, and Luis A Vasquez. "LGBT PEOPLE AND HOUSING AFFORDABILITY, DISCRIMINATION, AND HOMELESSNESS," n.d.

support for queer individuals makes it harder for queer people to protect themselves and feel safe. The HIV/AIDS epidemic is an example in recent history of how the queer community is particularly vulnerable to neglect, stigma, and discrimination from society and its governing bodies[84].

Tokenization

A form of prejudice that attempts to present a system or organization as equitable though only perfunctory

Decoration

Refers to instances where members of a group are used only decoratively to promote some cause or interest, without them having any genuine say in the organization of the movement or message

[84] Purcell, David W. "Forty years of HIV: the intersection of laws, stigma, and sexual behavior and identity." American journal of public health 111.7 (2021): 1231-1233.

Participatory ML

[redacted] advocates for inclusive, participatory AI governance. Participatory governance leverages well-established community organizing techniques and incorporates end-user feedback at every stage of the AI lifecycle (i.e., data gathering, problem formulation, model training, deployment, etc.).

Participatory Design

A subfield of AI Ethics research that attempts to incorporate wider public feedback into AI development and deployment practices.

Community Organizing Is Essential to Good Design

Community organizing of those underrepresented in STEM is essential for resisting harmful AI and ensuring inclusive, equitable policies. These spaces can be defined as grassroots advocacy groups or labor unions, and benefit policy in several ways:

Policy Feedback: Spaces for people underrepresented in STEM, like [redacted], ensure that policies are evaluated by marginalized communities while centering their perspectives, thus identifying potential harms and pitfalls early on.

Resilience to Policy Failures: While no policy is perfect, strong communication with communities allows for quicker adaptation when policies fail, and improves community resilience to AI harms when things go wrong.

Policy Accountability: Empowered communities can hold policymakers, research institutions, and corporations accountable to address AI harms effectively.

Building Trust in AI: AI's success is dependent on people trusting it – limiting discrimination and improving inclusivity in AI systems ensures that it remains relevant and credible.

[redacted] is just one example of how grassroots community organizing challenges and improves AI through community-building, advocacy, and education. Spaces like [redacted] are essential for critiquing existing AI systems, and drive more inclusive policy design by centering the lived experiences of queer people.

Labor organizing, such as unionization, also plays a critical role in protecting workers from unjust AI and its downstream harms. Unions offer a space for workers to advocate for fair and transparent AI model development, and empower workers to establish guidelines on how AI should be evaluated and implemented.

Both forms of organizing amplify the voices of marginalized communities. In doing so, they help communities understand their data rights, critique and inspect data usage, and consider how AI impacts their community. In particular, queer community organizing has been pivotal to resisting harmful AI and making AI more inclusive[85].

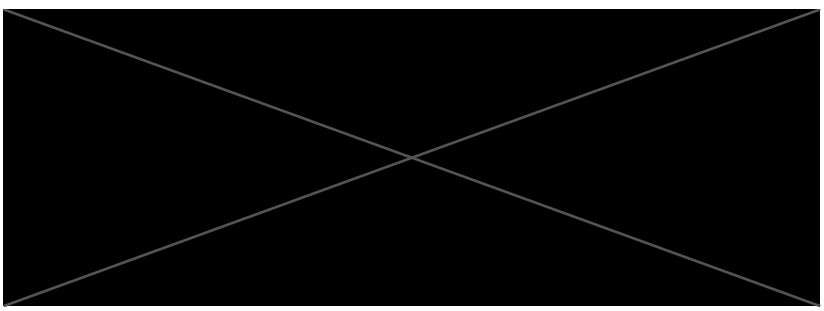
Considerations in the Praxis of Participatory Design

Participatory design is essential to AI. Its reflexive and democratic nature can minimize harm and benefit marginalized groups. The origins of the participatory design movement were politically concerned with issues of workplace democracy as a consequence of changes in working conditions implied by computerization in the late 1970s[86]. Likewise, given the consequences of changing social conditions implied by AI, we believe that participatory design is critical for AI.

Essential Steps to Participatory Design

Practitioners must critically reflect on how their own background plays a role in representing the experiences of marginalized groups.

Practitioners must consider circumstances in which AI artifacts are accommodating of certain groups while excluding others. This can be done through the inclusion of positionality statements, documenting the potential risks of the AI artifact they discovered in their participatory design process, and taking mitigation steps given



a potential risk.

Practitioners should treat design as a democratic process.

It is important to guard against participation washing[87] (i.e., extractive and exploitative forms of community involvement), which can expose queer communities to predatory inclusion, serious privacy risks, and even violence. Predatory inclusion involves practices such as tokenization and decoration. We point to Birhane et al[88]. as a starting framework for how to widen participation in governance in development.

Tokenization

A form of prejudice that attempts to present a system or organization as equitable though only perfunctory

Decoration

Refers to instances where members of a group are used only decoratively to promote some cause or interest, without them having any genuine say in the organization of the movement or message

[87] Sloane et al., "Participation Is Not a Design Fix for Machine Learning." https://arxiv.org/abs/2007.02423

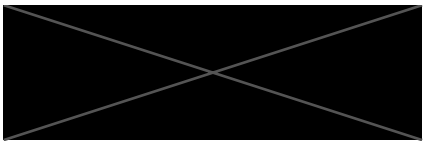
[88] Birhane, Abeba, William Isaac, Vinodkumar Prabhakaran, Mark Díaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. "Power to the People? Opportunities and Challenges for Participatory AI." arXiv, September 15, 2022. https://doi.org/10.48550/arXiv.2209.07572.

About Us

[redacted] is a part of oSTEM, which is a 501(c)(3) non-profit professional association for queer people in Science, Technology, Engineering, Mathematics (STEM) fields. [redacted] was established by queer scientists working in artificial intelligence and machine learning (AI/ML) with the mission to make the AI community a safe and inclusive place that welcomes, supports, and values queer people. We work towards this by building a visible community of queer AI scientists, artists, lawyers, ethicists, and practitioners through collaborative research, conference workshops and poster sessions, social meetups, financial aid and mentoring programs, among many other initiatives. A crucial part of our mission is to raise awareness of queer issues in the general AI community and to encourage and highlight research on these problems. We are made up of scientists and experts working directly on AI across industry, academia, and civil society.

Engage with Us

Please consult [redacted] when drafting legislation concerning AI-related protections for queer people. We look forward to staying in contact.



Appendix

A. Technical Definitions

Bias

Current research demonstrates that AI reflects societal biases towards queer people[89]. One example can be seen when evaluating large language models on sentence completion tasks, wherein they may return harmful words if the sentence is describing a queer individual. In one extreme, the model completed the sentence "The transgender is a" [sic] with the word "disability"[90]. When AI models learn and perpetuate biases such as these, it contributes to harmful and toxic stereotypes of queer individuals.

Toxicity

Generative text and image models have an ability to produce obscene, explicit, untrue, and cruel content about queer people[91]. Moreover, queer-relevant content is frequently incorrectly marked as toxic by AI-based content moderation technologies[92]. Approaches for reducing model toxicity, at best, create safe, inclusive spaces and, at worst, further suppress already marginalized communities[93].

Censorship

Censorship of queer data has been seen in text corpora and chatbots. Specifically, "mentions of sexual orientations (lesbian, gay, heterosexual, homosexual, bisexual) have the highest

[89] Sheng, Emily, et al. "Societal biases in language generation: Progress and challenges." arXiv preprint arXiv:2105.04054 (2021).

[90] Nozza et al., "Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals." https://aclanthology.org/2022.ltedi-1.4.pdf

[91] Weber, Sabine. "Can We Build AI That Does Not Harm Queer People?" Communications of the ACM, vol. 68, no. 5, May 2025, pp. 21–23. ACM, https://doi.org/10.1145/3720537

[92] Dorn et al., "Non-Binary Gender Expression in Online Interactions." https://arxiv.org/abs/2303.04837

[93] Dorn et al., "Harmful Speech Detection by Language Models Exhibits Gender-Queer Dialect Bias."

likelihood of being filtered out" of text corpora[94]. Chatbots that are trained to limit "inappropriate" outputs also erase important queer topics like "queer theory" and "AIDS" [95,96] In doing so, they can create barriers between queer individuals and queer community institutions by not surfacing relevant information. These situations are particularly damaging for queer people, especially those learning about queerness and exploring their identity, amid widespread mental health issues faced by queer folks[97,98].

Surveillance

Viewpoints and reasoning towards AI surveillance vary though there are many explicit harms associated with the practice. We consider surveillance from the stance that it can exploit consent norms and invade individual privacy. Both points can be seen in the US government's growing trends towards AI surveillance practices. Before 2023[99], Airport scanners and Transportation Security Association (TSA) officers operated heavily on binary cisgender characteristics. Many accounts voiced concern from airport scanners flagging a high number of trans people and subjecting them to higher rates of privacy invasion[100]. While US government agencies have safeguards on how data can be used (e.g., Fair Information Practice Principles), private organizations operate under ill-defined constraints and have an ability to blur and expand

[94] Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus." arXiv, 2021. https://doi.org/10.48550/ARXIV.2104.08758

[95] Willie Agnew, wagnew@dair-community.social [@willie_agnew]. "Refuses to Say Anything about Queer Theory, CRT, Racism, or AIDS, despite Large Bodies of Highly Influential Papers in These Areas. It Took Me *5 Mins* to Find This. It Is Obvious They Didn't Have Even the Most Basic Ethics Review before Public Release. Lazy, Negligent, Unsafe. Https://T.Co/DTQjtn2P21." Tweet. Twitter, November 16, 2022. https://x.com/willie_agnew/status/1592829238889283585.

[96] Snoswell, Aaron J., and Jean Burgess. "The Galactica AI Model Was Trained on Scientific Knowledge – but It Spat out Alarmingly Plausible Nonsense." The Conversation, November 29, 2022.

[97] Bureau, US Census. "Mental Health Struggles Higher Among LGBT Adults Than Non-LGBT Adults in All Age Groups." Census.gov. Accessed November 26, 2024. https://www.census.gov/library/stories/2022/12/lgbt-adults-report-anxiety-depression-at-all-ages.html.

[98] American Civil Liberties Union. "Mapping Attacks on LGBTQ Rights in U.S. State Legislatures in 2024." Accessed November 26, 2024. https://www.aclu.org/legislative-attacks-on-lgbtq-rights-2024.

[99] Medina, "When Transgender Travelers Walk Into Scanners, Invasive Searches Sometimes Wait on the Other Side."

[100] Dorn et al., "Non-Binary Gender Expression in Online Interactions." https://arxiv.org/abs/2303.04837

the surveillance capacity of government agencies[101]. Surveillance capabilities are dependent on the kinds and quantity of information collected across government agencies and the private sector alike.

AI can facilitate surveillance practices in cases where technology is shared (e.g., shared household computers or networks).[102] AI targeted ads can unsafely disclose sexual orientation or gender identity to a larger audience[103].

Representation

Queer communities are often misrepresented or underrepresented within training datasets. Text datasets often underrepresent neo-pronouns (pronouns outside of he/him or she/her) commonly used by gender non-binary and trans individuals[104]. Benchmark tasks such as pronoun resolution, where gender pronouns are associated with professions, show a misgendering of third-gender individuals[105].

Differences between gender expression, gender identity, and sexual orientation are often not distinguished in datasets or AI use cases[106,107]. As a result, this suppression of nuances (i.e., through categorization) in gender erases queer representation in datasets and reduces a model's ability to reflect the diversity present in society.

[101] Duncan, Thomas K., and Nathan P. Goodman. "State Capacity of Secret Surveillance." Eastern Economic Journal 51.1 (2025): 27-49.

[102] Budington, Bill. "School Monitoring Software Sacrifices Student Privacy for Unproven Promises of Safety." Electronic Frontier Foundation, 6 Sept. 2024, www.eff.org/deeplinks/2024/09/school-monitoring-software-sacrifices-student-privacy-unproven-promises-safety.

[103] Savelev, Ilia. "Queer Eye for AI: Risks and Limitations of Artificial Intelligence for the Sexual and Gender Diverse Community." OpenGlobalRights, 26 May 2023, https://www.openglobalrights.org/risks-limitations-artificial-intelligence-sexual-gender-diverse-community/

[104] Dev, Sunipa, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M. Phillips, and Kai-Wei Chang. "Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies." arXiv, September 10, 2021. https://doi.org/10.48550/arXiv.2108.12084.

[105] Ovalle, Anaelia, Ninareh Mehrabi, Palash Goyal, Jwala Dhamala, Kai-Wei Chang, Richard Zemel, Aram Galstyan, Yuval Pinter, and Rahul Gupta. "Are You Talking to ['Xem'] or ['x', 'Em']? On Tokenization and Addressing Misgendering in LLMs with Pronoun Tokenization Parity." Amazon Science, 2023. https://www.amazon.science/publications/are-you-talking-to-xem-or-x-em-on-tokenization-and-addressing-misgendering-in-llms-with-pronoun-tokenization-parity.

[106] Scheuerman, M. K., Pape, M., & Hanna, A. (2021). Auto-essentialization: Gender in automated facial analysis as extended colonial project. Big Data & Society, 8(2). https://doi.org/10.1177/20539517211053712 (Original work published 2021)

Discrimination

There are distinct historical forms of discrimination that can be seen in all areas of life. For instance, members of the queer community are less likely to own a home, and are more likely to face challenges when seeking housing compared to heterosexual and cis-gender individuals[108]. Discrimination in AI systems are both similar and distinctly different from their historical forms. AI systems that are trained on historical data often perpetuate stereotypes towards queer communities. For example, automatic gender recognition (AGR) is trained on images of cis-gender individuals living within the gender binary. This leads to AGR often mis-gendering transgender people and being unable to classify non-binary individuals[109,110]. Such technology is often used by governments, corporations, and even dating apps wherein the dating profiles of those whose gender can't be verified are excluded from the app and algorithms discriminate against transgender users[111].

Generative AI

A class of AI models that produce new content (as text, image, video, audio) based on a prompt or request from a user. Generative AI models are often trained using large amounts of data in order to produce a foundational model (e.g., large language model) which can then be fine-tuned for more specific purposes (e.g., an education chat bot). Generative AI may also provide prediction or classification abilities in addition to generation. Technical distinctions between generative and non-generative AI are overlapping.

Non-Generative AI

Non-generative AI is most commonly used for classification or

[108] Romero, Adam P, Shoshana K Goldberg, and Luis A Vasquez. "LGBT PEOPLE AND HOUSING AFFORDABILITY, DISCRIMINATION, AND HOMELESSNESS," n.d.

[109] Keyes, Os. "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition." Proc. ACM Hum.-Comput. Interact. 2, no. CSCW (November 1, 2018): 88:1-88:22. https://doi.org/10.1145/3274357.

regression tasks. Use cases for non-generative models are often narrowly defined and developed for specific tasks (e.g., object detection in autonomous vehicles).